ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΝΑΛΟΓΙΣΤΙΚΩΝ-
ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΩΝ ΜΑΘΗΜΑΤΙΚΩΝ

«ΤΟΠΟΛΟΓΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΦΑΡΜΟΓΗ ΣΤΗ ΒΑΘΜΟΝΟΜΗΣΗ ΠΙΣΤΟΛΗΠΤΙΚΗΣ ΙΚΑΝΟΤΗΤΑΣ»

Διπλωματική Εργασία για το Πρόγραμμα Μεταπτυχιακών Σπουδών
«Στατιστική και Αναλογιστικά-Χρηματοοικονομικά Μαθηματικά»

ΧΑΡΜΠΗ ΜΥΡΤΩ

ΧΑΡΜΠΗ ΜΥΡΤΩ

ΤΟΠΟΛΟΓΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΦΑΡΜΟΓΗ ΣΤΗ ΒΑΘΜΟΝΟΜΗΣΗ ΠΙΣΤΟΛΗΠΤΙΚΗΣ ΙΚΑΝΟΤΗΤΑΣ

09/02/2023

Διπλωματική Εργασία για το Πρόγραμμα Μεταπτυχιακών Σπουδών

«Στατιστική και Αναλογιστικά-Χρηματοοικονομικά Μαθηματικά»

Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών

Συγγραφέας : Χαρμπή Μυρτώ

Επιβλέπων : Ξανθόπουλος Στυλιανός

Μέλος επιτροπής: Λάππας Παντελής

Μέλος επιτροπής: Ταχτσής Ελευθέριος

ΣΑΜΟΣ

Ευχαριστίες

Εάν η αρχή είναι πράγματι το ήμισυ του παντός, θα πρέπει να ευχαριστήσω πρώτα από όλους τον πατέρα μου που μου έδειξε το σύμπαν και τα μαθηματικά.

Ευχαριστώ τη μητέρα μου και την αδερφή μου που σηκώνουν λίγο από το βάρος της καθημερινότητας.

Ευχαριστώ τη φίλη και συμφοιτήτριά μου Μαρία Τόκα που πιστεύει σε μένα περισσότερο από όσο πιστεύω εγώ.

Τέλος, θέλω να εκφράσω την ευγνωμοσύνη μου απέναντι στον κ. Ξανθόπουλο. Αρχικά για την βοήθεια, την καθοδήγηση και την ενθάρρυνση που μου παρείχε σε όλα τα στάδια της εργασίας και ακόμα περισσότερο για την κατανόηση και τον σεβασμό του απέναντι στις ανησυχίες και τους προβληματισμούς μου.

Περιεχόμενα

Περίληψη

Η τοπολογική ανάλυση δεδομένων αποτελεί ένα ανερχόμενο πεδίο της υπολογιστικής τοπολογίας, το οποίο παρέχει θεωρητικά θεμελιωμένες τεχνικές και υπολογιστικά εργαλεία για την ποσοτικοποίηση του σχήματος των δεδομένων. Στην παρούσα εργασία θα περιγράψουμε το θεωρητικό υπόβαθρο και τα εργαλεία της τοπολογικής ανάλυσης δεδομένων και θα προτείνουμε ένα τοπολογικό μοντέλο για τη βαθμονόμηση πιστοληπτικής ικανότητας. Σε αντίθεση με αρκετά μοντέλα μηχανικής εκμάθησης που χρησιμοποιούνται στην αξιολόγηση πιστωτικού κινδύνου, το συγκεκριμένο μοντέλο, καθώς παρέχει τη δυνατότητα οπτικοποίησης των δεδομένων, δε στερείται διαφάνειας και ερμηνευσιμότητας.

Λέξεις- κλειδιά: Τοπολογική Ανάλυση Δεδομένων, Credit Scoring, Ball Mapper, Εμμένουσα ομολογία

Abstract

Topological data analysis is a growing field of computational topology that provides theoretically established techniques and computational tools for the quantification of the shape of data. In this paper we are going to describe the theoretical background and the tools of topological data analysis and we are going to suggest a topological model in order to rate creditworthiness. Unlike several machine learning models used in credit risk assessment, this specific model, does not lack transparency and interpretability as it allows the visualization of data.

Keywords: Topological Data Analysis, Credit Scoring, Ball Mapper, Persistent Homology

Κατάλογος εικόνων

Κατάλογος πινάκων

Εισαγωγή

Η τοπολογία είναι ο κλάδος των μαθηματικών που ασχολείται με τη μελέτη σχημάτων. Σε αντίθεση όμως με τη γεωμετρία ενδιαφέρεται για τον τρόπο με τον οποίο συνδέεται εσωτερικά ένα σχήμα και όχι με το πως ακριβώς μοιάζει (Zomorodian, Topology for Computing 2005). Θεωρώντας ότι ένα σύνολο δεδομένων προέρχεται από έναν συνήθως υψηλής διάστασης άγνωστο χώρο είναι φυσικό να σκεφτούμε πως μέσα από την προσέγγιση του σχήματός του μπορούμε να εξάγουμε χρήσιμες πληροφορίες για τα δεδομένα.

Η τοπολογική ανάλυση δεδομένων (Topological Data Analysis-TDA) είναι ένα αναπτυσσόμενο πεδίο της υπολογιστικής τοπολογίας που παρέχει τα εργαλεία για τη μελέτη και ποσοτικοποίηση του σχήματος των δεδομένων. Σε γενικές γραμμές, τα κυριότερα εργαλεία που χρησιμοποιεί είναι η εμμένουσα ομολογία και ο αλγόριθμος Mapper (Michel 2015). Και οι δύο μεθοδολογίες βασίζονται στη δημιουργία του νεύρου ενός καλύμματος για την προσέγγιση του άγνωστου χώρου από τον οποίο προέρχονται τα δεδομένα. Η εμμένουσα ομολογία εξετάζει το σχήμα του χώρου μέσα από μία κλίμακα αποσκοπώντας στον διαχωρισμό των εγγενών τοπολογικών χαρακτηριστικών του από αυτά που αποτελούν «τοπολογικό θόρυβο». Ο αλγόριθμος Mapper είναι μία μέθοδος οπτικοποίησης της τοπολογικής δομής του χώρου, η οποία χρησιμοποιεί μία συνάρτηση φίλτρου για τη δημιουργία ενός καλύμματος και του νεύρου του (Frédéric Chazal 2021).

Το credit scoring ή credit rating ασχολείται με την αξιολόγηση του πιστωτικού κινδύνου, η οποία καταρχάς αντιμετωπίζεται ως ένα πρόβλημα δυαδικής ταξινόμησης. Αξιοποιώντας διάφορα χαρακτηριστικά, όπως η ηλικία, ο μισθός, ο σκοπός δανεισμού κλπ. αν πρόκειται για ιδιώτες δανειολήπτες λιανικής τραπεζικής ή χρηματοοικονομικούς δείκτες αν πρόκειται για επιχειρήσεις, βαθμολογεί τους δανειζόμενους, επιχειρώντας έτσι την ιεράρχηση των αιτούντων ανάλογα με την πιστοληπτική τους ικανότητα και κατ' επέκταση τον διαχωρισμό τους σε δύο κατηγορίες: «καλούς» και «κακούς» δανειζόμενους (Siddiqi 2017). Η ανάγκη για την δημιουργία ενός μοντέλου με υψηλή προβλεπτική ικανότητα είναι εμφανής. Παρόλα αυτά, αρκετά αποδοτικά αλλά πολύπλοκα μοντέλα αντιμετωπίζονται με δυσπιστία καθώς στερούνται διαφάνειας και ερμηνευσιμότητας. Επομένως, η κατασκευή ενός καλού μοντέλου credit scoring δεν έχει κάποια προφανή λύση (Lean Yu 2008).

Ο σκοπός της εργασίας είναι η μελέτη μιας εναλλακτικής μεθόδου κατηγοριοποίησης και ιεράρχησης των δανειοληπτών μέσω της κατασκευής του νεύρου ενός κατάλληλου καλύμματος του υποκείμενου τοπολογικού χώρου των επεξηγηματικών μεταβλητών, δηλαδή των χαρακτηριστικών τους. Η μέθοδος αυτή, χρησιμοποιεί τον αλγόριθμο Ball Mapper για τη δημιουργία ενός γραφήματος που παρέχει πληροφορίες σχετικά με τη θέση των δανειζόμενων στο χώρο. Η βασική ιδέα είναι ότι άτομα με παρόμοια χαρακτηριστικά (άρα άτομα που βρίσκονται στην ίδια γειτονιά) αναμένεται να έχουν παρόμοια συμπεριφορά. Επομένως, αν οι επεξηγηματικές μεταβλητές έχουν καλή προβλεπτική ικανότητα μπορούμε να εκτιμήσουμε την πιθανότητα αθέτησης για έναν καινούριο δανειζόμενο εντοπίζοντας κάποια «κατάλληλη» γειτονιά στην οποία ανήκει. Από τη στιγμή που εντοπίστηκε αυτή η κατάλληλη γειτονιά, εκτιμούμε την

πιθανότητα αθέτησης του δανειολήπτη ανάλογα με το ποσοστό των «κακών» που ζουν σε αυτή τη γειτονιά. Επιπλέον, τόσο η απλότητα της μεθόδου όσο και η δυνατότητα οπτικοποίησης των δεδομένων την καθιστούν ιδιαίτερα κατανοητή και εύκολα ερμηνεύσιμη. Εκτός από αυτή τη μέθοδο, θα εξετάσουμε και μια λεπτομερέστερη παραλλαγή της για τη δημιουργία καλύμματος του χώρου δειγματοληψίας και στη συνέχεια για την εκτίμηση των πιθανοτήτων αθέτησης των δανειοληπτών.

Το πρώτο κεφάλαιο περιλαμβάνει κάποιες βασικές τοπολογικές έννοιες ενώ στο δεύτερο κεφάλαιο παρουσιάζονται τα simplicial complexes, η έννοια του νεύρου ενός καλύμματος και το θεώρημα του νεύρου, το οποίο διαδραματίζει κεντρικό ρόλο στην τοπολογική ανάλυση δεδομένων. Στη συνέχεια, εισάγεται η έννοια της simplicial ομολογίας και στο τέταρτο κεφάλαιο η έννοια της εμμένουσας ομολογίας. Στο πέμπτο κεφάλαιο, παρουσιάζονται συνοπτικά οι αλγόριθμοι Mapper, οι οποίοι επιτρέπουν μια ενδιαφέρουσα οπτικοποίηση των δεδομένων και περιγράφεται ιδιαίτερα ο αλγόριθμος Ball Mapper τον οποίο και θα μελετήσουμε, θα αναλύσουμε και θα χρησιμοποιήσουμε εκτενώς στην εφαρμογή του επόμενου κεφαλαίου. Το έκτο κεφάλαιο περιλαμβάνει μία σύντομη αναφορά στο credit scoring και εστιάζει στην παρουσίαση των τοπολογικών μοντέλων που θα χρησιμοποιηθούν για την εκτίμηση πιθανοτήτων αθέτησης, την αναλυτική περιγραφή της εφαρμογής τους στο Australian Credit Dataset και την παρουσίαση των αποτελεσμάτων που προέκυψαν. Στο τελευταίο κεφάλαιο ολοκληρώνουμε με μία σύνοψη των αποτελεσμάτων, τη διατύπωση κάποιων προβληματισμών και συμπερασμάτων καθώς και σκέψεις για περαιτέρω έρευνα.

Κεφάλαιο 1: Βασικές Τοπολογικές Έννοιες

Η τοπολογία είναι ο κλάδος των μαθηματικών που ασχολείται με τη μελέτη σχημάτων με έναν πιο ελαστικό τρόπο σε σχέση με τη γεωμετρία. Μπορούμε να φανταζόμαστε ένα σχήμα σα να είναι φτιαγμένο από πλαστελίνη, το οποίο μπορούμε να τεντώσουμε ή να συρρικνώσουμε χωρίς να μεταβάλλουμε τα ουσιαστικά χαρακτηριστικά του, δηλαδή αυτά που το διακρίνουν από άλλα σχήματα. Δεν μπορούμε όμως να το κόψουμε ή να το κολλήσουμε. Με άλλα λόγια, η τοπολογία μελετά τον τρόπο με τον οποίο συνδέεται ένα σχήμα. Στο κεφάλαιο αυτό θα αναφερθούν ορισμένοι βασικοί ορισμοί της τοπολογίας που κρίνονται αναγκαίοι για την καλύτερη κατανόηση της τοπολογικής ανάλυσης δεδομένων.
Το παρόν κεφάλαιο βασίζεται στις πηγές (Zomorodian, TopologyforComputing 2005)**,** (Munkres 1984)**,** (TamalKrishnaDey 2022)**,** (Hatcher 2001)**.**

1.1 Τοπολογία

Ορισμός 1.1.1: Τοπολογικός Χώρος

Έστω X ένα σύνολο. Μία **τοπολογία** του X είναι μία οικογένεια $\boldsymbol{T}$**,** αποτελούμενη από υποσύνολα του X που ικανοποιεί τις ακόλουθες ιδιότητες:

1. Το X και το κενό σύνολο $\emptyset$ ανήκουν στην $\boldsymbol{T}$
2. Η τομή πεπερασμένης οικογένειας στοιχείων της $\boldsymbol{T}$ είναι στοιχείο της $\boldsymbol{T}$
3. Η ένωση αυθαίρετης οικογένειας στοιχείων της $\boldsymbol{T}$ είναι στοιχείο της $\boldsymbol{T}$

Το ζεύγος $(X, \boldsymbol{T})$ λέγεται **τοπολογικός χώρος.** Τα στοιχεία της $\boldsymbol{T}$ λέγονται **ανοικτά σύνολα** ως προς την $\boldsymbol{T}$.

Ορισμός 1.1.2: Κλειστό σύνολο

Έστω $(X, \boldsymbol{T})$ τοπολογικός χώρος και $A \subset X$. Το Α λέγεται **κλειστό** εάν το συμπλήρωμά του, $X \backslash A$ είναι ανοιχτό (δηλαδή εάν $X \backslash A \in \boldsymbol{T}$).

Ορισμός 1.1.3: Εσωτερικό συνόλου

Έστω $(X, \boldsymbol{T})$ τοπολογικός χώρος και $A \subset X$. Το **εσωτερικό** A^o του A είναι η ένωση όλων των ανοικτών συνόλων που περιέχονται στο A.
Δηλαδή, $A^o = \cup \{G \subset X : G \text{ ανοικτό και } G \in A\}$. Το A^o είναι ανοικτό και μάλιστα είναι το μεγαλύτερο ανοικτό σύνολο που περιέχεται στο A.

Ορισμός 1.1.4: Κλειστότητα συνόλου

Έστω $(X, \boldsymbol{T})$ τοπολογικός χώρος και $A \subset X$. Η **κλειστότητα** $\bar{A}$ του A είναι η τομή όλων των κλειστών συνόλων που περιέχουν το A. Δηλαδή,

$\bar{A} = \cap \{F \subset X : F$ κλειστό και $A \subset F\}$. Το $\bar{A}$ είναι το μικρότερο κλειστό σύνολο που περιέχει το A.

Ορισμός 1.1.5: Υπόχωρος τοπολογικού χώρου

Έστω $(X, \boldsymbol{T})$ τοπολογικός χώρος και $A \subset X$. Η οικογένεια $\boldsymbol{T}_A = \{G \cap A : G \in \boldsymbol{T}\}$
είναι μία τοπολογία του A, η **σχετική ή επαγόμενη τοπολογία** του A ως προς $\boldsymbol{T}$. Ο τοπολογικός χώρος $(A, \boldsymbol{T}_A)$ ονομάζεται **υπόχωρος** του A.

Ορισμός 1.1.6: Κάλυμμα και συμπαγής τοπολογικός χώρος

Ένα ανοικτό (κλειστό) **κάλυμμα** ενός τοπολογικού χώρου $(X, \boldsymbol{T})$ είναι μία συλλογή C ανοικτών (κλειστών) συνόλων τέτοια ώστε $X = \cup_{\boldsymbol{c} \in \boldsymbol{C}} \boldsymbol{c}$. Ο τοπολογικός χώρος $(X, \boldsymbol{T})$ καλείται **συμπαγής** αν κάθε ανοικτό κάλυμμα του C έχει πεπερασμένο υποκάλυμμα, δηλαδή, υπάρχει $C' \subset C$ τέτοιο ώστε $X = \cup_{c' \in \boldsymbol{C'}} c'$ και C' πεπερασμένο.

Ορισμός 1.1.7: Τοπολογία πηλίκο

Έστω Y ένα (αυθαίρετο) σύνολο, X ένας τοπολογικός χώρος και $\varphi : X \to Y$ μία συνάρτηση επί του Y. Η **τοπολογία πηλίκο** του Y ως προς την φ είναι η οικογένεια
$\boldsymbol{T}\varphi = \{G \subset Y : \varphi^{-1}(G)\ ανοικτό\ στον\ X\}$
Ο τοπολογικός χώρος $(Y, \boldsymbol{T}_\varphi)$ είναι ο **χώρος πηλίκο ως προς την $\boldsymbol{\varphi}$**.

Ορισμός 1.1.8: Μετρικός χώρος

Έστω X ένα αυθαίρετο σύνολο. Μία συνάρτηση $\rho : X \times X \to \mathbb{R}$ λέγεται **μετρική** στο X αν ικανοποιεί τις παρακάτω ιδιότητες :

i. $\rho(x, y) \geq 0$ για κάθε $x, y \in \boldsymbol{X}$ και $\rho(x, y) = 0$ αν-ν $x = y$
ii. $\rho(x, y) = \rho(y, x)$ για κάθε $x, y \in \boldsymbol{X}$
iii. $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$ για κάθε $x, y, z \in \boldsymbol{X}$

Το ζεύγος (X, ρ) καλείται **μετρικός χώρος**.

Ορισμός 1.1.9: Ανοικτή σφαίρα

Έστω (X, ρ) μετρικός χώρος, $x \in X$ και $\varepsilon > 0$. **Ανοικτή σφαίρα** κέντρου x και ακτίνας ε (ως προς τη μετρική ρ) είναι το σύνολο των στοιχείων του X που η απόσταση τους από το x είναι γνησίως μικρότερη του ε. Δηλαδή, $B(x, \varepsilon) = \{y \in X : \rho(x, y) < \varepsilon\}$.

Ορισμός 1.1.10: Ανοικτό σύνολο μετρικού χώρου

Ένα υποσύνολο G ενός μετρικού χώρου (X, ρ) καλείται **ανοικτό** εάν για κάθε $x \in G$ υπάρχει $\varepsilon > 0$ ώστε $B(x, \varepsilon) \subset G$.

Ορισμός 1.1.11: Τοπολογία μετρικού χώρου

Έστω ένας μετρικός χώρος (X, ρ). Η οικογένεια όλων των ανοικτών υποσυνόλων του X ορίζουν τοπολογία επί του X.

1.2 Απεικονίσεις, ομοιομορφισμοί και ομοτοπίες

Ορισμός 1.2.1: Συνεχής συνάρτηση

Μία συνάρτηση f από έναν τοπολογικό χώρο X σε έναν τοπολογικό χώρο Y είναι **συνεχής** αν για κάθε ανοικτό υποσύνολο $G \subset X$, το $f^{-1}(G)$ είναι ανοικτό.

Ορισμός 1.2.2: Ομοιομορφισμός και ομοιομορφικοί τοπολογικοί χώροι

Έστω X, Y τοπολογικοί χώροι. Μία συνάρτηση $f: X \to Y$ είναι **ομοιομορφισμός** αν είναι 1-1, επί, συνεχής και η αντίστροφη συνάρτηση $f^{-1}: Y \to X$ είναι συνεχής. Αν υπάρχει ομοιομορφισμός $f: X \to Y$ τότε οι χώροι X και Y λέγονται **ομοιομορφικοί** και γράφουμε $X \sim Y$.

Παράδειγμα 1.2: Ένας κύκλος και ένα τετράγωνο είναι ομοιομορφικοί χώροι, αφού μπορούμε με συνεχή τρόπο να απεικονίσουμε το ένα σχήμα στο άλλο. Από την άλλη κανένα από αυτά τα σχήματα δεν είναι ομοιομορφικό με ένα ευθύγραμμο τμήμα, αφού κάτι τέτοιο απαιτεί να κόψουμε (τον κύκλο ή το τετράγωνο) ή να κολλήσουμε (το ευθύγραμμο τμήμα), δηλαδή να καταστρέψουμε τη συνέχεια. Πιο συγκεκριμένα, είναι γνωστό από τη γενική τοπολογία ότι η συνεχής εικόνα ενός συνεκτικού συνόλου είναι συνεκτικό. Αν λοιπόν είχαμε ένα ομοιομορφισμό $f: S^1 \to [a, b]$ τότε επιλέγοντας ένα σημείο s στον S^1 που να μην απεικονίζεται ούτε στο a ούτε στο b θα είχαμε ότι και η συνάρτηση $f_s: S^1 - \{s\} \to [a, f(s)) \cup (f(s), b]$ θα ήταν επίσης ομοιομορφισμός μεταξύ των αντίστοιχων χώρων, οι οποίοι φυσικά είναι εφοδιασμένοι με τη σχετική τοπολογία που κληρονομούν από τους αντίστοιχους Ευκλείδειους χώρους στους οποίους ζουν. Όμως, το $S^1 - \{s\}$ εξακολουθεί να είναι συνεκτικό ενώ το $[a, f(s)) \cup (f(s), b]$ είναι προφανώς μη συνεκτικό (αφού τα $[a, f(s))$ και $(f(s), b]$ είναι ξένα και ανοικτά στη σχετική τοπολογία). Αυτό όμως θα σήμαινε ότι η συνεχής f_s απεικονίζει ένα συνεκτικό σε ένα μη συνεκτικό, το οποίο είναι άτοπο.

Ορισμός 1.2.3: Ισοτοπία

Η **ισοτοπία** μεταξύ δύο τοπολογικών χώρων $X \subseteq \mathbb{R}^d$ και $Y \subseteq \mathbb{R}^d$ είναι μία συνεχής απεικόνιση $\xi: X \times [0,1] \to \mathbb{R}^d$ τέτοια ώστε $\xi(X, 0) = X$, $\xi(X, 1) = Y$ και για κάθε $t \in [0,1]$, η $\xi(\cdot, t)$ να είναι ομοιομορφισμός ανάμεσα στον X και την εικόνα του $\{\xi(x, t): x \in X\}$.

Ορισμός 1.2.4 : Ομοτοπία

Έστω δύο απεικονίσεις $g: X \to Y$ και $h: X \to Y$. Μία **ομοτοπία** είναι μία συνεχής απεικόνιση $H: X \times [0,1] \to Y$ τέτοια ώστε $H(\cdot, 0) = g$ και $H(\cdot, 1) = h$. Δύο απεικονίσεις λέγονται **ομοτοπικές** αν υπάρχει ομοτοπία που να τις συνδέει.

Ορισμός 1.2.5: Ομοτοπικά ισοδύναμοι τοπολογικοί χώροι

Δύο τοπολογικοί χώροι X και Y λέγονται **ομοτοπικά ισοδύναμοι** εάν υπάρχουν απεικονίσεις $f: X \to Y$ και $g: Y \to X$ ώστε η σύνθεση $g \circ f$ να είναι ομοτοπική με την

ταυτοτική απεικόνιση $\iota_X: X \to X$ και η $f \circ g$ να είναι ομοτοπική με την ταυτοτική απεικόνιση $\iota_Y: Y \to Y$.

Ορισμός 1.2.6: Συσταλτός χώρος

Ένας χώρος είναι **συσταλτός** αν είναι ομοτοπικά ισοδύναμος με σημείο.

Για παράδειγμα, μία μπάλα στον ευκλείδειο χώρο είναι συσταλτό σύνολο. Πράγματι, εάν $B^n = \{x \in R^n: |x| \leq 1\}$ και $c: B^n \to B^n$ η σταθερή απεικόνιση με $c(x) = 0\ \forall x \in B^n$ τότε η $H: B^n \times [0,1] \to B^n$ με $H(x,t) = (1-t)x$ είναι ομοτοπία μεταξύ της $c: B^n \to B^n$ και της ταυτοτικής $id: B^n \to B^n$.

Πιο γενικά, όλα τα κυρτά υποσύνολα του ευκλείδειου χώρου είναι συσταλτά.

Ορισμός 1.2.7: Τοπολογικά αναλλοίωτες

Μία **τοπολογικά αναλλοίωτη** είναι μια απεικόνιση f που αποδίδει το ίδιο αντικείμενο σε ομοιομορφικούς χώρους, δηλαδή

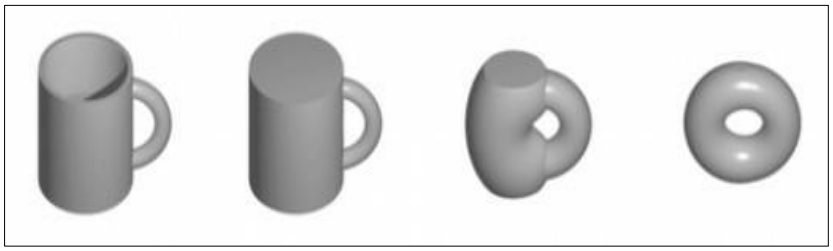
$$X \sim Y \Rightarrow f(X) = f(Y)$$

Σημείωση 1.2.1: Οι τοπολογικά αναλλοίωτες είναι χρήσιμες μόνο μέσω αντιθετοαντιστροφής. Δηλαδή, αν δύο χώροι έχουν διαφορετικές αναλλοίωτες δεν μπορούν να είναι ομοιομορφικοί.

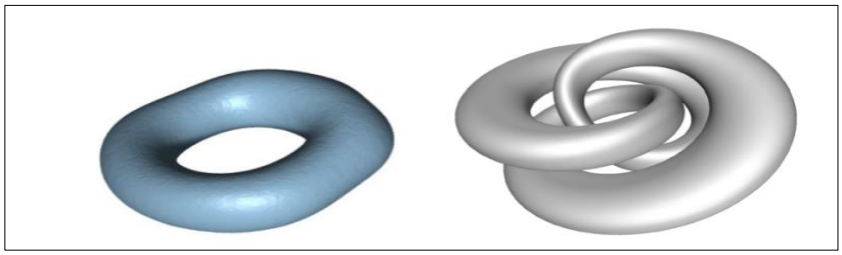
Παρατήρηση 1.2.1: Ο ομοιομορφισμός συνδέει τοπολογικούς χώρους ενώ η ομοτοπία συνδέει απεικονίσεις. Επιπλέον, οι ομοιομορφικοί χώροι έχουν την ίδια διάσταση, κάτι που δεν είναι απαραίτητο για τους ομοτοπικά ισοδύναμους (π.χ. μία κλειστή μπάλα και ένα σημείο). Η ομοτοπία είναι ασθενέστερη του ομοιομορφισμού και η ισοτοπία είναι ισχυρότερη. Επομένως, αν δύο χώροι είναι ομοιομορφικοί θα είναι και ομοτοπικά ισοδύναμοι.

Σημείωση 1.2.2: Η ομοτοπία διατηρεί κάποια μορφή συνδεσιμότητας (κάποια τοπολογικά χαρακτηριστικά- αναλλοίωτες) όπως συνεκτικές συνιστώσες, κύκλους και κενά. Για το λόγο αυτό ένας δίσκος είναι ομοτοπικά ισοδύναμος με ένα σημείο, αφού μπορούμε σταδιακά να τον «συρρικνώσουμε». Προφανώς, δεν είναι ομοιομορφικά.

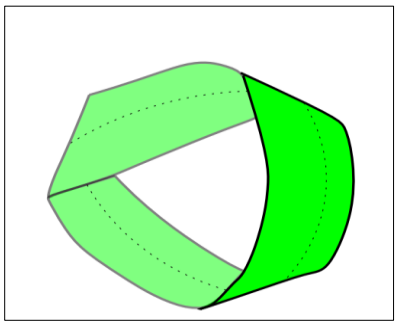
Σημείωση 1.2.3: Η ομοτοπία είναι μια συνεχής μονοπαραμετρική οικογένεια συνεχών συναρτήσεων (δε χρειάζεται να είναι καν αμφιμονοσήμαντη). Η ισοτοπία από την άλλη είναι μια συνεχής μονοπαραμετρική οικογένεια από ομοιομορφισμούς.



Εικόνα 1: Μετασχηματισμός μίας κούπας σε τόρο . Ο τόρος και η κούπα είναι ισοτοπικά ισοδύναμα. (Για τον ορισμό του τόρου βλέπετε https://en.wikipedia.org/wiki/Torus) (https://math.stackexchange.com/questions/2258389/how-can-a-mug-and-a-torus-be-equivalent-if-the-mug-is-chiral)



Εικόνα2: Ο τόρος και ο «δεμένος» τόρος είναι ομοιομορφικοί, όχι όμως και ισοτοπικά ισοδύναμοι. (Tamal Krishna Dey 2022)



Εικόνα 3: Η ταινία του Möbius είναι ομοτοπικά ισοδύναμη με τον κύκλο που παρουσιάζεται με διακεκομμένες γραμμές στο κέντρο της. (Tamal Krishna Dey 2022)

Κεφάλαιο 2: Simplicial Complexes

Στην περίπτωση της τοπολογικής ανάλυσης δεδομένων το μόνο που έχουμε στη διάθεσή μας είναι ένα σύνολο σημείων, έστω D, το οποίο έχει ληφθεί από έναν άγνωστο χώρο X. Σκοπός μας είναι να χρησιμοποιήσουμε το σύνολο δεδομένων για να μελετήσουμε την τοπολογία του **X**. Όμως, εν γένει, ένα σύνολο σημείων δε φέρει καμία μη-τετριμμένη τοπολογία (Michel 2015). Η κεντρική ιδέα του TDA είναι να προσεγγίσουμε τον άγνωστο χώρο **X** χρησιμοποιώντας ένα συνεχές αντικείμενο K, το οποίο «χτίζουμε» πάνω στα δεδομένα, και στη συνέχεια να υπολογίσουμε τις τοπολογικά αναλλοίωτες ιδιότητές του προκειμένου να μελετήσουμε τα αναλλοίωτα χαρακτηριστικά του X. Τα συνεχή αντικείμενα που θα χρησιμοποιήσουμε είναι τα simplicial complexes, τα οποία χάρη στην απλότητά τους είναι ιδιαίτερα δεδομένα στην τοπολογική ανάλυση δεδομένων. Το κεφάλαιο έχει βασιστεί στις πηγές (Munkres 1984) (Zomorodian, TopologicalDataAnalysis 2012), (Zomorodian, TopologyforComputing 2005), (TamalKrishnaDey 2022), (Frédéric Chazal 2021) και (Carlsson 2009)

2.1 Γεωμετρικά και αφηρημένα simplicial complexes

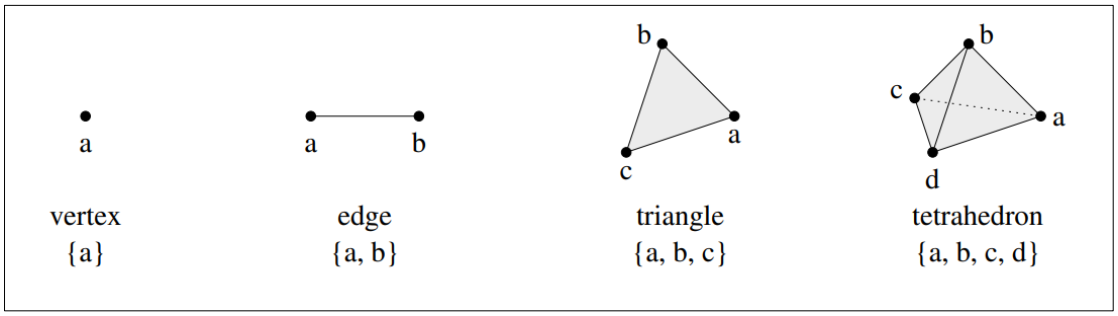
Ορισμός 2.1.1: Simplex

Για $k \geq 0$, **ένα k-simplex** σ στον ευκλείδειο χώρο $\mathbb{R}^m$ είναι η κυρτή θήκη (δηλαδή το μικρότερο κυρτό σύνολο) ενός συνόλου $P = \{v_0, v_1, \dots, v_k\}$ που αποτελείται από $k+1$ γεωμετρικά ανεξάρτητα σημεία στον $\mathbb{R}^m$. Δηλαδή, το σ αποτελείται από όλα τα σημεία x του $\mathbb{R}^m$ τέτοια ώστε:

$$x = \sum_{i=0}^{k} t_i v_i \ , \quad \sum_{i=0}^{k} t_i = 1, \qquad t_i \geq 0 \ \forall i$$

Οι αριθμοί t_i καθορίζονται μονοσήμαντα από το x και καλούνται βαρυκεντρικές συντεταγμένες.

Συγκεκριμένα, ένα 0-simplex είναι ένα σημείο, ένα 1-simplex ένα ευθύγραμμο τμήμα, ένα 2-simplex ένα τρίγωνο με το εσωτερικό του και ένα 3-simplex ένα τετράεδρο (με το εσωτερικό του). Ένα k-simplex έχει διάσταση k. Τα στοιχεία του P καλούνται **κορυφές** του σ.



Εικόνα 4: k-simplex, 0 ≤ k ≤ 3 (Zomorodian, Topology for Computing 2005)

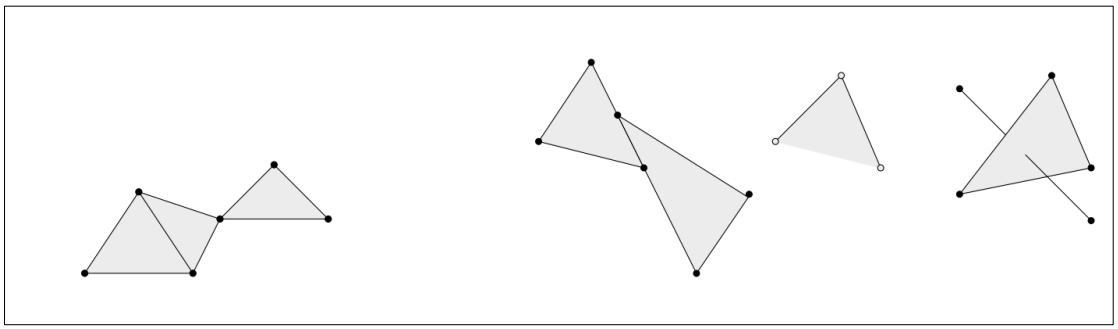
Παρατήρηση 2.1.1: Ένα simplex είναι συσταλτό.

Ορισμός 2.1.2: Γεωμετρικό simplicial complex

Ένα **γεωμετρικό simplicial complex** K είναι μία συλλογή από simplices στον $\mathbb{R}^m$ τέτοια ώστε:

- Το K να περιέχει κάθε όψη από κάθε simplex του K
- Για κάθε δύο simplices $\sigma, \tau \in K$ η τομή τους είναι κενή ή κοινή όψη και των δύο

Η διάσταση του K είναι η μέγιστη διάσταση των simplices του.



Εικόνα 5: Συλλογές από simplices. Το αριστερό σχήμα είναι γεωμετρικό simplicial complex, κάτι που δεν ισχύει για τα υπόλοιπα τρία. (Zomorodian, Topology for Computing 2005)

Ορισμός 2.1.3: Αφηρημένο simplicial complex

Έστω ένα σύνολο V. Μία συλλογή K από μη κενά υποσύνολα του V είναι ένα **αφηρημένο simplicial complex** εάν για κάθε στοιχείο $\sigma \in K$ όλα τα μη κενά υποσύνολά του είναι επίσης στοιχεία του K. Κάθε στοιχείο σ με $|\sigma| = k + 1$ λέγεται k-simplex . Κάθε υποσύνολο σ' του σ με $|\sigma'| = k' + 1$ λέγεται k΄-όψη του σ ή απλά όψη

του σ. Η διάσταση μίας όψης είναι ο πληθάριθμός της μείον ένα και η διάσταση ενός αφηρημένου simplicial complex είναι η μεγαλύτερη διάσταση των όψεών του.

Σύνδεση αφηρημένων και γεωμετρικών simplicial complexes:

Ένα γεωμετρικό simplicial complex K στον $\mathbb{R}^{\mathrm{m}}$ καλείται η **γεωμετρική πραγματοποίηση** ενός αφηρημένου simplicial complex K' αν και μόνο αν υπάρχει 1-1 συνάρτηση $e : V(K') \to \mathbb{R}^m$ η οποία αντιστοιχεί κάθε k-simplex $[v_0, v_1, \dots, v_k]$ του K' σε ένα k-simplex του K που είναι το κυρτό περίβλημα των σημείων $e(v_0), e(v_1), \dots, e(v_k)$.

Ορισμός 2.1.4: Υποκείμενος χώρος ενός simplicial complex

Έστω K ένα (γεωμετρικό) simplicial complex στον $\mathbb{R}^{\mathrm{m}}$ και |K| η ένωση όλων των simplices του. Εφοδιάζοντας κάθε simplex με τη φυσιολογική τοπολογία που κληρονομεί από τον $\mathbb{R}^{\mathrm{m}}$, ορίζουμε τοπολογία επί του |K| ως εξής: κάθε υποσύνολο $A \subset |K|$ είναι κλειστό αν-ν το $A \cap \sigma$ είναι κλειστό στο σ, για κάθε $\sigma \in K$. Ο χώρος $|K|$ καλείται **υποκείμενος χώρος** του K.

Επομένως, ένα simplicial complex K μπορεί να εμβαπτισθεί στον ευκλείδειο χώρο ως η ένωση των (γεωμετρικών) simplices του, με τρόπο ώστε οι τομές τους να είναι είτε κενές είτε κοινή όψη τους. Διαφαίνεται, λοιπόν, η διττή φύση των simplicial complexes αφού μπορούν να αντιμετωπιστούν τόσο ως συνδυαστικά αντικείμενα όσο και ως τοπολογικοί χώροι. Η δεύτερη ιδιότητά τους δίνει τη δυνατότητα εξαγωγής τοπολογικών πληροφοριών ενώ η πρώτη καθιστά ευκολότερους τους υπολογισμούς.

Ορισμός 2.1.5: Subcomplex και p-σκελετός

Έστω K ένα simplicial complex και L ένα υποσύνολο του K που περιέχει όλες τις όψεις των στοιχείων του. Τότε το L είναι simplicial complex και καλείται **subcomplex** του K. Ένα subcomplex του K το οποίο αποτελείται από όλα τα simplices διάστασης έως και p ονομάζεται $\boldsymbol{p}$**-σκελετός** του K. Συμβολίζεται με K^p. Τα στοιχεία του συνόλου K^0 είναι οι κορυφές του K.

Ορισμός 2.1.6: Simplicial απεικονίσεις

Έστω K_1, K_2 δύο simplicial complexes. Μία απεικόνιση $f : K_1 \to K_2$ ονομάζεται **simplicial** εάν για κάθε simplex $\{v_0, \dots, v_k\} \in K_1$ έχουμε ένα simplex $\{f(v_0), \dots, f(v_k)\} \in K_2$. Μια simplicial απεικόνιση καλείται απεικόνιση κορυφών εάν το πεδίο ορισμού και το πεδίο τιμών είναι τα σύνολα κορυφών $V(K_1)$ και $V(K_2)$ αντίστοιχα.

Παρατήρηση 2.1.2: Κάθε simplicial απεικόνιση μπορεί να συσχετιστεί με μία απεικόνιση κορυφών. Ωστόσο, μία απεικόνιση κορυφών $f: V(K_1) \to V(K_2)$ δεν επεκτείνεται απαραίτητα σε μια απεικόνιση από το K_1 στο K_2.

2.2 Νεύρα, Čech και Vietoris-Rips Complex

Σε αυτό το σημείο τίθενται δύο ερωτήματα: Πρώτον, πώς μπορούμε να κατασκευάσουμε ένα simplicial complex πάνω σε ένα σύνολο δεδομένων; Και δεύτερον, πώς γνωρίζουμε ότι αυτό αποτελεί μια καλή προσέγγιση του χώρου από τον οποίο προέρχονται τα δεδομένα; Η απάντηση στο πρώτο ερώτημα είναι ότι μελετούμε τον χώρο τοπικά, κατασκευάζοντας ένα κάλυμμά του. Η ένωση των συνόλων του καλύμματος είναι η προσέγγιση του άγνωστου χώρου. Ένα κάλυμμα ενός τοπολογικού χώρου ορίζει ένα simplicial complex που καλείται το νεύρο του. Βέβαια, η δημιουργία ενός καλύμματος που θα συλλαμβάνει (κάποια από) τα τοπολογικά χαρακτηριστικά του άγνωστου χώρου είναι ένα ζήτημα που θα αναλυθεί παρακάτω. Ας υποθέσουμε, προς το παρόν, ότι έχουμε κατασκευάσει ένα τέτοιο κάλυμμα. Την απάντηση στο δεύτερο ερώτημα, δίνει το Θεώρημα του Νεύρου, το οποίο παρέχει τις συνθήκες υπό τις οποίες ένας χώρος και το νεύρο ενός καλύμματός του είναι ομοτοπικά ισοδύναμα.

Ορισμός: Νεύρο

Δοθείσης μίας πεπερασμένης συλλογής συνόλων $\mathcal{U} = \{U_a\}, a \in A$, ορίζουμε το **νεύρο** του $\mathcal{U}$ να είναι το simplicial complex $N(\mathcal{U})$ το οποίο έχει ως σύνολο κορυφών το σύνολο δεικτών A και ένα υποσύνολο $\{a_0, a_1, \dots, a_k\} \subseteq A$ ορίζει ένα k-simplex του $N(\mathcal{U})$ αν και μόνο αν $U_{a_o} \cap U_{a_1} \cap \dots \cap U_{a_k} \neq \emptyset$.

Η ένωση των συνόλων ενός καλύμματος του υποκείμενου τοπολογικού χώρου αποτελεί την προσέγγισή μας για τον άγνωστο χώρο. Εάν το κάλυμμα είναι «καλό», το νεύρο του καλύμματος αιχμαλωτίζει την τοπολογία του άγνωστου χώρου από τον οποίο προέρχονται τα δεδομένα (Zomorodian, Topological Data Analysis 2012).
Το παρακάτω θεώρημα συνδέει –υπό κάποιες προϋποθέσεις– την τοπολογία του νεύρου ενός καλύμματος με την τοπολογία της ένωσης των συνόλων του καλύμματος.

Το Θεώρημα του Νεύρου: Έστω $\mathcal{U} = \{U_a\}, a \in A$ ένα (πεπερασμένο) κάλυμμα ενός τοπολογικού χώρου X τέτοιο ώστε οποιαδήποτε τομή $\bigcap_{i=0}^{k} U_{a_i}$ να είναι είτε κενή είναι συσταλτή. Τότε, ο X και το νεύρο $N(\mathcal{U})$ είναι ομοτοπικά ισοδύναμα.

Για τα παρακάτω θα θεωρήσουμε ότι ο άγνωστος χώρος από τον οποίο λάβαμε τα δεδομένα είναι ένας μετρικός χώρος.

Δοθέντος ενός πεπερασμένου υποσυνόλου P ενός μετρικού χώρου (M, ρ) μπορούμε να κατασκευάσουμε ένα αφηρημένο simplicial complex με κορυφές στο P χρησιμοποιώντας την έννοια του νεύρου.

Ορισμός 2.2.2: Čech complex

Έστω (M, ρ) μετρικός χώρος, P ένα πεπερασμένο υποσύνολο του M και ένας πραγματικός αριθμός $r > 0$. Το **Čech complex**, $\check{C}ech_r(P)$, είναι το νεύρο του συνόλου $\{B(p_i, r)\}$, όπου $B(p_i, r) = \{x \in M \mid d(p_i, x) \leq r\}$ είναι η κλειστή μπάλα με κέντρο p_i και ακτίνα r.

Στην περίπτωση που το M είναι ο ευκλείδειος χώρος οι μπάλες που χρησιμοποιούνται για την κατασκευή του Čech complex είναι κυρτά σύνολα και συνεπώς οι τομές τους είναι συσταλτές. Επομένως, το Čech complex είναι ομοτοπικά ισοδύναμο με τον M.

Το Čech complex δεν χρησιμοποιείται στην πράξη εξαιτίας της υψηλής υπολογιστικής πολυπλοκότητάς του. Αντίθετα, το Vietoris-Rips complex είναι ευρέως διαδεδομένο στην τοπολογική ανάλυση δεδομένων λόγω της ευκολίας κατασκευής του.

Ορισμός 2.2.3: Vietoris-Rips complex

Έστω (P, ρ) ένας πεπερασμένος μετρικός χώρος και $r > 0$. Το **Vietoris-Rips complex**, $VR_r(P)$, είναι το αφηρημένο simplicial complex το οποίο περιλαμβάνει ένα simplex σ αν και μόνο αν $\rho(x, y) \leq 2r$ για κάθε ζεύγος κορυφών του σ.

Ο 1-σκελετός του $VR_r(P)$ (δηλαδή το σύνολο κορυφών και ακμών) καθορίζει όλα τα simplices του.

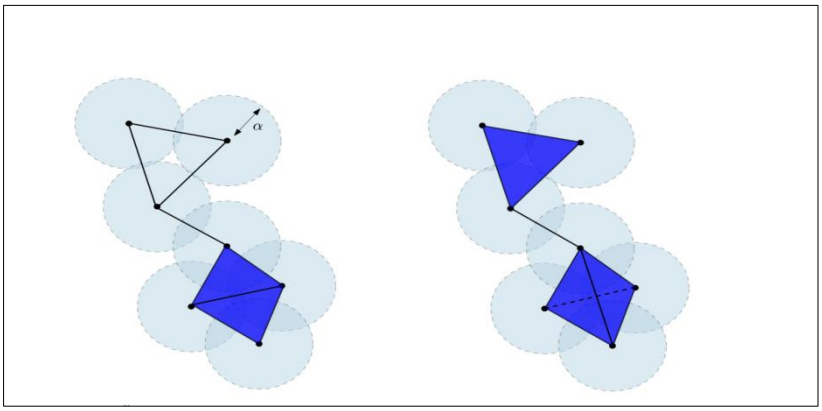
Παρατήρηση 2.2.1: Έστω (M, ρ) ένας μετρικός χώρος ο οποίος ικανοποιεί την εξής ιδιότητα: για κάθε πραγματικό αριθμό $r > 0$ και δύο οποιαδήποτε σημεία $x, y \in M$ με $d(x, y) \leq 2r$ οι μπάλες $B(x, r)$ και $B(y, r)$ έχουν μη κενή τομή. Αν P είναι ένα πεπερασμένο υποσύνολο αυτού του μετρικού χώρου τότε τα $VR_r(P)$ και $\check{C}ech_r(P)$ έχουν τον ίδιο 1-σκελετό. Στην περίπτωση που ο M είναι ευκλείδειος χώρος, προφανώς ικανοποιείται η ανωτέρω συνθήκη, και επομένως το Vietoris-Rips και το Čech complex έχουν τον ίδιο 1-σκελετό (για το ίδιο r).

Πρόταση 2.2: Έστω P ένα πεπερασμένο υποσύνολο ενός μετρικού χώρου (M, ρ). Τότε

$$\check{C}ech_r(P) \subseteq VR_r(P) \subseteq \check{C}ech_{2r}(P)$$

Απόδειξη: Έστω $[v_o, v_1, \dots, v_k] \in \check{C}echr(P)$. Τότε $\bigcap_{i=0}^{k} B(v_i, r) \neq \emptyset$ και επομένως $\rho(v_i, v_j) \leq 2r \ \forall (i, j)$ με $0 \leq i, j \leq k$.
Άρα $[v_o, v_1, \dots, v_k] \in VR_r(P)$, δηλαδή $\check{C}ech_r(P) \subseteq VR_r(P)$.
Έστω τώρα ένα simplex $[v_o, v_1, \dots, v_k] \in VR_r(P)$. Εξ' ορισμού θα ισχύει ότι $\rho(v_i, v_0) \leq 2r$ για κάθε $v_i, i = 0,1, \dots, k$ και επομένως $\bigcap_{i=0}^{k} B(v_i, 2r) \supset \{v_0\} \neq \emptyset$. Άρα, $[v_o, v_1, \dots, v_k]$ είναι simplex του $\check{C}ech_{2r}(P)$, δηλαδή $VR_r(P) \subseteq \check{C}ech_{2r}(P)$.

Παρατήρηση 2.2.2: Παρόλο που το VR_r complex δεν είναι πάντα ομοτοπικά ισοδύναμο με το $\check{C}ech_r$ complex η παραπάνω πρόταση μας επιτρέπει να το δούμε ως μία προσέγγιση του $\check{C}ech_r$.



Εικόνα 6:Čech$_\alpha$ και VR$_\alpha$ complexes. Όπως φαίνεται στο σχήμα, τα δύο complexes έχουν τον ίδιο 1-σκελετό. Όμως, το Čech$_\alpha$ έχει διάσταση 2 ενώ το VR$_\alpha$ διάσταση 3. (Frédéric Chazal 2021)

Κεφάλαιο 3: Simplicial Ομολογία

Ο βασικός μας στόχος είναι να εξάγουμε πληροφορίες για το σχήμα του χώρου προέλευσης των δεδομένων. Αυτό θα γίνει μέσα από τον υπολογισμό τοπολογικά αναλλοίωτων χαρακτηριστικών. Η ομολογία χρησιμοποιεί ελεύθερες αβελιανές ομάδες για την περιγραφή της τοπολογίας ενός χώρου, μέσω της οποίας μπορούμε να υπολογίσουμε αναλλοίωτες του, όπως συνεκτικές συνιστώσες, κύκλους και κενά (Zomorodian, Topology for Computing 2005). Στην πράξη, ο υπολογισμός της ομολογίας ενός χώρου δεν είναι πάντα εύκολος. Αυτός είναι ένας από τους λόγους για τους οποίους χρησιμοποιούμε simplicial complexes για την προσέγγιση του τοπολογικού χώρου (Nina Otter 2017). Επομένως, θα περιοριστούμε μόνο στην simplicial ομολογία, δηλαδή στον υπολογισμό ομάδων ομολογίας για simplicial complexes. Στο πρώτο μέρος του κεφαλαίου θα αναφερθούν κάποιες βασικές αλγεβρικές έννοιες και στη συνέχεια οι ομάδες ομολογίας.

3.1 Αλγεβρικές έννοιες

Η συγκεκριμένη ενότητα βασίζεται στις πηγές (Tamal Krishna Dey 2022), (Munkres 1984).

Ορισμός 3.1.1: Ομάδα

Ένα σύνολο G εφοδιασμένο με μία πράξη «+» καλείται **ομάδα** και συμβολίζεται $(G, +)$ εάν ικανοποιεί τις ακόλουθες ιδιότητες:

i. Για κάθε $a, b \in G$ το $a + b \in G$
ii. $(a + b) + c = a + (b + c)$
iii. Υπάρχει ουδέτερο στοιχείο $0 \in G$ τέτοιο ώστε $a + 0 = 0 + a = a$
iv. Για κάθε $a \in G$ υπάρχει αντίθετο στοιχείο $-a \in G$ τέτοιο ώστε $a + (-a) = (-a) + a = 0$

Εάν ικανοποιείται η αντιμεταθετική ιδιότητα, δηλαδή $a + b = b + a \ \forall a, b \in G$ η ομάδα καλείται **αβελιανή**.

Ορισμός 3.1.2: Υποομάδα

Έστω $(G, +)$ μία ομάδα. Ένα υποσύνολο $H \subset G$ καλείται **υποομάδα** αν το $(H, +)$ είναι ομάδα.

Ορισμός 3.1.3 : Ελεύθερη αβελιανή ομάδα

Μια αβελιανή ομάδα $(G, +)$ ονομάζεται **ελεύθερη** εάν υπάρχει υποσύνολο $B \subseteq G$ ώστε κάθε στοιχείο του G να γράφεται με μοναδικό τρόπο ως πεπερασμένο άθροισμα στοιχείων του B και των αντιθέτων τους. Το σύνολο B καλείται **βάση** του G και το πλήθος των στοιχείων του ονομάζεται **βαθμός**. Στην περίπτωση που κάθε στοιχείο του

G γράφεται ως πεπερασμένο άθροισμα στοιχείων του B αλλά όχι απαραίτητα μονοσήμαντα το B λέγεται **γεννήτορας** του G.

Ορισμός 3.1.4: Κλάσεις και πηλίκο

Έστω μία ομάδα $(G, +)$. Για κάθε υποομάδα της H ≤ G και για κάθε στοιχείο $\alpha \in G$ η αριστερή κλάση είναι $aH = \{a + b | b \in H\}$ και η δεξιά κλάση $Ha = \{b + a | b \in H\}$. Στις αβελιανές ομάδες η αριστερή κλάση ταυτίζεται με τη δεξιά. Το σύνολο όλων των αριστερών (δεξιών) κλάσεων $modH$ το λέμε αριστερό (δεξί) σύνολο πηλίκο της G δια την H. Εάν η G είναι αβελιανή, η **ομάδα πηλίκο** G δια την υποομάδα H συμβολίζεται με G/H και είναι $G/H = \{aH | a \in G\}$. Η ομάδα πηλίκο κληρονομεί την πρόσθεση από την G ως εξής: $aH + bH = (a + b)H, \forall a, b \in G$.

Ορισμός 3.1.5: Ομομορφισμός ομάδων

Μια απεικόνιση $h: G \to H$ μεταξύ δύο ομάδων $(G, +)$ και $(H, *)$ καλείται **ομομορφισμός** εάν $h(a + b) = h(a) * h(b)\ \forall a, b \in G$. Εάν, επιπλέον, η απεικόνιση είναι 1-1 και επί τότε καλείται **ισομορφισμός**. Δύο ομάδες G και H που συνδέονται με κάποιον ισομορφισμό καλούνται **ισομορφικές** και συμβολίζουμε $G \cong H$.

Ορισμός 3.1.6: Πυρήνας και εικόνα ομομορφισμού

Έστω ένας ομομορφισμός $h: G \to H$ μεταξύ των ομάδων $(G, +)$ και $(H, *)$. Ο πυρήνας είναι η υποομάδα της G που ορίζεται ως $ker(h) = \{g \in G | h(g) = 0\}$ και η εικόνα είναι η υποομάδα της H που ορίζεται ως $Im(h) = \{a \in H | \exists g \in G\ με\ h(g) = a\}$.

Ορισμός 3.1.7: Δακτύλιος

Ένα σύνολο R εφοδιασμένο με δύο πράξεις, πρόσθεση «+» και πολλαπλασιασμό «·», ονομάζεται **δακτύλιος** αν ισχύουν τα παρακάτω:

i. Το $(R, +)$ είναι αβελιανή ομάδα
ii. Ισχύει η προσεταιριστική ιδιότητα του πολλαπλασιασμού
iii. Ισχύει η επιμεριστική ιδιότητα του πολλαπλασιασμού ως προς την πρόσθεση
iv. Υπάρχει ουδέτερο στοιχείο για τον πολλαπλασιασμό

Από τον ορισμό των αβελιανών ομάδων έπεται ότι η πρόσθεση είναι αντιμεταθετική. Αν ισχύει και η αντιμεταθετικότητα ως προς τον πολλαπλασιασμό, ο δακτύλιος ονομάζεται **αντιμεταθετικός**. Εάν επιπλέον κάθε μη μηδενικό στοιχείο είναι αντιστρέψιμο το $(R, +, \cdot)$ είναι **σώμα**.

Ορισμός 3.1.8: Πρότυπο

Έστω αντιμεταθετικός δακτύλιος R με ουδέτερο στοιχείο πολλαπλασιασμού το 1. Ένα **R-πρότυπο** είναι μία αβελιανή ομάδα M εφοδιασμένη με μία απεικόνιση $R \times M \to M$ η οποία ικανοποιεί τις ακόλουθες ιδιότητες για κάθε $r, r' \in R$ και $x, y \in M$:

i. $r \cdot (x + y) = r \cdot x + r \cdot y$
ii. $(r + r') \cdot x = r \cdot x + r' \cdot x$
iii. $1 \cdot x = x$
iv. $(r \cdot r') \cdot x = r \cdot (r' \cdot x)$

Εάν το R είναι σώμα, κάθε μη μηδενικό έχει πολλαπλασιαστικό αντίστροφο και το R-πρότυπο γίνεται διανυσματικός χώρος.

Ορισμός 3.1.9: Διανυσματικός χώρος

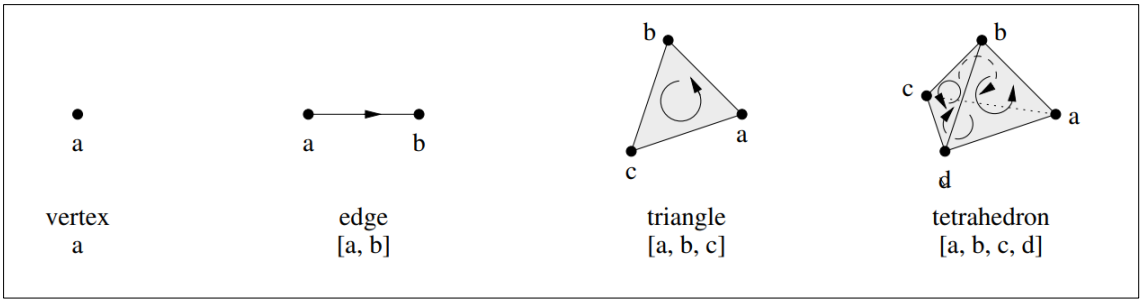
Ένα R-πρότυπο V λέγεται **διανυσματικός χώρος** εάν το R είναι σώμα. Ένα σύνολο στοιχείων $\{g_1, g_2, \dots, g_k\}$ λέγεται ότι παράγουν τον διανυσματικό χώρο V αν κάθε στοιχείο $a \in V$ μπορεί να γραφεί ως $a = a_1 g_1 + \cdots + a_k g_k$ με $a_1, \dots, a_k \in R$. Αν αυτός ο τρόπος γραφής είναι μοναδικός το σύνολο $\{g_1, g_2, \dots, g_k\}$ καλείται **βάση** του διανυσματικού χώρου. Όλες οι βάσεις ενός διανυσματικού χώρου έχουν το ίδιο πλήθος στοιχείων που ονομάζεται **διάσταση** του V.

3.2 Αλυσίδες, κύκλοι και σύνορα

Για την παρούσα και την επόμενη ενότητα έχουν χρησιμοποιηθεί οι πηγές (Tamal Krishna Dey 2022), (Zomorodian, Topology for Computing 2005), (Munkres 1984).

Ορισμός3.2.1: Προσανατολισμένο simplex

Έστω σ ένα simplex. Ορίζουμε δύο διαφορετικές διατάξεις των κορυφών του να είναι ισοδύναμες αν διαφέρουν κατά μία άρτια μετάθεση. Επομένως, στην περίπτωση $dim(\sigma) > 0$ οι διατάξεις των κορυφών του χωρίζονται σε δύο κλάσεις ισοδυναμίας. Κάθε μία από αυτές τις κλάσεις καλείται προσανατολισμός του σ. Αν το σ είναι 0-simplex υπάρχει μόνο ένας προσανατολισμός. Ένα προσανατολισμένο simplex σ είναι ένα simplex σ μαζί με κάποιον προσανατολισμό του σ.



Εικόνα 7: Προσανατολισμένο k-simplex, 0≤k≤3 (Zomorodian, Topology for Computing 2005)

Ορισμός 3.2.2: p-αλυσίδα

Έστω K ένα simplicial k-complex και n_p ο αριθμός των p-simplices του K, $p < k$. Μία **p-αλυσίδα** c στο K είναι ένα τυπικό (πεπερασμένο) άθροισμα από p-simplices πολλαπλασιασμένων με κάποιους συντελεστές, δηλαδή,

$c = \sum_{i=1}^{n_p} a_i \sigma_i$, όπου σ_i είναι τα p-simplices και a_i οι συντελεστές.

Μπορούμε να προσθέσουμε p-αλυσίδες $c = \sum_{i=1}^{n_p} a_i \sigma_i$ και $c' = \sum_{i=1}^{n_p} a_i{}' \sigma_i$ για να πάρουμε μία άλλη p-αλυσίδα $c + c' = \sum_{i=1}^{n_p} (a_i + a_i{}') \sigma_i$.

Ο **χώρος των p-αλυσίδων** ενός simplicial complex K συμβολίζεται με $C_p(K)$. Στην περίπτωση που οι συντελεστές ανήκουν σε κάποιον δακτύλιο R ο C_p αποτελεί R-πρότυπο, ενώ αν οι συντελεστές προέρχονται από κάποιο σώμα ο C_p είναι διανυσματικός χώρος.

Μια από τις συνηθέστερες και (βολικότερες) επιλογές για το σύνολο των συντελεστών είναι το $\mathbb{Z}_2$. Η **ομάδα των p-αλυσίδων με συντελεστές στο** $\mathbb{Z}_2$ έχει ως ουδέτερο στοιχείο την αλυσίδα $0 = \sum_{i=1}^{n_p} 0\sigma_i$ και το αντίθετο στοιχείο οποιασδήποτε αλυσίδας c είναι το ίδιο το c, δηλαδή, $c + c = 0$ και αφού το $\mathbb{Z}_2$ είναι σώμα ο C_p είναι διανυσματικός χώρος.

Ορισμός 3.2.3: Συνοριακός τελεστής

Αν $\sigma = [v_0, v_1, \dots, v_p]$ είναι ένα p-simplex με $p > 0$ ορίζουμε

$$\partial_p(\sigma) = \sum_{i=0}^{p} (-1)^i [v_0, \dots \widehat{v_i}, \dots v_p]$$

όπου ο συμβολισμός $\widehat{v_i}$ δηλώνει ότι η i-οστή κορυφή παραλείπεται.

Σημείωση 3.2: Επί της ουσίας, ο συνοριακός τελεστής στέλνει ένα p-simplex σε μία (p-1)-αλυσίδα που έχει μη μηδενικούς συντελεστές μόνο για τις (p-1)-όψεις του σ.

Παρατήρηση 3.2.1: Στην περίπτωση που οι συντελεστές προέρχονται από το $\mathbb{Z}_2$ ισχύει ότι

$\partial_p(\sigma) = \sum_{i=0}^{p} [v_0, \dots \widehat{v_i}, \dots v_p]$, αφού $1 = (-1)\ mod\ 2$

Η παραπάνω απεικόνιση μπορεί να επεκταθεί σε έναν ομομορφισμό $\partial_p : C_p \to C_{p-1}$ μεταξύ των ομάδων C_p και C_{p-1}, ο οποίος καλείται **συνοριακός τελεστής** και απεικονίζει μία p-αλυσίδα $c = \sum_{i=1}^{n_p} a_i \sigma_i \in C_p$ σε μία (p-1)-αλυσίδα ως εξής:

$$\partial_p(c) = \sum_{i=1}^{n_p} a_i (\partial_p(\sigma_i))$$

Παρατήρηση 3.2.2: Το σύνορο μίας κορυφής είναι κενό. Επιπλέον, η ομάδα C_{-1} περιέχει μόνο ένα στοιχείο, το 0. Τέλος, αν το K είναι k-complex τότε το C_p είναι 0 για $p > k$.

Πρόταση 3.2: Για $p > 0$ και οποιαδήποτε p-αλυσίδα c ισχύει $\partial_{p-1} \circ \partial_p(c) = 0$

Απόδειξη: Έστω $\sigma = [v_0, v_1, \dots, v_p]$ ένα p-simplex. Έχουμε:

$$\partial_{p-1} \circ \partial_p(\sigma) = \sum_{i=0}^{p} (-1)^i \partial_{p-1}\left[v_0, \dots \widehat{v_i}, \dots v_p\right] = \sum_{j<i} (-1)^i (-1)^j \left[\dots, \widehat{v_j}, \dots, \widehat{v_i}, \dots\right] + \sum_{j>i} (-1)^i (-1)^{j+1} \left[\dots, \widehat{v_i}, \dots, \widehat{v_j}, \dots\right]$$

Οι όροι αυτών των αθροισμάτων διαγράφονται ανά δύο.

Επεκτείνοντας τον συνοριακό τελεστή, αποκτούμε την εξής ακολουθία ομομορφισμών μεταξύ των ομάδων αλυσίδων:

$$0 = C_{k+1} \xrightarrow{\partial_{p+1}} C_k \xrightarrow{\partial_p} C_{k-1} \xrightarrow{\partial_{p-1}} C_{k-2} \dots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} C_{-1} = 0$$

Ορισμός 3.2.4: Κύκλοι και σύνορα

Ο πυρήνας του ομομορφισμού $\partial_p : C_p(K) \to C_{p-1}(K)$ ονομάζεται **ομάδα των p-κύκλων** και συμβολίζεται με $Z_p(K)$. Η εικόνα του ομομορφισμού $\partial_{p+1} : C_{p+1}(K) \to C_p(K)$ ονομάζεται **ομάδα των p-συνόρων** και συμβολίζεται με $B_p(K)$. Και οι δύο είναι υποομάδες της $C_p(K)$. Βάσει της προηγούμενης πρότασης, κάθε σύνορο μιας (p+1)-αλυσίδας είναι p-κύκλος. Επομένως, $B_p(K) \subseteq Z_p(K)$.

3.3 Ομάδες ομολογίας

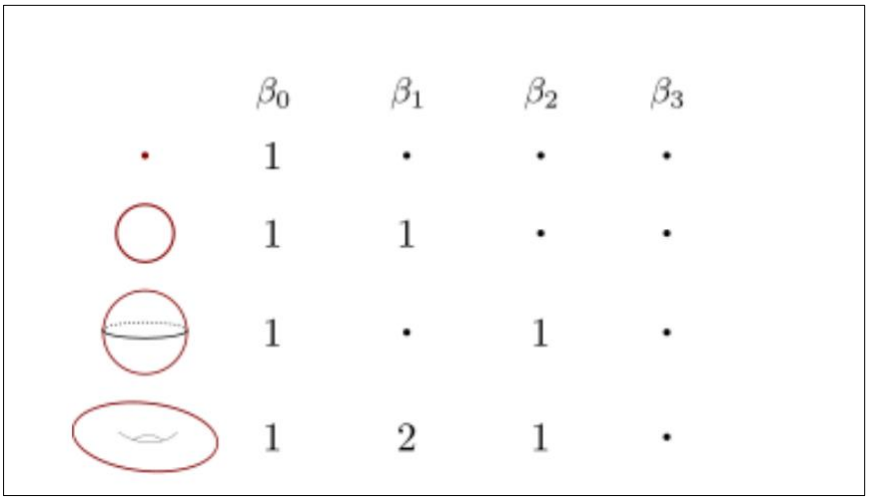
Ορισμός 3.3.1: Ομάδες ομολογίας

Για $p \geq 0$, η **p ομάδα ομολογίας** είναι η ομάδα πηλίκο $H_p(K) = Z_p(K)/B_p(K)$. Ο βαθμός αυτής της ομάδας ονομάζεται p-οστός **αριθμός Betti**, δηλαδή $\beta_p = rank\left(H_p(K)\right) = rank\left(Z_p(K)\right) - rank\left(B_p(K)\right)$

Παρατήρηση 3.3.1: Οι ομάδες ομολογίας διαχωρίζουν τους «πραγματικούς» p-κύκλους από αυτούς που προέκυψαν ως σύνορα αλυσίδων μεγαλύτερης διάστασης. Αυτό επιτυγχάνεται θεωρώντας το πηλίκο της ομάδας των κύκλων δια την ομάδα των συνόρων, το οποίο επιτρέπεται γιατί η ομάδα συνόρων είναι (κανονική) υποομάδα της ομάδας κύκλων.

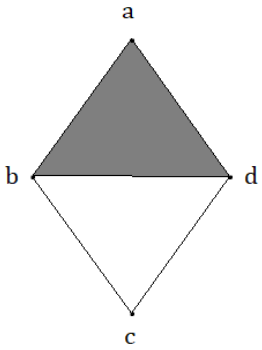
Παρατήρηση 3.3.2: Στην περίπτωση που οι συντελεστές προέρχονται από το $\mathbb{Z}_2$ ο $H_p(K)$ είναι διανυσματικός χώρος και επομένως ο βαθμός του είναι απλά η διάστασή του.

Κάθε στοιχείο του $H_p(K)$ προκύπτει από την πρόσθεση ενός p-κύκλου $c \in Z_p(K)$ με ολόκληρη την ομάδα συνόρων $B_p(K)$. Όλοι οι κύκλοι που προκύπτουν από την πρόσθεση του c με ένα στοιχείο του $B_p(K)$ σχηματίζουν την κλάση ομολογίας του c, $[c]$. Δύο κύκλοι c και c^* στην ίδια κλάση ομολογίας καλούνται ομόλογοι και , προφανώς, $[c] = [c^*]$. Εξ' ορισμού $[c] = [c^*]$ αν και μόνο αν $c \in c^* + B_p(K)$ και αν έχουμε συντελεστές στο $\mathbb{Z}_2$ αυτό σημαίνει ότι $c + c^* \in B_p(K)$.



Εικόνα 8: Αριθμοί Betti για γνωστά σχήματα (Munch 2017)

Παράδειγμα: Θεωρούμε το simplicial complex *K={{a},{b},{c},{d},{a,b},{a,d},{b,d}, {b,c},{c,d},{a,b,d}}* που φαίνεται στο παρακάτω σχήμα. Θα υπολογίσουμε τις ομάδες ομολογίας του και τους αντίστοιχους αριθμούς Betti. (με συντελεστές στο $\mathbb{Z}_2$)



Αρχικά, θα καταγράψουμε κάποιες από τις 0,1,2-αλυσίδες του.

0-αλυσίδες: $\{a\}, \{b\}, \{d\}, \{c\}, \{a\} + \{b\}, \{a\} + \{d\} + \{c\}, \ldots$

1-αλυσίδες: $\{a, b\}, \{a, d\}, \{b, d\}, \{b, c\}, \{c, d\}, \{a, b\} + \{b, d\} + \{a, d\}, \{b, c\} + \{c, d\} + \{b, d\}, \{a, b\} + \{b, c\} + \{c, d\} + \{a, d\}, \ldots$

Παρατηρείστε ότι:

$\partial_1(\{a, b\} + \{b, d\} + \{a, d\}) = \{a\} + \{b\} + \{b\} + \{d\} + \{a\} + \{d\} = 0$

$\partial_1(\{b, c\} + \{c, d\} + \{b, d\}) = \{b\} + \{c\} + \{c\} + \{d\} + \{b\} + \{d\} = 0$

$\partial_1(\{a, b\} + \{b, c\} + \{c, d\} + \{a, d\}) =$
$= \{a\} + \{b\} + \{b\} + \{c\} + \{c\} + \{d\} + \{a\} + \{d\} = 0$

2-αλυσίδες: $\{a, b, d\}$

Επιπλέον, $\partial_2(\{a, b, d\}) = \{a, b\} + \{b, d\} + \{a, d\}$, δηλαδή η συγκεκριμένη 1-αλυσίδα προκύπτει ως σύνορο ενός simplexυψηλότερης διάστασης.

Υπολογισμός της $H_0(K) = Z_0(K) \,/\, B_0(K)$:

$Z_0(K) = ker\partial_0 = span(\{a\}, \{b\}, \{d\}, \{c\}) \cong (\mathbb{Z}_2)^4$

$B_0(K) = im\partial_1 = span(\{a\} + \{b\}, \{a\} + \{d\}, \{b\} + \{d\}, \{b\} + \{c\}, \{c\} + \{d\})$

Όμως, $\{b\} + \{c\} = \{\alpha\} + \{b\} + \{a\} + \{c\}$ και $\{c\} + \{d\} = \{a\} + \{c\} + \{a\} + \{d\}$

επομένως $B_0(K) = span(\{a\} + \{b\}, \{a\} + \{d\}, \{b\} + \{d\}) \cong (\mathbb{Z}_2)^3$

Τελικά, $H_0(K) = (\mathbb{Z}_2)^4 / (\mathbb{Z}_2)^3 = \mathbb{Z}_2$ και $\beta_0 = 1$

Υπολογισμός της $H_1(K) = Z_1(K) \,/\, B_1(K)$:

$Z_1(K) = ker\partial_1 = span(\{b, c\} + \{c, d\} + \{b, d\}, \{a, b\} + \{a, d\} + \{b, d\}) \cong (\mathbb{Z}_2)^2$

$B_1(K) = im\partial_2 = span(\{a, b\} + \{b, d\} + \{c, d\}) \cong \mathbb{Z}_2$

Επομένως, $H_1(K) = (\mathbb{Z}_2)^2 / \mathbb{Z}_2 = \mathbb{Z}_2$ και $\beta_1 = 1$.

Συνεπώς, το συγκεκριμένο simplicial complex έχει μία συνεκτική συνιστώσα και έναν κύκλο, όπως άλλωστε ήταν εμφανές από το σχήμα.

3.4 Επαγόμενη Ομολογία (Ομομορφισμοί επαγόμενοι από simplicial απεικονίσεις)

Οι συνεχείς απεικονίσεις από έναν τοπολογικό χώρο σε έναν άλλο, απεικονίζουν τους κύκλους σε κύκλους και τα σύνορα σε σύνορα. Επομένως, επάγουν μία απεικόνιση ανάμεσα στις ομάδες ομολογίας τους. Η παρούσα ενότητα βασίζεται στο βιβλίο (Tamal Krishna Dey 2022)

Ορισμός 3.4.1: Απεικόνιση αλυσίδων

Έστω μία simplicial απεικόνιση $f: K_1 \to K_2$. Η **απεικόνιση αλυσίδων** $f_\#: C_p(K_1) \to C_p(K_2)$ που αντιστοιχεί στην f ορίζεται ως εξής: Αν $c = \sum a_i \sigma_i$ είναι μία p-αλυσίδα, τότε $f_\#(c) = \sum a_i \tau_i$ όπου

$$\tau_i = \begin{cases} f(\sigma_i), \text{αν } f(\sigma_i) \text{ είναι p} - \text{simplex στο } K_2 \\ 0, \text{διαφορετικά} \end{cases}$$

Πρόταση 3.4: Έστω $f: K_1 \to K_2$ simplicial απεικόνιση και $\partial_p^{K_1}$, $\partial_p^{K_2}$ οι συνοριακοί ομομορφισμοί διάστασης p. Τότε, οι επαγόμενες απεικονίσεις αλυσίδων μετατίθενται με τους συνοριακούς ομομορφισμούς, δηλαδή:

$$f_\# \circ \partial_p^{K_1} = \partial_p^{K_1} \circ f_\#$$

Η παραπάνω πρόταση μπορεί να παρασταθεί μέσω του ακόλουθου μεταθετικού διαγράμματος:

$$\begin{array}{ccc} C_p(K_1) & \xrightarrow{f_\#} & C_p(K_2) \\ \downarrow \partial_p^{K_1} & & \downarrow \partial_p^{K_2} \\ C_{p-1}(K_1) & \xrightarrow{f_\#} & C_{p-1}(K_2) \end{array}$$

Επιπλέον, $f_\#(Z_p(K_1)) \subseteq Z_p(K_2)$ και $f_\#(B_p(K_1)) \subseteq B_p(K_2)$. Επίσης, επειδή $B_p(K_1) \subseteq Z_p(K_1)$ έπεται ότι $f_\#(B_p(K_1)) \subseteq f_\#(Z_p(K_1))$ και, επομένως, η επαγόμενη απεικόνιση του στον χώρο πηλίκο,

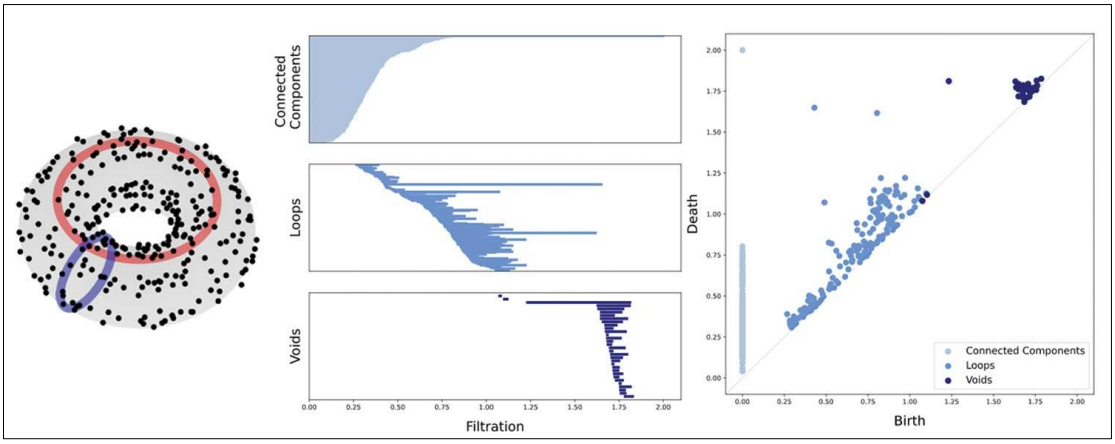
$$f_*(Z_p(K_1)/B_p(K_1)) := f_\#(Z_p(K_1))/f_\#(B_p(K_1))$$

είναι καλώς ορισμένη.

Ορίζουμε, λοιπόν, την επαγόμενη απεικόνιση μεταξύ ομάδων ομολογίας $f_*: H_p(K_1) \to H_p(K_2)$ η οποία απεικονίζει κάθε στοιχείο $c + B_p(K_1)$ της $H_p(K_1)$ σε κάποιο στοιχείο $f_\#(c) + B_p(K_2)$ της $H_p(K_2)$.

Κεφάλαιο 4: Εμμένουσα Ομολογία

Ας υποθέσουμε ότι έχουμε ένα point cloud από έναν τόρο και θέλουμε να μελετήσουμε τα τοπολογικά χαρακτηριστικά του υποκείμενου χώρου. Σύμφωνα με τα προηγούμενα, μπορούμε να τοποθετήσουμε μπάλες γύρω από κάθε σημείο προκειμένου να δημιουργήσουμε ένα κάλυμμα του χώρου προέλευσης των δεδομένων και στη συνέχεια να κατασκευάσουμε το αντίστοιχο Čech ή το Vietoris-Rips complex και να υπολογίσουμε τις ομάδες ομολογίας και τους αντίστοιχους αριθμούς Betti. Όμως, ποια θα πρέπει να είναι η ακτίνα που θα χρησιμοποιήσουμε ώστε να αιχμαλωτίσουμε τα βασικά χαρακτηριστικά του τόρου; Η απάντηση είναι να χρησιμοποιήσουμε όλες τις τιμές για την ακτίνα και για κάθε μία να κατασκευάσουμε το αντίστοιχο simplicial complex, να υπολογίσουμε τις ομάδες ομολογίας του και να εξετάσουμε ποια είναι τα χαρακτηριστικά που επιμένουν (Munch 2017). Αυτό ακριβώς έγινε για τα δεδομένα της παρακάτω εικόνας και αυτό που βλέπουμε είναι ότι υπάρχει μια τρύπα διάστασης 0 (συνεκτική συνιστώσα) , 2 τρύπες διάστασης 1 (που αντιπροσωπεύουν τους δύο κύκλους) και 1 τρύπα διάστασης 2 (που αντιπροσωπεύει το κενό που περικλείει ο τόρος) οι οποίες επιμένουν. Μπορούμε, επομένως, να εξάγουμε πληροφορίες για έναν χώρο εξετάζοντας για πόσο επιμένει ένα χαρακτηριστικό (κλάση ομολογίας) όταν το κοιτάζουμε μέσα από μια αύξουσα ακολουθία από simplicial complexes. Με αυτόν τον τρόπο μπορούμε να διαχωρίσουμε τα «σημαντικά» χαρακτηριστικά από αυτά που αποτελούν απλώς «τοπολογικό θόρυβο» (Michel 2015). Αυτός είναι με λίγα λόγια ο τρόπος με τον οποίο λειτουργεί η εμμένουσα ομολογία. Χρησιμοποιεί μία συνάρτηση που ορίζεται επί ενός τοπολογικού χώρου (simplicial complex) και ποσοτικοποιεί τις αλλαγές στις κλάσεις ομολογίας καθώς τα sublevel σύνολα (subcomplexes) «μεγαλώνουν» όσο αυξάνονται οι τιμές της συνάρτησης. Παρακάτω θα αναφερθούμε μόνο στα simplicial complexes. Δηλαδή, στην περίπτωση στην οποία ασχολούμαστε με μία φωλιασμένη ακολουθία από simplicial complexes, την οποία ονομάζουμε **simplicial διήθηση** (simplicial filtration).



Εικόνα 9: Ένα point cloud από έναν τόρο και το αντίστοιχο barcode και persistence diagram (https://chance.amstat.org/2021/04/topological-data-analysis/)

Κάποια από τα πιο σημαντικά πλεονεκτήματα της εμμένουσας ομολογίας είναι η ευκολία υπολογισμού της (όταν ο άγνωστος χώρος προσεγγίζεται από simplicial complexes) και η ανθεκτικότητά της απέναντι σε μικρές «διαταραχές» ή στην παρουσία θορύβου στα δεδομένα. (Munch 2017), (NinaOtter 2017). Το παρόν κεφάλαιο βασίζεται στις πηγές (Tamal Krishna Dey 2022), (Zomorodian, Topology for Computing 2005).

4.1 Διηθήσεις και εμμένουσα ομολογία

Ορισμός 4.1.1: Διήθηση χώρου (space filtration)

Έστω T ένας τοπολογικός χώρος και $f: T \to \mathbb{R}$ μία πραγματική συνάρτηση. Συμβολίζουμε με $T_\alpha = f^{-1}(-\infty, a]$ το υποσύνολο του T για την τιμή α της συνάρτησης. Προφανώς, για $a \leq b$ έχουμε $T_\alpha \subseteq T_b$. Θεωρούμε τώρα την ακολουθία πραγματικών αριθμών $a_1 \leq a_2 \leq \cdots \leq a_n$, οι οποίες συχνά επιλέγονται να είναι οι κρίσιμες τιμές στις οποίες η ομάδα ομολογίας αλλάζει. Θεωρώντας τα υποσύνολα του T για αυτές τις τιμές της συνάρτησης και $a_0 = -\infty$ με $T_{\alpha_0} = \emptyset$, προκύπτει μία φωλιασμένη ακολουθία υποχώρων του T η οποία δημιουργεί μία **διήθηση**:

$$\mathcal{F}_f: \emptyset = T_{a_0} \hookrightarrow T_{a_1} \hookrightarrow \cdots \hookrightarrow T_{a_n}$$

Επιπλέον, εάν $\iota: T_{a_i} \to T_{a_j}, i \leq j$ είναι η απεικόνιση συμπερίληψης $x \longmapsto x$, προκύπτει ο ομομορφισμός των αντίστοιχων ομάδων ομολογίας:

$h_p^{i,j}: H_p(T_{a_i}) \to H_p(T_{a_j})$ για κάθε $p \geq 0, 0 \leq i \leq j \leq n$.

Επομένως, έχουμε την εξής ακολουθία ομομορφισμών:

$$0 = H_p(T_{a_0}) \to H_p(T_{a_1}) \to \cdots \to H_p(T_{a_n})$$

Η διήθηση μπορεί να εφαρμοστεί και σε point clouds ως εξής:

Έστω P ένα σύνολο σημείων σε έναν μετρικό χώρο (M, d). Ορίζουμε τη συνάρτηση $f: M \to \mathbb{R}$, $x \longmapsto d(x, p)$, όπου $p = arg\min_{q \in P} d(x, q)$. Όλα τα σύνολα $f^{-1}(-\infty, a]$ προκύπτουν ως η ένωση κλειστών μπαλών ακτίνας α και κέντρου στο P. Έχουμε, λοιπόν, ακριβώς τις ίδιες συνθήκες με πριν: ο χώρος T αντικαθίσταται από τον M και τα υποσύνολά του T_α με την ένωση μπαλών που μεγαλώνουν καθώς μεγαλώνει το α.

Θα αναφερθούμε τώρα στο διακριτό ανάλογο της τοπολογικής εμμονής. Για ένα point cloud, η ένωση μπαλών αντικαθίσταται από το νεύρο τους, μέσω της κατασκευής simplicial complex (Čech ή Vietoris-Rips complex). Η φωλιασμένη ακολουθία τοπολογικών χώρων μεταφράζεται σε μία φωλιασμένη ακολουθία από simplicial complexes .

Ορισμός 4.1.2: Simplicial Filtration

Μία διήθηση $\mathscr{F} = \mathscr{F}(K)$ ενός simplicial complex K είναι μία φωλιασμένη ακολουθία από subcomplexes του K:

$$\mathscr{F}: \emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$$

η οποία μπορεί επίσης να γραφεί ως :

$$\mathscr{F}: \emptyset = K_0 \hookrightarrow K_1 \hookrightarrow \cdots \hookrightarrow K_n = K$$

Η διήθηση $\mathscr{F}$ καλείται **simplex-wise** αν $K_i \backslash K_{i-1}$ είναι κενό ή μόνο ένα simplex για κάθε $i \in \{1, 2, \dots, n\}$. (Παρατήρηση: δύο διαδοχικά simplicial complexes ενδέχεται να είναι ίδια.)

Καταλήγουμε στην εξής ακολουθία ομομορφισμών μεταξύ των ομάδων ομολογίας για τα simplicial complexes της διήθησης:

$$H_p\mathscr{F}: 0 = H_p(K_0) \to H_p(K_1) \to \cdots H_p(K_i) \xrightarrow{h_p^{i,j}} H_p\big(K_j\big) \dots \to H_p(K_n) = H_p(K)$$

Ορισμός 4.1.3: Simplex-wise μονότονη συνάρτηση

Έστω K ένα simplicial complex και μία simplex-wise συνάρτηση $f: K \to \mathbb{R}$ επί αυτού. Η συνάρτηση f ονομάζεται simplex-wise μονότονη αν για κάθε $\sigma' \subseteq \sigma$ έχουμε $f(\sigma') \leq f(\sigma)$. Η ιδιότητα αυτή εξασφαλίζει ότι τα σύνολα $f^{-1}(-\infty, a]$ είναι subcomplexes του K για κάθε $a \in \mathbb{R}$. Θεωρώντας $K_i = f^{-1}(-\infty, a_i]$ και $a_0 = -\infty$ προκύπτει η διήθηση:

$$\mathscr{F}: \emptyset = K_0 \hookrightarrow K_1 \hookrightarrow \cdots \hookrightarrow K_n = K$$

Ορισμός 4.1.4: Εμμένων αριθμός Betti (Persistent Betti Number)

Οι p-οστές ομάδες ομολογίας είναι οι εικόνες των ομομορφισμών $H_p^{i,j} = im\big(h_p^{i,j}\big), 0 \leq i \leq j \leq n$. Οι **p-οστοί εμμένοντες αριθμοί Betti** είναι οι διαστάσεις $\beta_p^{i,j} = \dim(H_p^{i,j})$ των αντίστοιχων διανυσματικών χώρων $H_p^{i,j}$ (έχουμε διανυσματικούς χώρους στην περίπτωση που δουλεύουμε με συντελεστές από τον $\mathbb{Z}_2$. Διαφορετικά – και γενικότερα – είναι το rank της ομάδας ομολογίας $H_p^{i,j}$, δηλαδή το μικρότερο πλήθος γεννητόρων της ομάδας)

Σημείωση 4.1: Οι p-οστές ομάδες ομολογίας παρέχουν πληροφορίες για το πότε μία κλάση ομολογίας δημιουργείται και το πότε εξαφανίζεται. Το ζήτημα της γέννησης και του θανάτου μίας κλάσης γίνεται πιο πολύπλοκο γιατί όταν μία νέα κλάση δημιουργείται, μαζί με αυτή δημιουργούνται και άλλες κλάσεις που είναι το άθροισμα της καινούριας με αυτές που ήδη υπήρχαν. Με παρόμοιο τρόπο, όταν μία κλάση παύει να υπάρχει, μαζί με αυτή μπορεί να εξαφανιστούν και άλλες.

Παρατήρηση 4.1.1: Τα μη τετριμμένα στοιχεία της $H_p^{i,j}$ απαρτίζονται από τις κλάσεις που επιβιώνουν από το K_i μέχρι το K_j επομένως $H_p^{i,j} = Z_p(K_i)/(B_p(K_j) \cap Z_p(K_i))$.

Ορισμός 4.1.5: Γέννηση και θάνατος μίας κλάσης ομολογίας

Μία μη-τετριμμένη p-οστή κλάση ομολογίας $\xi \in H_p(K_a)$ **γεννιέται** στο $K_i, i \leq a$ αν $\xi \in H_p^{i,a}$ αλλά $\xi \notin H_p^{i-1,a}$. Παρόμοια, μία μη-τετριμμένη p-οστή κλάση ομολογίας $\xi \in H_p(K_a)$ **πεθαίνει** στο $K_j, j > a$ αν $h_p^{a,j-1}(\xi) \neq 0$ αλλά $h_p^{a,j}(\xi) = 0$.

Παρατήρηση 4.1.2: Έστω $[c] \in H_p(X_{j-1})$ μία p κλάση ομολογίας που πεθαίνει στο X_j. Τότε, έχει γεννηθεί στο X_i αν και μόνο αν υπάρχει ακολουθία $i_1 \leq i_2 \leq \cdots \leq i_k = i$ για κάποιο $k \geq 1$ τέτοια ώστε:

i. η μη τετριμμένη κλάση $[c_{i_l}] \in H_p(X_{j-1})$ γεννιέται στο X_{i_l} για κάθε $l \in \{1, \dots, k\}$.
ii. $[c] = [c_{i_1}] + [c_{i_2}] + \cdots + [c_{i_k}]$

Η παραπάνω παρατήρηση μπορεί να ερμηνευθεί ως εξής. Όταν μία κλάση πεθαίνει, μπορούμε να τη σκεφτόμαστε σαν τη συγχώνευση διάφορων κλάσεων η νεότερη από της οποίες καθορίζει το σημείο της γέννησης.

4.2 Persistence Diagram

Το persistence diagram είναι ένας τρόπος να αναπαραστήσουμε γραφικά την εμμένουσα ομολογία για μία φωλιασμένη ακολουθία από simplicial complexes. Για την κατασκευή του θεωρούμε το επεκταμένο επίπεδο $\overline{\mathbb{R}^2} = \mathbb{R}^2 \cup \{\pm\infty\}$ πάνω στο οποίο θα απεικονίσουμε τη γέννηση και το θάνατο μίας κλάσης σαν ένα σημείο. Οι συντεταγμένες θα καθοριστούν από την persistence pairing function, η οποία ορίζεται παρακάτω και μετρά το πλήθος σημείων που γεννιούνται και πεθαίνουν συγκεκριμένες στιγμές. Αυστηρά θετικές τιμές της συνάρτησης αντιστοιχούν σε πολλαπλότητες του persistence diagram, δηλαδή κλάσεις με κοινές ημερομηνίες γέννησης και θανάτου. Προκειμένου να λάβουμε υπόψη κλάσεις που δεν πεθαίνουν, επεκτείνουμε την ακολουθία από simplicial complexes προσθέτοντας στα δεξιά το K_{n+1} και θεωρώντας $H_p(K_{n+1}) = 0$.

Ορισμός 4.2.1: Persistence Pairing Function

Για $0 < i < j \leq n + 1$ ορίζουμε:

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j})$$

Η πρώτη διαφορά μετρά τον αριθμό των ανεξάρτητων κλάσεων που γεννιούνται στο ή πριν από το K_i και πεθαίνουν μπαίνοντας στο K_j. Η δεύτερη διαφορά μετρά των αριθμό

ανεξάρτητων κλάσεων που γεννιούνται στο ή πριν το K_{i-1} και πεθαίνουν μπαίνοντας στο K_j. Τελικά, η διαφορά τους, $\mu_p^{i,j}$, μετρά τον αριθμό κλάσεων που γεννιούνται στο K_i και πεθαίνουν μπαίνοντας στο K_j. Όταν $j = n+1$, η $\mu_p^{i,n+1}$ μετρά τον αριθμό των κλάσεων που γεννιούνται στο K_i και πεθαίνουν μπαίνοντας στο K_{n+1}, δηλαδή παραμένουν ζωντανές για όλη την αρχική διήθηση. Λέμε ότι αυτές οι κλάσεις δεν πεθαίνουν ποτέ. Για να δώσουμε περισσότερη έμφαση στο γεγονός ότι οι κλάσεις που υπάρχουν στο K_n δεν πεθαίνουν, εξισώνουμε το $n+1$ με ∞ και παίρνουμε $a_{n+1} = a_\infty = \infty$.

Ορισμός 4.2.2: Class Persistence

Για $\mu_p^{i,j} \neq 0$ η επιμονή $Pers([c])$ μίας κλάσης $[c]$ που γεννιέται στο K_i και πεθαίνει μπαίνοντας στο K_j είναι: $Pers([c]) = a_j - a_i$

Όταν $j = n+1$, $Pers([c]) = a_{n+1} - a_i = \infty$

Ορισμός 4.2.3: Persistence diagram

Το **διάγραμμα επιμονής (persistence diagram)** $\mathrm{Dgm_p}(\mathcal{F}_\mathrm{f})$ μίας διήθησης $\mathcal{F}_\mathrm{f}$, επαγόμενης από μία συνάρτηση f , αποτελείται από όλα τα σημεία (a_i, a_j) με μη μηδενική πολλαπλότητα $\mu_p^{i,j}$, $i < j$, στο επεκταμένο επίπεδο $\overline{\mathbb{R}^2} = \mathbb{R}^2 \cup \{\pm\infty\}$. Τα σημεία της διαγωνίου Δ: $\{(x,x)\}$ προστίθενται με άπειρη πολλαπλότητα.

Μία κλάση που γεννιέται στο a_i και δεν πεθαίνει ποτέ αναπαρίσταται ως ένα σημείο $(a_i, a_{n+1}) = (a_i, \infty)$.

Παρατηρήσεις 4.2.1:

1. Κάθε μη διαγώνιο σημείο θα βρίσκεται πάνω από τη διαγώνιο αφού προφανώς ισχύει ότι $a_i < a_j$.
2. Αν μία κλάση έχει επιμονή s τότε το σημείο που την αναπαριστά στο persistence diagram θα απέχει $\frac{s}{\sqrt{2}}$ από τη διαγώνιο Δ
3. Αν m_i είναι η πολλαπλότητα ενός σημείου (a_i, ∞) στο persistence diagram $\mathrm{Dgm_p}(\mathcal{F}_\mathrm{f})$, όπου $\mathcal{F}_\mathrm{f}$ είναι μία διήθηση του $K = K_n$ τότε ο p-οστός αριθμός Betti είναι: $\sum_i m_i = dim\left(H_p(K)\right) = \beta_p$

Barcode:

Το διάγραμμα **barcode** αποτελεί έναν εναλλακτικό τρόπο αναπαράστασης της επιμονής μίας κλάσης. Κάθε ζεύγος γέννησης-θανάτου (a_i, a_j) αναπαρίσταται ως ένα ευθύγραμμο τμήμα $[a_i, a_j)$.

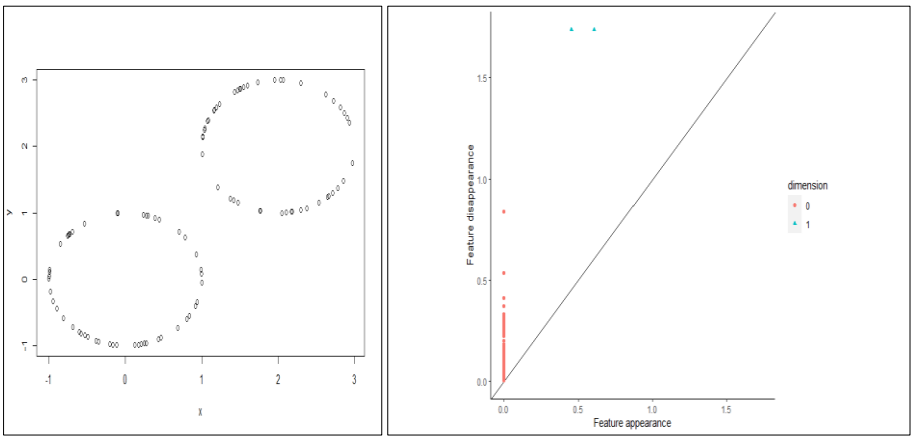
Παρατήρηση 4.2.4: Τα ευθύγραμμα τμήματα $[a_i, a_j)$ που αναπαριστούν τη γέννηση και τον θάνατο μίας κλάσης είναι ανοιχτά δεξιά για να δηλώσουμε ότι η κλάση που πεθαίνει εισερχόμενη στο K_j δεν υπάρχει στο K_j.

Θεώρημα 4.2: Για κάθε ζεύγος δεικτών $0 \leq k \leq l \leq n$ και κάθε p, ο p-οστός εμμένων αριθμός Betti είναι: $\beta_p^{k,l} = \sum_{i \leq k} \sum_{j > l} \mu_p^{i,j}$.

Παρατηρήσεις 4.2.2:

1. Μία κλάση που γεννιέται στο K_iκαι πεθαίνει μπαίνοντας στο K_j, συμμετέχει στον καθορισμό του $\beta_p^{k,l}$ αν και μόνο αν $i \leq k$ και $j > l$.
2. Το $\beta_p^{k,l}$ είναι ο αριθμός των σημείων που βρίσκονται στο πάνω δεξιά τεταρτοκύκλιο της γωνίας με κορυφή το (a_k, a_l). Το τεταρτοκύκλιο είναι κλειστό δεξιά και ανοιχτό κάτω.

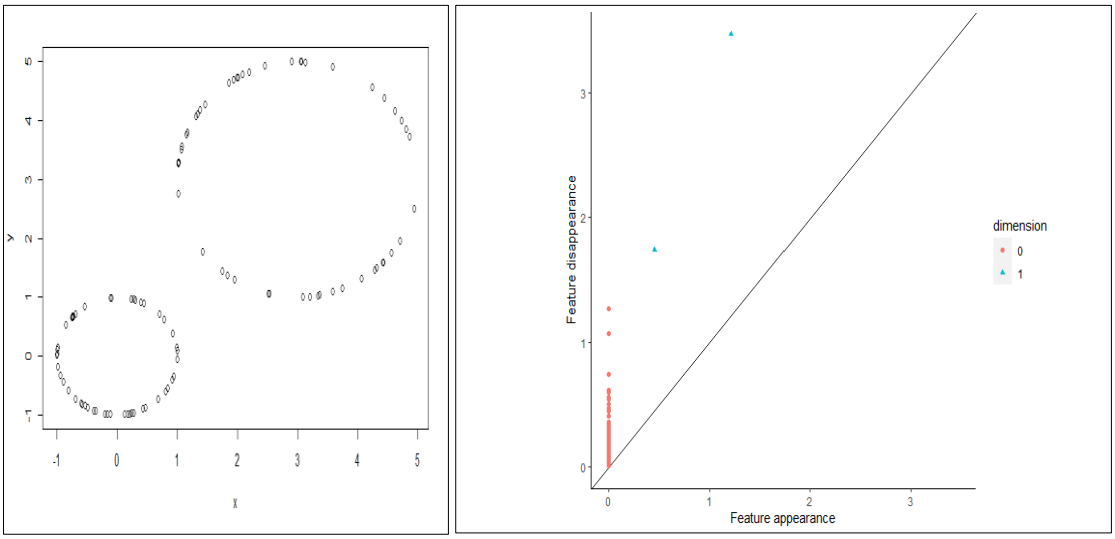
Παράδειγμα: Θα θεωρήσουμε αρχικά ένα σύνολο δεδομένων το οποίο αποτελείται από 50 παρατηρήσεις που έχουν ληφθεί από τον μοναδιαίο κύκλο με κέντρο το σημείο O(0,0) και 50 παρατηρήσεις από έναν μοναδιαίο κύκλο με κέντρο το σημείο (2,2) όπως φαίνεται στην παρακάτω εικόνα και θα κατασκευάσουμε το persistence diagram με χρήση του πακέτου "TDAstats". Αυτό που περιμένουμε να δούμε είναι δύο συνεκτικές συνιστώσες και δύο κύκλους.



Εικόνα 10: Σύνολο δεδομένων από δύο ίσους κύκλους και το αντίστοιχο persistence diagram

Αυτό που βλέπουμε είναι ότι επιμένουν δύο συνεκτικές συνιστώσες (από 0.54 μέχρι 0.84 υπάρχουν 2 συνεκτικές συνιστώσες) και δύο κύκλοι, όπως άλλωστε αναμέναμε. Μάλιστα, οι δύο κύκλοι πεθαίνουν σχεδόν ταυτόχρονα (κοντά στο 1.73) το οποίο σχετίζεται (και) με το γεγονός ότι τα δεδομένα προέρχονται από κύκλους με την ίδια ακτίνα.

Στη συνέχεια, θα θεωρήσουμε το σύνολο δεδομένων που αποτελείται από 50 παρατηρήσεις που προέρχονται από τον μοναδιαίο κύκλο και 50 παρατηρήσεις που έχουν ληφθεί από έναν κύκλο με κέντρο (3,3) και ακτίνα 2, όπως φαίνεται παρακάτω, για το οποίο θα κατασκευάσουμε το persistence diagram.



Εικόνα 11: Σύνολο δεδομένων από δύο κύκλους και το αντίστοιχο persistence diagram

Σε αυτήν την περίπτωση βλέπουμε ότι υπάρχουν δύο κύκλοι που επιμένουν. Ο πρώτος δημιουργείται για ε = 0.50 και ο δεύτερος για ε=1.10. Ο κύκλος που εμφανίζεται πρώτος αντιπροσωπεύει τον μικρό κύκλο του αριστερού σχήματος. Υπάρχει ένα εύρος τιμών του ε για το οποίο οι δύο κύκλοι συνυπάρχουν. Φαίνεται να υπάρχουν τρεις συνεκτικές συνιστώσες που επιμένουν και αυτό εξηγείται από τα «κενά» που παρουσιάζει ο μεγάλος κύκλος. Εάν είχαμε λάβει περισσότερα σημεία, το αποτέλεσμα θα ήταν πιο ακριβές. Επομένως, το πόσο ικανοποιητικά θα ανακτήσουμε την τοπολογία του υποκείμενου χώρου των δεδομένων εξαρτάται και από το πλήθος τους και από τον τρόπο που είναι κατανεμημένα στο χώρο.

4.3 Ευστάθεια των Persistence Diagrams

Απόσταση μεταξύ persistence diagrams

Έστω $Dgm_p(\mathscr{F}_f)$ και $Dgm_p(\mathscr{F}_g)$ δύο persistence diagrams για δύο συναρτήσεις f, g. Θέλουμε να θεωρήσουμε όλες τις 1-1 και επί αντιστοιχίες μεταξύ των σημείων των δύο διαγραμμάτων. Ωστόσο, ενδέχεται να μην έχουν το ίδιο πλήθος μη διαγώνιων σημείων. Θεωρήσαμε, προηγουμένως, ότι τα σημεία της διαγωνίου Δ έχουν άπειρη πολλαπλότητα. Μπορούμε επομένως να «δανειζόμαστε» σημεία από τη διαγώνιο - όταν χρειάζεται - για να ορίσουμε 1-1 και επί συναρτήσεις.

Ορισμός 4.3.1: Απόσταση Bottleneck

Έστω $\Pi = \{\pi: Dgm_p(\mathscr{F}_f) \to Dgm_p(\mathscr{F}_g)\}$ το σύνολο όλων των 1-1 και επί απεικονίσεων μεταξύ των persistence diagrams. Θεωρούμε επιπλέον την απόσταση μεταξύ δύο σημείων $x = (x_1, x_2)$ και $y = (y_1, y_2)$ στην L_∞ - νόρμα, δηλαδή, $||x - y||_\infty = \max\{|x_1 - x_2|, |y_1 - y_2|\}$ και, επιπλέον, $\infty - \infty = 0$. Η **απόσταση bottleneck** μεταξύ των δύο διαγραμμάτων είναι:

$$d_b\left(Dgm_p(\mathscr{F}_f), Dgm_p(\mathscr{F}_g)\right) = inf_{\pi\in\Pi} sup_{x\in Dgm_p(\mathscr{F}_f)} ||x - \pi(x)||_\infty$$

Παρατήρηση 4.3.1: Η d_b είναι μετρική στον χώρο των persistence diagrams. Ισχύει ότι $d_b(X, Y) = 0$ αν-ν $X = Y$. Επιπλέον, $d_b(X, Y) = d_b(Y, X)$ και $d_b(X, Y) \leq d_b(X, Z) + d_b(Z, Y)$.

Θεώρημα 4.3.1: Ευστάθεια για simplicial διηθήσεις

Έστω $f, g: K \to \mathbb{R}$ δύο simplex-wise μονότονες συναρτήσεις και $\mathscr{F}_f, \mathscr{F}_g$ οι αντίστοιχες simplicial διηθήσεις. Τότε, για κάθε $p \geq 0$:

$$d_b\left(Dgm_p(\mathscr{F}_f), Dgm_p(\mathscr{F}_g)\right) \leq ||f - g||_\infty$$

όπου $||f - g||_\infty \coloneqq \sup_{x\in K} |f(x) - g(x)|$

Παρατήρηση 4.3.2:Το παραπάνω θεώρημα μας εξασφαλίζει ότι το persistence diagram ενός συνόλου δεδομένων με θόρυβο είναι «κοντά» στο persistence diagram που θα προέκυπτε από το χωρίς θόρυβο σύνολο δεδομένων. (Munch 2017)

Ορισμός 4.3.2: Απόσταση Wasserstein

Έστω $\Pi = \{\pi: Dgm_p(\mathscr{F}_f) \to Dgm_p(\mathscr{F}_g)\}$ το σύνολο όλων των 1-1 και επί απεικονίσεων μεταξύ των persistence diagrams. Για κάθε $p \geq 0$ η **q-Wasserstein απόσταση** είναι:

$$d_{W,q}\left(Dgm_p(\mathscr{F}_f), Dgm_p(\mathscr{F}_g)\right) = inf_{\pi\in\Pi}\Big[\sum_{x\in Dgm_p(\mathscr{F}_f)} \left(||x - \pi(x)||_q\right)^q\Big]^{\frac{1}{q}}$$

όπου $||x - y||_q = (|x_1 - x_2|^q + |y_1 - y_2|^q)^{\frac{1}{q}}$ για δύο σημεία $x = (x_1, x_2)$, $y = (y_1, y_2)$

Παρατήρηση 4.3.3: Η απόσταση $d_{W,q}$ είναι μετρική στον χώρο των persistence diagrams.

Θεώρημα 4.3.2: Ευστάθεια των persistence diagrams (απόσταση Wasserstein)

Έστω $f, g: K \to \mathbb{R}$ δύο simplex-wise μονότονες συναρτήσεις και $\mathscr{F}_f, \mathscr{F}_g$ οι αντίστοιχες simplicial διηθήσεις. Τότε, για κάθε $p \geq 0$:

$$d_{W,q}\left(Dgm_p(\mathscr{F}_f), Dgm_p(\mathscr{F}_g)\right) \leq ||f - g||_q$$

όπου $||f - g||_q = (\sum_{\sigma \in K} |f(\sigma) - g(\sigma)|^q)^{\frac{1}{q}}$

Κεφάλαιο 5: Ο αλγόριθμος Ball Mapper

Οι αλγόριθμοι Mapper είναι χρήσιμοι αλγόριθμοι που, όπως και τα persistence diagrams, επιχειρούν να παράσχουν μια «καλή τοπολογική περίληψη» των δεδομένων και ταυτόχρονα να οπτικοποιήσουν με κάποιο τρόπο τη συνδεσιμότητά τους αλλά και τα επιμέρους χαρακτηριστικά τους, μέσω ενός κατάλληλου γραφήματος το οποίο διαθέτει ευελιξία ως προς τις επιλογές των παραμέτρων του.

5.1 Ο αλγόριθμος Mapper

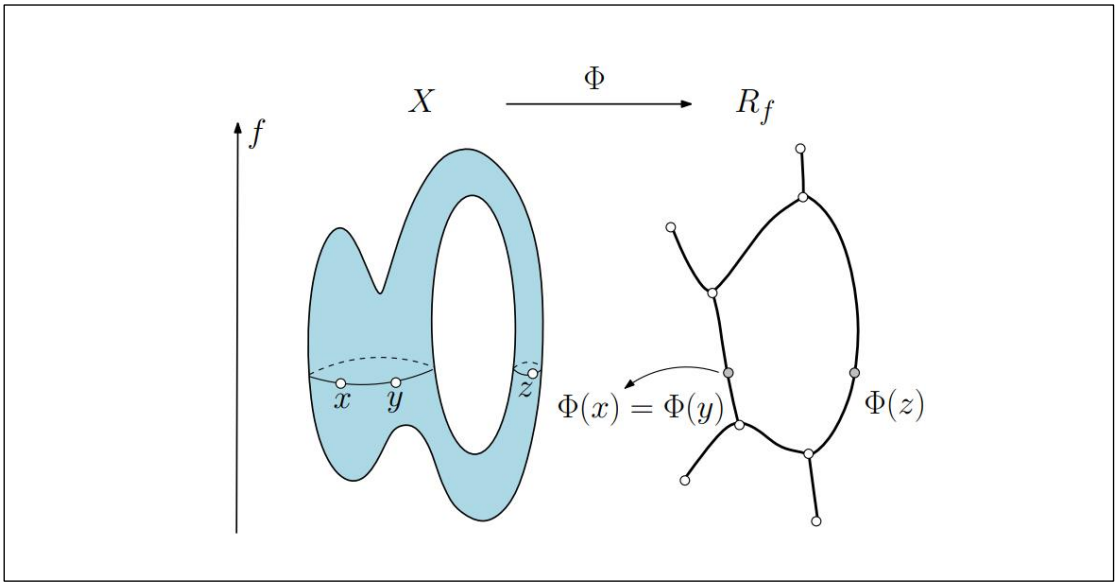
Ο αλγόριθμος Mapper αποτελεί μία γενίκευση της ιδέας ενός γραφήματος Reeb. Τα γραφήματα Reeb παρέχουν μία μονοδιάστατη περίληψη ενός τοπολογικού χώρου μέσα από μία συνάρτηση φίλτρου. Φυσικά, δεν προσφέρουν πληροφορίες για χαρακτηριστικά υψηλότερης διάστασης αλλά είναι εύκολα κατασκευάσιμα και δίνουν τη δυνατότητα οπτικοποίησης του χώρου. Στην περίπτωση του αλγορίθμου Mapper, ο τοπολογικός χώρος αντικαθίσταται από ένα σύνολο δεδομένων και δοθείσης μιας (κατάλληλης) συνάρτησης φίλτρου και ενός καλύμματος της εικόνας της κατασκευάζουμε ένα γράφημα που παρέχει πληροφορίες για το σχήμα των δεδομένων. Ο αλγόριθμος μπορεί να χρησιμοποιηθεί για συσταδοποίηση και επιλογή χαρακτηριστικών. Συγκεκριμένα, μας ενδιαφέρουν σχηματισμοί όπως βρόγχοι και εξάρσεις. Τέτοιοι σχηματισμοί μπορούν να αξιοποιηθούν για να εντοπίσουμε ενδιαφέρουσες συστάδες και να εξετάσουμε ποια είναι τα χαρακτηριστικά που διαφοροποιούν τα δεδομένα στις διάφορες ομάδες (Frédéric Chazal 2021). Παρακάτω θα γίνει μία σύντομη περιγραφή του αλγορίθμου Mapper και των παραμέτρων του, ξεκινώντας φυσικά από τα γραφήματα Reeb. Η ενότητα βασίζεται στις πηγές (P. Y. Lum 2013), (Frédéric Chazal 2021), (Tamal Krishna Dey 2022).

Ορισμός: Reeb graph

Έστω X ένας τοπολογικός χώρος και $f: X \to \mathbb{R}$ μία πραγματική συνεχής συνάρτηση. Ορίζουμε τη σχέση ισοδυναμίας $\sim$ θεωρώντας ότι $x \sim y$ αν και μόνο αν:

i. $f(x) = f(y) = a$
ii. Τα x και y ανήκουν στην ίδια συνεκτική συνιστώσα του συνόλου $f^{-1}(a) = \{v \in X: f(v) = a\}$ (δηλαδή αν το σύνολο $f^{-1}(a)$ μπορεί να γραφεί ως ένωση δύο ξένων συνόλων, έστω U και V, τότε, τα x, y θα ανήκουν σε ένα από τα δύο).

Συμβολίζοντας με $[x]$ την κλάση ισοδυναμίας του $x \in X$, το γράφημα Reeb R_f με συνάρτηση φίλτρου την f είναι ο χώρος πηλίκο $X/\sim$.



Εικόνα 12:**Το γράφημα Reeb με συνάρτηση φίλτρου f(x,y,z)= z.Τα ισοϋψή σημεία που βρίσκονται στην ίδια συνεκτική συνιστώσα του συνόλου $f^{-1}(a) = \{(x, y, z) \in X: f(x, y, z) = a\}$ ανήκουν στην ίδια κλάση ισοδυναμίας (Tamal Krishna Dey 2022)**

Επιθυμούμε να κατασκευάσουμε το διακριτό ανάλογο ενός γραφήματος Reeb για ένα σύνολο δεδομένων Y που έχουμε λάβει από τον χώρο X και έστω η συνάρτηση $f: X \rightarrow \mathbb{R}$, η οποία καλείται **συνάρτηση φίλτρου**. Προφανώς, η αντίστροφη εικόνα ενός σημείου $x \in Y$ είναι εν γένει κενή. Υπολογίζουμε αρχικά το εύρος τιμών της f, έστω I, για τα σημεία του συνόλου Y και στη συνέχεια χωρίζουμε το διάστημα αυτό σε μικρότερα αλληλεπικαλυπτόμενα διαστήματα I_j (συνήθως ίσου μήκους). Το μήκος των διαστημάτων (και το πλήθος τους) και το ποσοστό αλληλοκάλυψης διαδοχικών διαστημάτων είναι παράμετροι που καθορίζονται από τον χρήστη. Στη συνέχεια, για κάθε διάστημα I_j, προσδιορίζουμε το σύνολο $Y_j = \{y | f(y) \in I_j\} = f^{-1}(I_j)$. Σε κάθε σύνολο Y_j εφαρμόζουμε έναν αλγόριθμο συσταδοποίησης και βρίσκουμε τις συστάδες Y_{jk} (σε αντιστοιχία με τις συνεκτικές συνιστώσες του γραφήματος Reeb). Κάθε συστάδα απεικονίζεται ως κορυφή του γραφήματος Mapper και προσθέτουμε μία ακμή μεταξύ δύο κορυφών αν οι αντίστοιχες συστάδες έχουν μη κενή τομή.

Παράμετροι του αλγορίθμου

Ο αλγόριθμος Mapper, αν και είναι απλός στην υλοποίηση του εγείρει πολλά ερωτηματικά καθώς απαιτεί τον καθορισμό πολλών παραμέτρων από τον χρήση. Διαφορετικές επιλογές ενδέχεται να οδηγήσουν σε διαφορετικά αποτελέσματα.

1. Επιλογή της συνάρτησης φίλτρου

   Η επιλογή της συνάρτησης φίλτρου έχει μεγάλη σημασία καθώς μπορεί να αποκαλύψει ενδιαφέρουσες δομές των δεδομένων. Η συνάρτηση αυτή δεν είναι απαραίτητο να περιοριστεί στο σύνολο των πραγματικών αριθμών. Αντίθετα,

μπορεί να απεικονίζει τα δεδομένα σε μετρικούς χώρους υψηλότερης διάστασης αποτυπώνοντας έτσι χαρακτηριστικά μεγαλύτερης διάστασης. Κάποιες από τις συνηθέστερες επιλογές είναι οι παρακάτω:

- Εκτιμητές πυκνότητας

Η πυκνότητα μπορεί να εκτιμηθεί θεωρώντας έναν Γκαουσιανό πυρήνα. Για $\varepsilon > 0$ η συνάρτηση ορίζεται ως εξής:

$$f_\varepsilon(x) = C_\varepsilon \sum_y \exp\left(-\frac{d(x,y)^2}{\varepsilon}\right)$$

όπου $x, y \in X$ και C_ε είναι κατάλληλη σταθερά ώστε $\int f_\varepsilon(x)dx = 1$

- Συνάρτηση εκκεντρότητας

Η οικογένεια των συναρτήσεων εκκεντρότητας, όπως και η πυκνότητα, φέρει πληροφορίες για το σχήμα των δεδομένων. Η βασική ιδέα είναι να εντοπίσουμε τα σημεία που βρίσκονται σε μεγάλη απόσταση από το κέντρο, χωρίς όμως να υπολογίσουμε κάποιο πραγματικό κέντρο. Για $1 \leq p < \infty$ η συνάρτηση εκκεντρότητας ορίζεται ως εξής:

$$E_p(x) = \left(\sum_{y \in X} \frac{d(x,y)^p}{N}\right)^{\frac{1}{p}}$$

όπου $x, y \in X$ και N είναι το μέγεθος του δείγματος.

Η συνάρτηση εκκεντρότητας παίρνει μεγάλες τιμές σε σημεία που βρίσκονται μακριά από το «κέντρο».

- Graph Laplacians

Θεωρούμε έναν Λαπλασιανό τελεστή στο σύνολο όλων των σημείων του συνόλου δεδομένων (που αποτελούν τις κορυφές του γραφήματος) και ορίζουμε το βάρος της ακμής που συνδέει τα σημεία x και y να είναι:

$$w(x,y) = k(d(x,y))$$

όπου d είναι η απόσταση μεταξύ σημείων και k είναι κάποιος πυρήνας ομαλοποίησης, για παράδειγμα ένας γκαουσιανός πυρήνας.

Στη συνέχεια κατασκευάζουμε τον κανονικοποιημένο Λαπλασιανό πίνακα διάστασης $N \times N$ τα στοιχεία του οποίου δίνονται ως εξής:

$$L(x,y) = \frac{w(x,y)}{\sqrt{\sum_z w(x,z)}\sqrt{\sum_z w(y,z)}}$$

Τα ιδιοδιανύσματα του πίνακα L μπορούν να αποκαλύψουν χρήσιμες πληροφορίες για το σχήμα των δεδομένων και χρησιμοποιούνται ως συναρτήσεις φίλτρου.

- Συνάρτηση συντεταγμένων

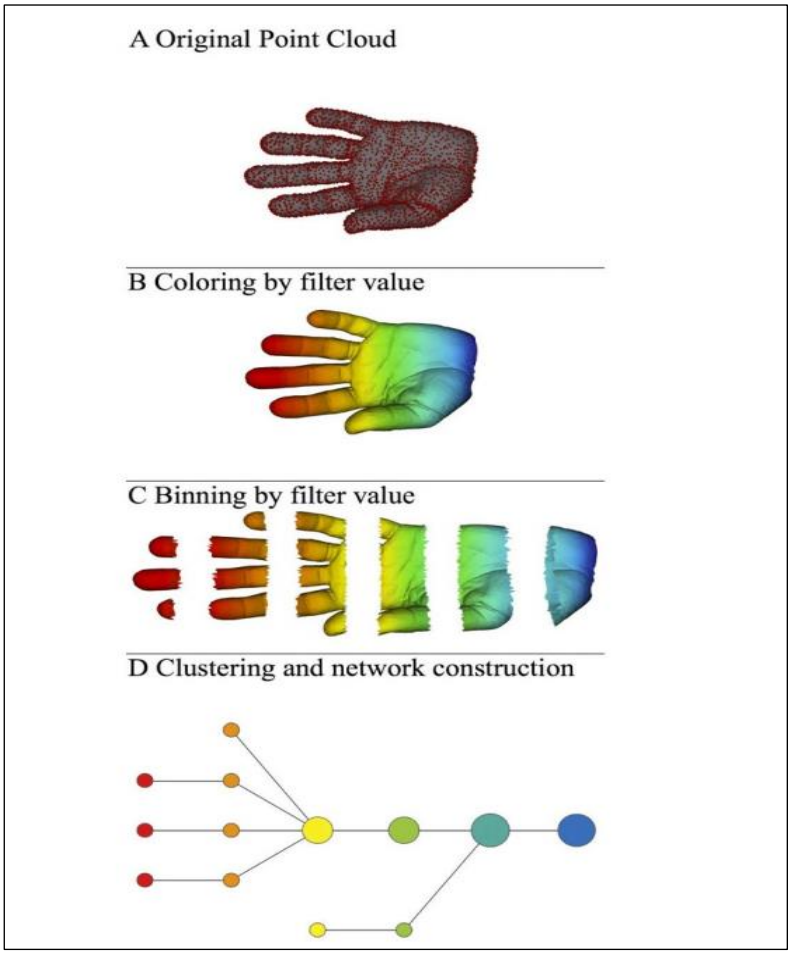
Η επιλογή κάποιας από τις συντεταγμένες (ή και περισσότερων) ως συνάρτηση φίλτρου αποτελεί μία από τις απλούστερες επιλογές, παράλληλα όμως είναι αρκετά διαισθητική και μπορεί να δώσει πληροφορίες για τη σημασία της εκάστοτε μεταβλητής (συντεταγμένης).

2. Επιλογή καλύμματος του $\mathbb{R}$

Όπως έχει ήδη αναφερθεί, όταν η f είναι πραγματική συνάρτηση, το κάλυμμα του πεδίου τιμών επιλέγεται να είναι ένα σύνολο διαστημάτων ίσου μήκους r και ποσοστού επικάλυψης g. Ωστόσο, δεν υπάρχει κάποιο σωστό μήκος ή ποσοστό κάλυψης . Μάλιστα, το αποτέλεσμα του αλγορίθμου είναι ευαίσθητο σε αλλαγές αυτών των παραμέτρων και ορισμένες τιμές ενδέχεται να μην αποδίδουν ικανοποιητικά τα χαρακτηριστικά του προς μελέτη χώρου. Μία μέθοδος αντιμετώπισης αυτού του προβλήματος είναι η δοκιμή ενός εύρους τιμών των παραμέτρων r και g και κατόπιν η επιλογή αυτών που θεωρούμε ότι προσφέρουν τις περισσότερες πληροφορίες. (Frédéric Chazal 2021)

3. Επιλογή αλγορίθμου συσταδοποίησης

Για την κατασκευή του γραφήματος Mapper είναι απαραίτητη η ομαδοποίηση των σημείων του συνόλου δεδομένων μέσα από τη συνάρτηση φίλτρου, δηλαδή τη συσταδοποίηση των σημείων που προκύπτουν ως αντίστροφη εικόνα των διαστημάτων που χρησιμοποιήθηκαν για την κάλυψη του πεδίου τιμών. Οι διάφορες μέθοδοι συσταδοποίησης μπορούν να χωριστούν στις εξής κατηγορίες: representative based clustering, ιεραρχική συσταδοποίηση και συσταδοποίηση βάσει πυκνότητας.



Εικόνα 13: Κατασκευή του Mapper για ένα τρισδιάστατο χέρι που αναπαρίσταται ως ένα νέφος σημείων. Η συνάρτηση φίλτρου f είναι η τετμημένη κάθε σημείου. Το πεδίο τιμών της χωρίζεται σε αλληλεπικαλυπτόμενα διαστήματα και τα σημεία χρωματίζονται βάσει της τιμής f στα διαστήματα αυτά. Τα σημεία χωρίζονται χρησιμοποιώντας την αντίστροφη εικόνα της f και χωρίζονται σε συστάδες με χρήση κάποιου αλγορίθμου συσταδοποίησης. Κάθε συστάδα αναπαρίσταται ως κορυφή (χρωματισμένη ανάλογα με την τιμή της συνάρτησης φίλτρου) και προστίθενται ακμές ανάμεσα στις συστάδες που έχουν κοινά στοιχεία.(P. Y. Lum 2013)

5.2 Ο αλγόριθμος Ball-Mapper

Ο αλγόριθμος Mapper απαιτεί τον καθορισμό πολλών παραμέτρων από τον χρήστη προκειμένου να εξαχθούν χρήσιμες πληροφορίες από τα δεδομένα. Ο αλγόριθμος Ball Mapper αποτελεί ένα απλό εργαλείο που μπορεί να δώσει χρήσιμες πληροφορίες για το σχήμα αλλά και για τα ίδια τα δεδομένα μέσω της οπτικοποίησής τους. Η βασική ιδέα του είναι η δημιουργία ενός καλύμματος του τοπολογικού χώρου προέλευσης των δεδομένων, έστω X. Οι παράμετροι του αλγορίθμου είναι ο καθορισμός μίας μετρικής (συνήθως επιλέγεται η ευκλείδεια) και η επιλογή ακτίνας ε. Αρχικά γίνεται η επιλογή της παραμέτρου ε και στη συνέχεια κατασκευάζονται μπάλες ακτίνας ε με κέντρα κάποια από τα σημεία του X προκειμένου να καλύψουμε ολόκληρο το X. Θα

συμβολίζουμε με C το σύνολο των κέντρων και με $B(C) = \bigcup_{x \in C} B(x, \varepsilon)$ το κάλυμμα που προκύπτει για το συγκεκριμένο σύνολο κέντρων και τη συγκεκριμένη ακτίνα. Επιπλέον, με $B(X, \varepsilon)$ θα συμβολίζουμε το σύνολο σημείων του X μαζί με μία λίστα κέντρων μπαλών έτσι ώστε για κάθε σημείο x του X οι μπάλες ακτίνας ε οι μπάλες που έχουν το κέντρο τους στο $B(X, \varepsilon)[x]$ να καλύπτουν το x, ενώ οι υπόλοιπες μπάλες όχι. Δηλαδή, το σύνολο $B(X, \varepsilon)$ μας λέει σε ποιες μπάλες ανήκει κάθε σημείο του X. Η επιλογή του ε είναι μία από τις προκλήσεις της μεθόδου καθώς γίνεται από τον χρήστη και δεν υπάρχουν πληροφορίες για το αν υπάρχει κάποιο «κατάλληλο» εύρος ώστε να αποκαλυφθούν ενδιαφέρουσες δομές του χώρου των δεδομένων. Είναι εμφανές, ότι πολύ μικρές τιμές για το ε θα δώσουν ένα πολύ λεπτομερές γράφημα χωρίς όμως κάποια ιδιαίτερη σημασία μιας και θα σχηματιστούν μπάλες σχεδόν γύρω από όλα τα σημεία. Από την άλλη, μία πολύ μεγάλη τιμή για την ακτίνα θα οδηγήσει στον σχηματισμό λίγων μπαλών με αποτέλεσμα τα δεδομένα να είναι αρκετά ανακατεμένα χωρίς να παρέχουν κάποια πληροφορία για το σχήμα του χώρου. Εάν το ε είναι κατάλληλο, το $B(C) = \bigcup_{x \in C} B(x, \varepsilon)$ είναι ένα καλό κάλυμμα του X δεδομένου ότι οι οι τομές των συνόλων του καλύμματος είναι είτε κενές είτε συσταλτά σύνολα. Συνεπώς, μπορούμε να κατασκευάσουμε το νεύρο του καλύμματος θεωρώντας κάθε κέντρο $c_1, c_2, \dots, c_n$ ως κορυφή και προσθέτοντας ένα k-simplex $[c_{i0}, c_{i1}, \dots, c_{ik}]$ όταν $B(c_{i0}, \varepsilon) \cap B(c_{i1}, \varepsilon) \cap \dots \cap B(c_{ik}, \varepsilon) \neq \emptyset$. Βάσει του Θεωρήματος του Νεύρου το X και το νεύρο N είναι ομοτοπικά ισοδύναμα εάν το κάλυμμα είναι «καλό». Το αποτέλεσμα αυτής της κατασκευής είναι το simplicial complex/γράφημα Ball-Mapper, δηλαδή ένα γράφημα με κορυφές τις μπάλες που προέκυψαν και ακμές μεταξύ των μπαλών που έχουν μη κενή τομή. Το μέγεθος κάθε μπάλας/κορυφής αντικατοπτρίζει τον αριθμό των παρατηρήσεων που περιέχει. Τέλος, οι μπάλες μπορούν να χρωματιστούν χρησιμοποιώντας μία συνάρτηση (π.χ. κάποια από τις συντεταγμένες). Επειδή, τελικά, αυτό που κατασκευάζουμε είναι ο 1-σκελετός του simplicial complex, αναφερόμαστε σε αυτό ως γράφημα Ball Mapper. (Dlotko 2019) (Pawel Dlotko 2022) (Wanling Qiu 2020)

Υπάρχουν δύο εναλλακτικές για την επιλογή των κέντρων των μπαλών:

1. Η κατασκευή ενός ε-net
2. Με χρήση ενός αλγορίθμου συσταδοποίησης όπως ο αλγόριθμος K-Means.

Ακολουθεί μία σύντομη αναφορά στους τρόπους κατασκευής ε-net (Dlotko 2019):

- **Αλγόριθμος 1**: Greedy ε-net

Είσοδος: Σύνολο δεδομένων X, $\varepsilon > 0$
Δημιουργία ενός αρχικά κενού διανύσματος $B(X, \varepsilon)$
Επανάλαβε: Διάλεξε ένα σημείο $p \in X$, το οποίο δεν ανήκει σε καμία μπάλα
Για κάθε $x \in B(p, \varepsilon) \cap X$, πρόσθεσε το p στο $B(X, \varepsilon)[x]$
Μέχρι να καλυφθούν όλα τα σημεία του X
Έξοδος: $B(X, \varepsilon)$

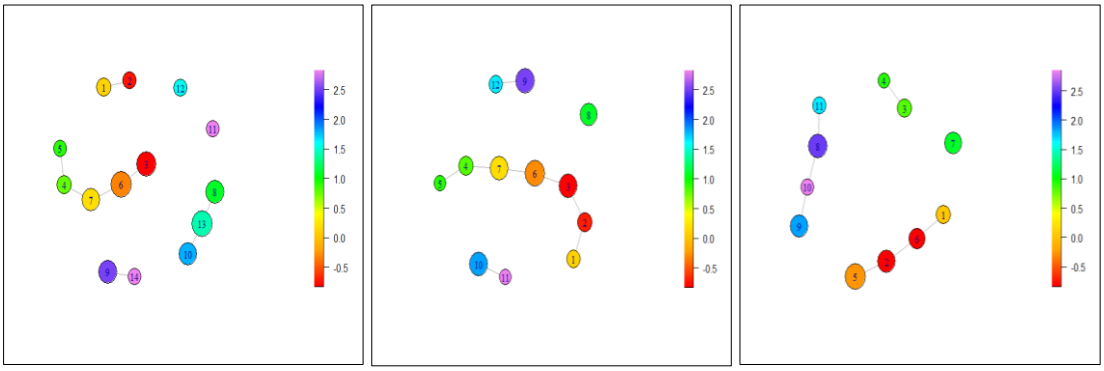
- **Αλγόριθμος 2:**Max-min ε-net

Είσοδος: Σύνολο δεδομένων $X, \varepsilon > 0$
Διάλεξε ένα αυθαίρετο σημείο $p \in X$ και θέσε $C = \{p\}$
Επανάλαβε: Βρες το σημείο $p^* \in X \backslash C$ που βρίσκεται πιο μακριά από το C
$d = dist(p^*, C)$
Θέσε C=$C \cup \{p^*\}$
Μέχρι $d \leq \varepsilon$
Δημιουργία ενός αρχικά κενού διανύσματος $B(X, \varepsilon)$
Για κάθε $p \in C$ βρες όλα τα σημεία $x \in B(p, \varepsilon) \cap X$ και πρόσθεσε το p στο $B(X, \varepsilon)[x]$
Έξοδος: $B(X, \varepsilon)$

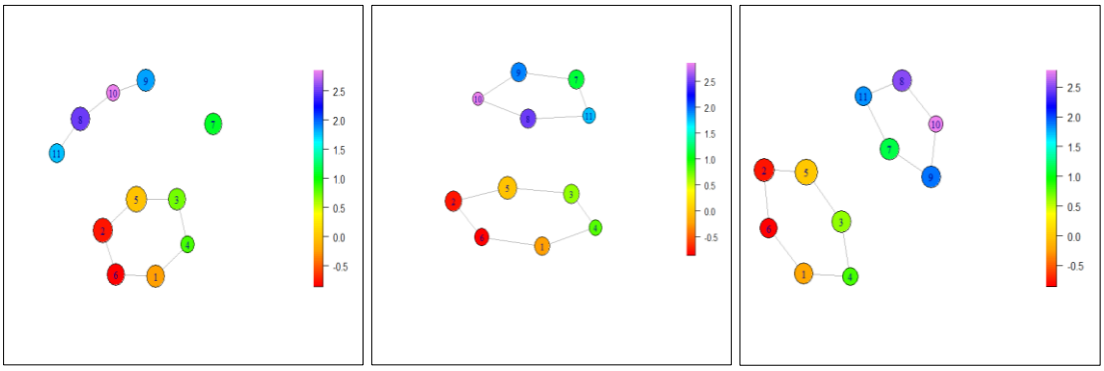
Σύνδεση του Mapper και του Ball-Mapper

Ας υποθέσουμε ότι η συνάρτηση φίλτρου $f: X \to \mathbb{R}$, όπου (X, ρ) μετρικός χώρος, που έχει επιλεχθεί για την κατασκευή του Mapper είναι ομοιόμορφα συνεχής, δηλαδή για κάθε x, y και για κάθε $\varepsilon > 0$ υπάρχει $\delta > 0$ τέτοιο ώστε $\rho(x, y) < \delta \Rightarrow |f(x) - f(y)| < \varepsilon$. Επιπλέον, θεωρούμε ότι ο αλγόριθμος συσταδοποίησης που χρησιμοποιείται είναι single linkage παραμέτρου ε και ότι το κάλυμμα του $\mathbb{R}$ αποτελείται από διαστήματα μήκους 6δ που αλληλοκαλύπτονται σε υποδιαστήματα μήκους δ. Τότε, τα σημεία που βρίσκονται κοντά στο Ball Mapper αναμένεται να βρίσκονται κοντά και στο Mapper, εφόσον η συνάρτηση φίλτρου που έχει επιλεχθεί για την κατασκευή του είναι συνεχής. Από την άλλη, τα σημεία που βρίσκονται κοντά στο Mapper θα συνδέονται στο Ball Mapper αλλά μπορεί να βρίσκονται οσοδήποτε μακριά. (Dlotko 2019)

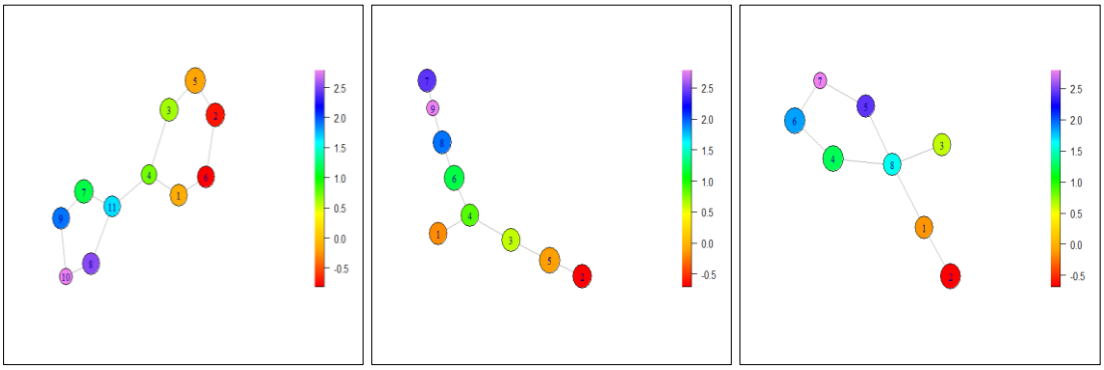
Παράδειγμα: Θεωρούμε ένα τυχαίο δείγμα το οποίο προέρχεται από δύο μοναδιαίους κύκλους με κέντρα (0,0) και (2,2) (από κάθε κύκλο έχουν επιλεγεί με τυχαίο τρόπο 50 παρατηρήσεις). Θα κατασκευάσουμε το γράφημα Ball Mapper για διάφορες τιμές του ε. Κάθε μπάλα χρωματίζεται βάσει της μέσης της των τετμημένων των παρατηρήσεων που περιέχει. Για την κατασκευή των γραφημάτων χρησιμοποιήθηκε το πακέτο "BallMapper" τηςR



(a) ε = 0.60 (b) ε = 0.65 (c) ε = 0.70



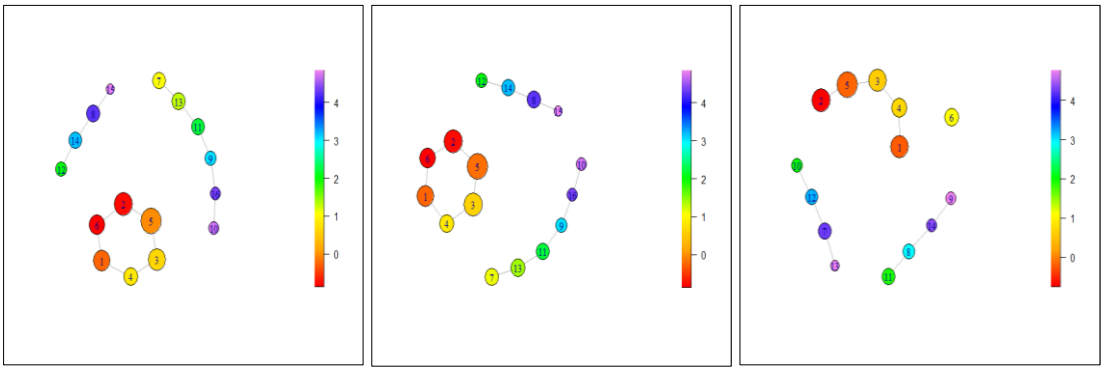
(d) ε = 0.75 (e) ε = 0.80 (f) ε = 0.84



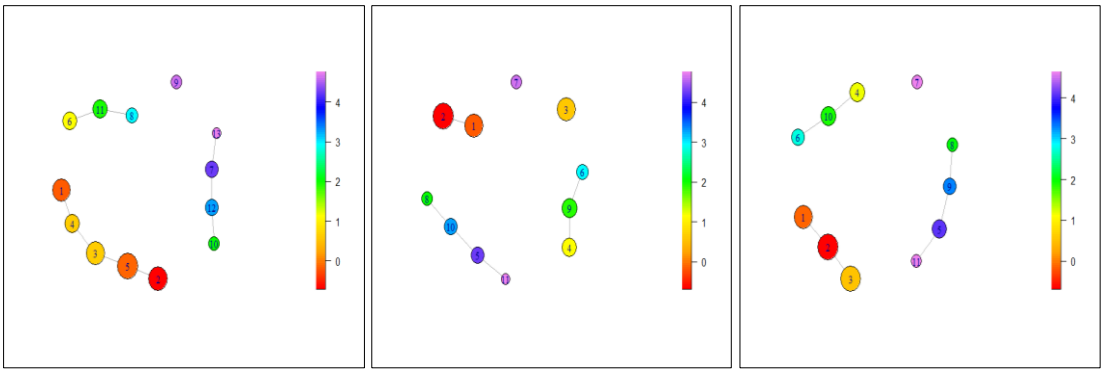
(g) ε = 0.88 (h) ε = 0.90 (i) ε =1.00

Εικόνα 14: Γραφήματα BallMapper για ένα σύνολο δεδομένων από δύο ίσους κύκλους για διάφορες τιμές του ε.

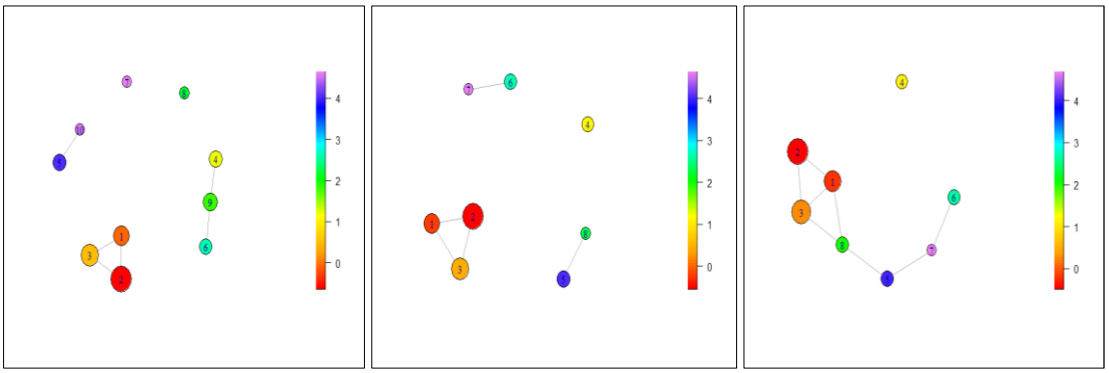
Στη συνέχεια, κατασκευάζουμε το Ball Mapper για ένα σύνολο δεδομένων που αποτελείται από 50 παρατηρήσεις που έχουν ληφθεί με τυχαίο τρόπο από τον μοναδιαίο κύκλο και 50 παρατηρήσεις που έχουν ληφθεί τυχαία από έναν κύκλο με ακτίνα 2 και κέντρο (3,3). Χρησιμοποιήθηκαν διάφορες τιμές για το ε και κάθε μπάλα χρωματίστηκε όπως και προηγουμένως.



(a) ε = 0.80 (b) ε = 0.85 (c) ε = 0.90



(d) ε = 0.95 (e) ε =1.00 (f) ε =1.10



(g)ε = 1.20 (h) ε =1.30 (i) ε =1.50

Εικόνα 15: Γραφήματα BallMapper για σύνολο δεδομένων από δύο κύκλους για διάφορες τιμές του ε.

6. Εφαρμογή της τοπολογικής ανάλυσης δεδομένων στο credit scoring

6.1 Credit Scoring

Το credit scoring ή credit rating ασχολείται με την αξιολόγηση του πιστωτικού κινδύνου, η οποία καταρχάς αντιμετωπίζεται ως ένα πρόβλημα δυαδικής ταξινόμησης. Αξιοποιώντας διάφορα χαρακτηριστικά όπως η ηλικία, ο μισθός, ο σκοπός δανεισμού κ.λπ. αν πρόκειται για ιδιώτες δανειολήπτες λιανικής τραπεζικής ή χρηματοοικονομικούς δείκτες αν πρόκειται για επιχειρήσεις, βαθμολογεί τους δανειζόμενους, επιχειρώντας έτσι την ιεράρχηση των αιτούντων ανάλογα με την πιστοληπτική τους ικανότητα και κατ' επέκταση τον διαχωρισμό τους σε δύο κατηγορίες: «καλούς» και «κακούς» (Siddiqi 2017), (Lean Yu 2008).

Οι τεχνικές που χρησιμοποιούνται μπορούν να χωριστούν στις εξής κατηγορίες (Lean Yu 2008):

1. Στατιστικές μέθοδοι, όπως η λογιστική παλινδρόμηση και η γραμμική διαχωριστική ανάλυση
2. Μέθοδοι μαθηματικού προγραμματισμού, όπως ο γραμμικός προγραμματισμός
3. Τεχνικές τεχνητής νοημοσύνης, όπως τα νευρωνικά δίκτυα και τα support vector machines
4. Υβριδικές μέθοδοι και μέθοδοι συλλογών ταξινομητών (ensembles)

Δεδομένης της υψηλής σημασίας της αξιολόγησης πιστωτικού κινδύνου, έχουν αναπτυχθεί και δοκιμαστεί πολλά μοντέλα ταξινόμησης. Μερικές από τις δημοφιλέστερες κλασσικές μεθόδους αποτελούν η λογιστική παλινδρόμηση και τα δέντρα αποφάσεων ενώ οι καινοτόμες μέθοδοι που συναντώνται συχνότερα είναι τα SVMs, τα νευρωνικά δίκτυα καθώς και οι συλλογές ταξινομητών (A. Markov 2022). Εν γένει, οι πιο μοντέρνες μέθοδοι και κυρίως οι συλλογές ταξινομητών έχουν καλύτερη απόδοση σε σχέση με τις κλασσικές μεθόδους, όπως η λογιστική παλινδρόμηση. (Stefan Lessmann 2013), (Lean Yu 2008), (X. Dastile 2020). Παρόλα αυτά η εύρεση του «καλύτερου» μοντέλου credit scoring αποδεικνύεται αρκετά πιο πολύπλοκο ζήτημα λαμβάνοντας υπόψη την απαίτηση ερμηνευσιμότητας, επεξηγησιμότητας και διαφάνειας (Lean Yu 2008), (Gero 2017), (Michael Bücker 2022). Οι αιτούντες δάνειο που απορρίπτονται έχουν το δικαίωμα να γνωρίζουν τον λόγο πίσω από την απόφαση κάτι που καθίσταται ιδιαίτερα δύσκολο όταν τα μοντέλα που χρησιμοποιούνται χαρακτηρίζονται από υψηλή πολυπλοκότητα. Το μοντέλο της λογιστικής παλινδρόμησης παραμένει ιδιαίτερα διαδεδομένο αφού λόγω της γραμμικής σχέσης μεταξύ των επεξηγηματικών μεταβλητών πληροί αυτές τις απαιτήσεις. (Gero 2017), (X. Dastile 2020). Από την άλλη, οι δύο τελευταίες κατηγορίες αν και είναι υποσχόμενες έχουν δεχθεί κριτική τόσο για την έλλειψη διαφάνειας και ερμηνευσιμότητας όσο και για την ενδεχόμενη τάση τους για overfitting (Yu, 2008), (Siddiqi 2017), (X. Dastile 2020), (Maria Rocha Sousa 2013).

6.2 Περιγραφή του τοπολογικού μοντέλου

Σκοπός μας είναι η μελέτη μιας εναλλακτικής μεθόδου κατηγοριοποίησης, μέσω της κατασκευής του νεύρου ενός κατάλληλου καλύμματος του υποκείμενου τοπολογικού χώρου των επεξηγηματικών μεταβλητών. Η μέθοδος αυτή έχει στέρεο θεωρητικό-μαθηματικό υπόβαθρο, είναι αρκετά απλή και εύκολα κατανοητή, και καθώς επιτρέπει την οπτικοποίηση των δεδομένων προσφέρει διαίσθηση για τη σχετική θέση ενός δανειζόμενου στον χώρο. Η βασική ιδέα είναι ότι άτομα που με παρόμοια χαρακτηριστικά (άρα άτομα που βρίσκονται στην ίδια γειτονιά) αναμένεται να έχουν παρόμοια συμπεριφορά. Επομένως, αν οι επεξηγηματικές μεταβλητές έχουν καλή προβλεπτική ικανότητα, μπορούμε να εκτιμήσουμε την πιθανότητα αθέτησης για έναν καινούριο δανειζόμενο εντοπίζοντας τη γειτονιά στην οποία ανήκει.

Σε αυτή τη βάση:

- θα κατασκευάσουμε ένα τοπολογικό μοντέλο για την εκτίμηση πιθανότητας αθέτησης
- θα το αξιολογήσουμε
- θα ασχοληθούμε με κάποιες από τις αδυναμίες και τους περιορισμούς του
- θα χρησιμοποιήσουμε την εμμένουσα ομολογία ως έναν αρχικό τρόπο περιγραφής των δεδομένων και ως μέσο διερεύνησης για την επιλογή της παραμέτρου ε του Ball Mapper.
- θα συγκρίνουμε την απόδοση του μοντέλου με το μοντέλο της λογιστικής παλινδρόμησης χρησιμοποιώντας ως μέτρο απόδοσης το εμβαδό κάτω από την καμπύλη (AUC) και το στατιστικό μέτρο Kolmogorov-Smirnov.
- θα μελετήσουμε έναν ακόμη τρόπο κατασκευής καλύμματος του υποκείμενου χώρου των χαρακτηριστικών και εκτίμησης πιθανοτήτων αθέτησης, την απόδοση του οποίου θα συγκρίνουμε με αυτή του Ball Mapper.

Τοπολογικό Μοντέλο Ι : Κατασκευή του Ball Mapper και πιθανότητες αθέτησης

Όπως έχει ήδη αναφερθεί, η κατασκευή του Ball Mapper απαιτεί τον καθορισμό μίας μετρικής και μίας παραμέτρου, της ακτίνας ε (Dlotko 2019). Για την εφαρμογή επιλέχθηκε η Ευκλείδεια μετρική. Ελπίζουμε ότι το ε που θα επιλέξουμε να είναι κατάλληλο για τη δημιουργία ενός «καλού» καλύμματος του υποκείμενου τοπολογικού χώρου από τον οποίο έγινε η δειγματοληψία . Η επιλογή του ε γίνεται από τον χρήστη χωρίς ωστόσο να υπάρχει κάποια ιδανική τιμή (Pawel Dlotko 2022). Αυτό το πρόβλημα θα αναλυθεί παρακάτω. Η ιδέα είναι η εξής: αφού γίνει η επιλογή του ε, χρησιμοποιούμε τις επεξηγηματικές μεταβλητές για να κατασκευάσουμε το γράφημα Ball Mapper και τη μεταβλητή απόκρισης (δηλαδή, αθέτηση=1/όχι αθέτηση=0) για τον χρωματισμό του. Το χρώμα κάθε μπάλας αντιστοιχεί στο ποσοστό «κακών» που βρίσκονται σε αυτήν.

Στη συνέχεια, γίνεται η εκτίμηση της πιθανότητας αθέτησης για κάθε νέα παρατήρηση ως εξής:

- εξετάζουμε σε ποια ή ποιες μπάλες ανήκει η καινούρια παρατήρηση και προσδιορίζουμε τις παρατηρήσεις που ανήκουν στην ένωση αυτών των μπαλών
- αποδίδουμε ως πιθανότητα αθέτησης στον καινούριο δανειζόμενο την μέση τιμή της μεταβλητής απόκρισης αυτών των παρατηρήσεων.

Στην περίπτωση που μία νέα παρατήρηση δεν βρίσκεται σε κάποια μπάλα ακολουθήσαμε δύο διαφορετικές προσεγγίσεις:

1. εντοπίζουμε την κοντινότερη παρατήρηση και βρίσκουμε την μπάλα ή τις μπάλες στις οποίες αυτή ανήκει και αποδίδουμε ως πιθανότητα αθέτησης στην νέα παρατήρηση την αναλογία κακών στις συγκεκριμένες μπάλες (τρόπος Α).
2. αντιμετωπίζουμε μία τέτοια παρατήρηση σα να βρίσκεται σε γκρίζα ζώνη και της αποδίδουμε ως πιθανότητα αθέτησης την αναλογία κακών του συνολικού πληθυσμού (τρόπος Β).

Ο αλγόριθμος Ball Mapper έχει χρησιμοποιηθεί σε εφαρμογές με οικονομικά δεδομένα (Wanling Qiu 2020), (Pawel Dlotko 2022)], με σκοπό όμως την οπτικοποίηση και κατανόησή τους, και όχι ως μοντέλο ταξινόμησης και πρόβλεψης.

Τοπολογικό Μοντέλο ΙΙ

Επιπλέον, θα εξετάσουμε ένα εναλλακτικό τοπολογικό μοντέλο, το οποίο χρησιμοποιεί παρόμοιο τρόπο με το persistence diagram για τη δημιουργία ενός καλύμματος του χώρου των χαρακτηριστικών. Πιο συγκεκριμένα, θα δημιουργήσουμε μπάλες ακτίνας ε γύρω από κάθε σημείο του συνόλου δεδομένων για να καλύψουμε το χώρο. Φυσικά, η επιλογή του ε παίζει καθοριστικό ρόλο στην επιτυχία του μοντέλου, και θα διερευνηθεί παρακάτω. Ωστόσο, το persistence diagram μπορεί να μας δώσει χρήσιμες πληροφορίες για το εύρος τιμών του. Αφού γίνει η επιλογή του ε και η κατασκευή του καλύμματος ακολουθούμε την ακόλουθη διαδικασία για την απόδοση πιθανοτήτων αθέτησης στις καινούριες παρατηρήσεις:

- υπολογίζουμε την απόσταση της νέας παρατήρησης από όλα τα σημεία του training set και βρίσκουμε τις μπάλες στις οποίες ανήκει (δηλαδή τις αποστάσεις της από τα κέντρα που είναι μικρότερες της ακτίνας που επιλέχθηκε)
- εξετάζουμε σε ποιες άλλες μπάλες ανήκει το κάθε κέντρο που επιλέχθηκε στο πρώτο βήμα
- αποδίδουμε ως πιθανότητα αθέτησης στην καινούρια παρατήρηση το ποσοστό κακών δανειζόμενων στο σύνολο των παρατηρήσεων που επιλέχθηκαν στα δύο προηγούμενα βήματα

Στην περίπτωση που μία νέα παρατήρηση δεν ανήκει σε καμία μπάλα, της προσδίδουμε πιθανότητα αθέτησης την αναλογία «κακών» στον αρχικό πληθυσμό.

6.3 Αναλυτική περιγραφή της εφαρμογής

6.3.1 Australian Credit Dataset

Για την αξιολόγηση του μοντέλου μελετήθηκε το Australian credit dataset, στο οποίο υπάρχει ελεύθερη πρόσβαση και χρησιμοποιείται ευρέως στην ανάπτυξη και στη μελέτη μοντέλων αξιολόγησης πιστοληπτικής ικανότητας. (A. Markov 2022).

Αποτελείται από 690 παρατηρήσεις σε 14 μεταβλητές και μία μεταβλητή που αντιπροσωπεύει την κλάση των παρατηρήσεων (1= κακός, 0= καλός). Το πλήθος «κακών» δανειζόμενων είναι 307 και των καλών 383 (πρόκειται δηλαδή για ένα σχετικά ισορροπημένο σύνολο δεδομένων). Επιπλέον, δεν υπάρχουν ελλείπουσες τιμές. Αν και το συγκεκριμένο σύνολο δεδομένων συναντάται συχνά σε σχετικές μελέτες, δεν υπάρχουν πληροφορίες σχετικά με την φύση των χαρακτηριστικών που εξετάζονται.

6.3.2 Επεξεργασία Δεδομένων

Αρχικά έγινε κανονικοποίηση των δεδομένων, προκειμένου οι τιμές για όλα τα χαρακτηριστικά να βρίσκονται στο εύρος [0,1] (A. Markov 2022). Οι παρατηρήσεις μετασχηματίστηκαν βάσει του τύπου:

$$x'_{i,j} = \frac{x_{i,j} - min_i x_{i,j}}{max_i x_{i,j} - min_i x_{i,j}}$$

όπου $x_{i,j}$ είναι η αρχική τιμή της i-παρατήρησης για το j χαρακτηριστικό, $min_i x_{i,j}$ και $max_i x_{i,j}$ είναι η ελάχιστη και μέγιστη τιμή του j χαρακτηριστικού αντίστοιχα.

Προκειμένου να αξιολογήσουμε την απόδοση των μοντέλων που χρησιμοποιήθηκαν το σύνολο δεδομένων χωρίστηκε σε training set και testing set σε αναλογία 70/30 και τηρήθηκε η αρχική αναλογία κακών δανειζόμενων στο δείγμα.

Η κατασκευή του Ball Mapper στην R βασίζεται στη σειρά με την οποία εμφανίζονται τα δεδομένα αφού ο αλγόριθμος ξεκινά κατασκευάζοντας μία μπάλα γύρω από την πρώτη παρατήρηση. Η δεύτερη μπάλα κατασκευάζεται γύρω από την επόμενη παρατήρηση που δε βρίσκεται μέσα στην πρώτη μπάλα και η διαδικασία επαναλαμβάνεται μέχρι να καλυφθούν όλα τα σημεία του συνόλου δεδομένων. Επομένως, διαφορετική σειρά εμφάνισης των δεδομένων αλλάζει το αποτέλεσμα του Ball Mapper. Συνιστάται, λοιπόν, να γίνονται αρκετά ανακατέματα του συνόλου δεδομένων προκειμένου να διαπιστωθεί η ευστάθεια του Ball Mapper (Pawel Dlotko

2022). Προκειμένου να μελετήσουμε αυτήν την «αδυναμία» και την επίδρασή της στην απόδοση του μοντέλου ανακατέψαμε 50 φορές το training set (αλλάξαμε, δηλαδή, τη σειρά εμφάνισης των παρατηρήσεων) και υπολογίσαμε κάθε φορά το εμβαδό κάτω από την καμπύλη που προκύπτει για το ίδιο testing set χρησιμοποιώντας και τους δύο τρόπους απόδοσης πιθανοτήτων αθέτησης. Για το ανακάτεμα χρησιμοποιήθηκε η εντολή set.seed για να είναι εφικτή η αναπαραγωγή και ο έλεγχος των αποτελεσμάτων.

6.3.3 Καθορισμός παραμέτρων

Καθορισμός παραμέτρων για το τοπολογικό μοντέλο Ι

Όπως έχει ήδη ειπωθεί η μοναδική παράμετρος (εκτός από την επιλογή μετρικής) που απαιτείται για την κατασκευή του Ball Mapper είναι η ακτίνα ε. Καθώς δεν υπάρχει συγκεκριμένη μεθοδολογία για τον προσδιορισμό της καταλληλότερης τιμής, ακολουθήσαμε την εξής διαδικασία: από το αρχικό training set δημιουργήσαμε πολλά βοηθητικά training-testing sets. Συγκεκριμένα, το αρχικό training set χωρίστηκε 50 φορές σε training set II και testing set II με αναλογία 70/30 και σε κάθε ένα από αυτά η αναλογία καλών-κακών είναι αυτή του αρχικού πληθυσμού. Σε κάθε επανάληψη εφαρμόστηκε ο αλγόριθμος Ball Mapper για το αντίστοιχο training set II για διάφορες τιμές του ε (από 0.8-1.79 με βήμα 0.01) και υπολογίστηκε το AUC, αποδίδοντας πιθανότητες αθέτησης στις νέες παρατηρήσεις (testing set II) και με τους δύο τρόπους που περιγράφονται παραπάνω. Στη συνέχεια, βρέθηκε το μέσο, ελάχιστο και μέγιστο AUC για κάθε τιμή του ε και η τυπική απόκλισή του. Χρησιμοποιήσαμε ως κριτήριο για την επιλογή του ε την τιμή που δίνει το μεγαλύτερο μέσο AUC. Με λίγα λόγια η προσέγγιση μας αποτελεί και πρόταση για τον τρόπο επιλογής του ε και συγκεκριμένα έγκειται στη βελτιστοποίηση μιας συνάρτησης του ε (εν προκειμένω της AUC) σε κάποιο σύνολο εκπαίδευσης (εν προκειμένω το εκάστοτε training set που χρησιμοποιούμε)

Καθορισμός παραμέτρων του τοπολογικού μοντέλου ΙΙ

Για τον καθορισμό της παραμέτρου ε ακολουθήσαμε τη ίδια διαδικασία. Το αρχικό training set χωρίστηκε 50 φορές σε training set II και testing set II με αναλογία 70/30 και σε κάθε ένα από αυτά η αναλογία καλών-κακών είναι αυτή του αρχικού πληθυσμού. Σε κάθε επανάληψη χρησιμοποιούμε διάφορες τιμές του ε (από 0.71 έως 1.20 με βήμα 0.01) για την κατασκευή καλύμματος (του χώρου του αντίστοιχου training set II) και στη συνέχεια αποδίδουμε τις πιθανότητες αθέτησης (στο testing set II) και υπολογίζουμε το AUC. Καταγράφουμε το μέσο, ελάχιστο και μέγιστο AUC, και την τυπική απόκλιση για κάθε τιμή του ε. Τέλος, επιλέγουμε το ε αντιστοιχεί στο μεγαλύτερο μέσο AUC.

6.3.4 Μέτρα αξιολόγησης του μοντέλου και cross validation

Ο σκοπός μας εδώ είναι να αξιολογήσουμε την απόδοση του Ball Mapper και γενικότερα της τοπολογικής προσέγγισης σε σχέση με κλασσικές μεθόδους ταξινόμησης, και συγκεκριμένα με τη λογιστική παλινδρόμηση, η οποία παραμένει ιδιαίτερα δημοφιλής στην αξιολόγηση πιστοληπτικής ικανότητας (Michael Bücker 2022). Ως μέτρο αξιολόγησης και σύγκρισης των μοντέλων χρησιμοποιήθηκε το εμβαδό κάτω από την καμπύλη (AUC), δηλαδή, η ικανότητα του εκάστοτε μοντέλου να βαθμολογεί υψηλότερα έναν κακό από έναν καλό (δεδομένου ότι η βαθμολογία μεταφράζεται ως πιθανότητα αθέτησης) και το στατιστικό Kolmogorov-Smirnov (KS).

Εμβαδόν κάτω από την καμπύλη (AUC):

Για έναν δυαδικό ταξινομητή η καμπύλη ROC είναι η γραφική παράσταση του FPR (False Positive Rate) στον οριζόντιο άξονα με την ευαισθησία TPR (True Positive Rate) στον κατακόρυφο άξονα για κάθε δυνατό κατώφλι. Το εμβαδό κάτω από την καμπύλη ROC (Receiver Operating Characteristic) εκφράζει την πιθανότητα ο ταξινομητής να βαθμολογήσει έναν κακό δανειζόμενο υψηλότερα από έναν καλό (δεδομένου ότι με 1 συμβολίζουμε τους κακούς και με 0 τους καλούς). Το AUC παίρνει τιμές στο [0.5,1].

Στατιστικό Kolmogorov-Smirnov:

Το στατιστικό KS είναι ένα ακόμη μέτρο της διαχωριστικής ικανότητας ενός ταξινομητή. Ο τύπος για τον υπολογισμό του είναι:

$\mathrm{KS} = \max_t(|\mathrm{TPR(t)} - \mathrm{FPR(t)}|)$, όπου t είναι το κατώφλι.

Όσο μεγαλύτερη η τιμή του, τόσο καλύτερη είναι η απόδοση του ταξινομητή.

Ένα σημαντικό πλεονέκτημα αυτών των μέτρων αξιολόγησης είναι δεν επηρεάζονται από ανισορροπία μεταξύ των κλάσεων. Τόσο το AUC όσο και το στατιστικό KS χρησιμοποιούν τα TPR και FPR τα οποία δεν εξαρτώνται από την αναλογία καλών-κακών (Mohammed J. Zaki 2020).

Cross validation:

Για την αξιολόγηση του μοντέλου Ι το αρχικό σύνολο δεδομένων χωρίστηκε 50 φορές σε training και testing set σε αναλογία 70/30 διατηρώντας την αναλογία καλών-κακών στο αρχικά δεδομένα. Σε κάθε επανάληψη, χρειάστηκε να επιλέξουμε την τιμή της παραμέτρου ε. Για τον σκοπό αυτό το κάθε ένα από τα 50 training sets χωρίστηκε 10 φορές σε training set ΙΙ και testing set ΙΙ (70/30 διατηρώντας την αρχική αναλογία καλών-κακών) προκειμένου να γίνει η επιλογή του ε (με κριτήριο το AUC) ακριβώς όπως περιγράφεται παραπάνω. Για το ε που προκύπτει εφαρμόζεται ο αλγόριθμος Ball Mapper για το training set, αποδίδονται πιθανότητες αθέτησης στις νέες παρατηρήσεις και με τους δύο τρόπους και υπολογίζεται το εμβαδόν κάτω από την καμπύλη.

Για το μοντέλο ΙΙ ακολουθήσαμε την ίδια διαδικασία με κάποιες διαφορές. Συγκεκριμένα, σε κάθε επανάληψη το αντίστοιχο training set χωρίστηκε σε training set ΙΙ και testing set ΙΙ (70/30 διατηρώντας την αναλογία καλών-κακών) 5 φορές προκειμένου να γίνει η επιλογή του ε. Για κάθε ένα από τα βοηθητικά training set ΙΙ δοκιμάσαμε 25 τιμές του ε (από 0.71-1.20, με βήμα 0.02) για να επιλέξουμε αυτό που δίνει το μεγαλύτερο μέσο AUC χρησιμοποιώντας το αντίστοιχο testing set ΙΙ. Στη συνέχεια, βρήκαμε το μέσο AUC για κάθε ε (των 5 διαφορετικών βοηθητικών συνόλων) και καταγράψαμε το ε που μεγιστοποιεί το μέσο AUC. Σε κάθε μία από τις 50 επαναλήψεις υπολογίσαμε τις πιθανότητες αθέτησης για τις παρατηρήσεις του αντίστοιχου (αρχικού) testing set και στη συνέχεια το εμβαδό κάτω από την καμπύλη.

6.3.5 Λογιστική Παλινδρόμηση

Το μοντέλο της λογιστικής παλινδρόμησης χρησιμοποιείται συχνά στο credit scoring καθώς λόγω της γραμμικής σχέσης των επεξηγηματικών μεταβλητών είναι εύκολα ερμηνεύσιμο (Gero 2017), (Michael Bücker 2022).

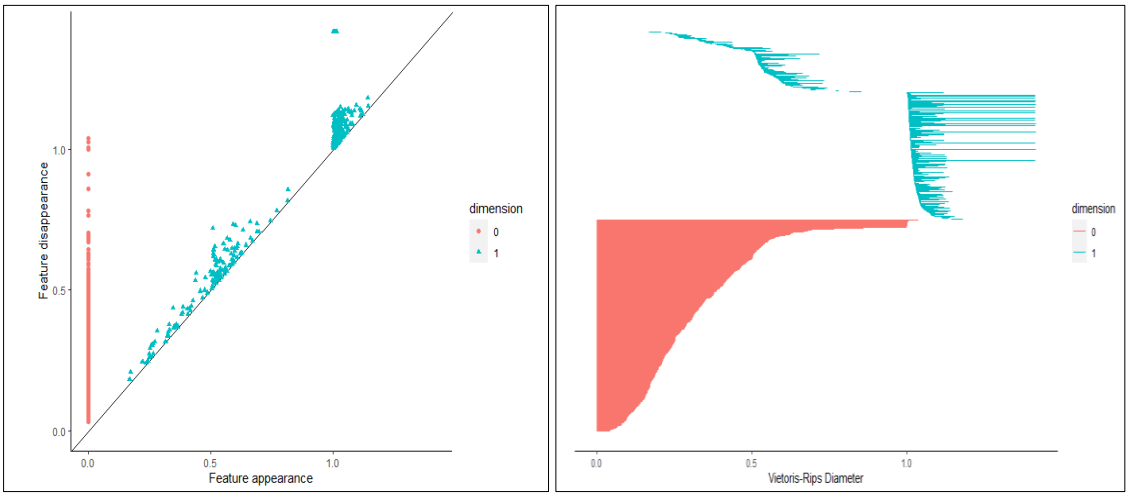
Αν υποθέσουμε ότι η μεταβλητή απόκρισης Y είναι δίτιμη (και άρα κωδικοποιείται με 0 και 1) η λογιστική παλινδρόμηση εκτιμά, χρησιμοποιώντας τις ανεξάρτητες μεταβλητές, την πιθανότητα να πάρει κάθε μία από αυτές τις τιμές. Οι πιθανότητες $P(Y = 1|X = x) = \pi, P(Y = 0|X = x) = 1 - \pi$ υπολογίζονται ως (Kantardzic 2020) :

$$logit(\pi) = log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_\nu X_\nu$$

Αφού κατασκευάστηκε το μοντέλο της λογιστικής παλινδρόμησης για το training set εκτιμήθηκε η πιθανότητα αθέτησης για κάθε νέο δανειολήπτη του testing set. Στη συνέχεια, αξιολογήθηκε η απόδοση του μοντέλου με cross validation και τα αποτελέσματα συγκρίθηκαν με αυτά των τοπολογικών μοντέλων που εξετάστηκαν.

6.4 Αποτελέσματα

Αρχικά, παρουσιάζεται το persistence diagram και το barcode για το training set το οποίο θα μας δώσει κάποιες πρώτες πληροφορίες για το σχήμα των δεδομένων. Για την κατασκευή του persistence diagram και του barcode χρησιμοποιήθηκε το πακέτο "TDAstats"της R.



Εικόνα 16: Persistence diagram και barcode για το training set.

Αυτό που παρατηρούμε είναι ότι υπάρχουν κάποιες συνεκτικές συνιστώσες που επιμένουν (αυτές που πεθαίνουν λίγο μετά το 0.75) και κάποιοι κύκλοι οι οποίοι βρίσκονται αρκετά μακριά από τη διαγώνιο (οι οποίοι δημιουργούνται λίγο μετά το 1). Σημειώνεται ότι υπολογίστηκε και η εμμένουσα ομολογία διάστασης 2, χωρίς όμως να παρατηρηθεί κάτι ενδιαφέρον.

Στον παρακάτω πίνακα παρουσιάζονται οι μέσοι χρόνοι ζωής και οι τυπικές αποκλίσεις τους για τις συνεκτικές συνιστώσες και τους κύκλους:

Πίνακας 1: Μέση τιμή και τυπική απόκλιση χρόνου ζωής συνεκτικών συνιστωσών και κύκλων (training set)

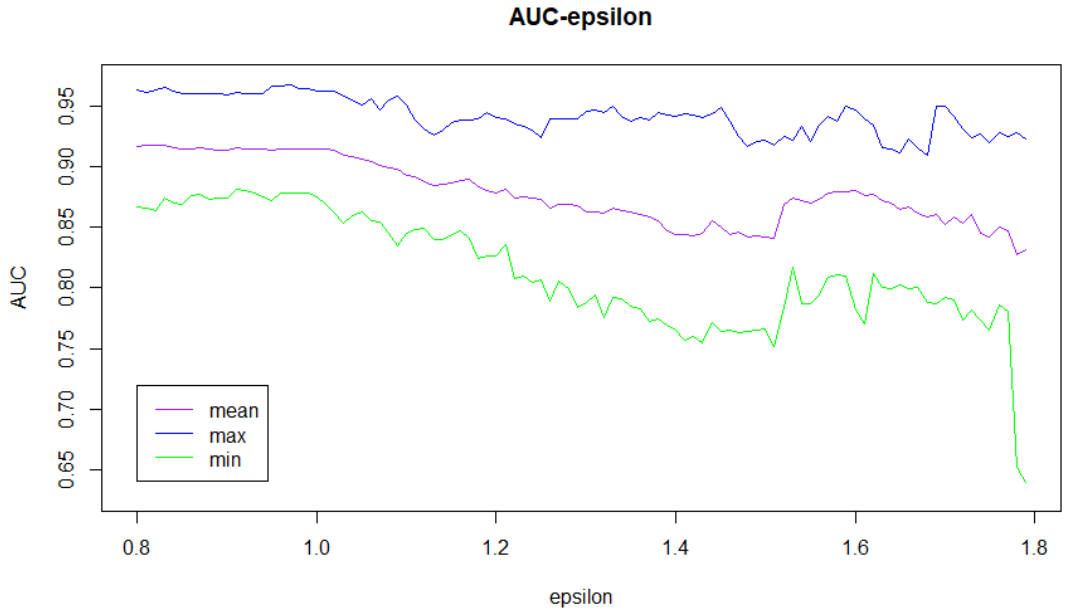
Διάσταση	0	1
Μέσος χρόνος ζωής	0.3327887	0.05128132
Τυπική απόκλιση χρόνου ζωής	0.2036795	0.08090239

Τοπολογικό μοντέλο I (Ball Mapper)

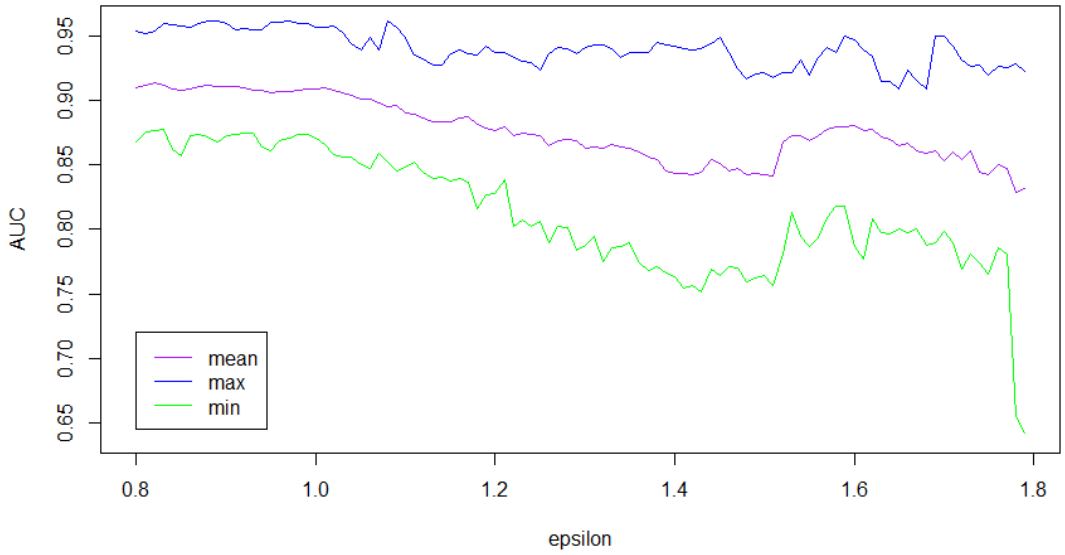
Επιλογή του ε και κατασκευή γραφημάτων Ball Mapper:

Προκειμένου να καθορίσουμε το ε για την κατασκευή του BallMapper χωρίσαμε το training set σε training II και testing II 50 φορές, χρησιμοποιώντας την εντολή set.seed και κατασκευάσαμε το BM για 100 τιμές του ε, από 0.80-1.79 με βήμα 0.01. Επιλέξαμε το συγκεκριμένο εύρος τιμών για το ε για τρεις λόγους. Ο πρώτος είναι ότι το persistence diagram για το training set μας έδωσε μια πρώτη εικόνα για τις συνεκτικές συνιστώσες που επιμένουν. Παρόλο που το ε που απαιτεί ο αλγόριθμος Ball Mapper είναι διαφορετικό από αυτό που χρησιμοποιείται για την κατασκευή του persistence diagram αναμένουμε πως σε αυτήν την (ευρύτερη) περιοχή το γράφημα που προκύπτει

θα αποκαλύψει ενδιαφέρουσες πληροφορίες για τις συνεκτικές συνιστώσες του χώρου των χαρακτηριστικών. Επιπλέον, η κατασκευή του BM για μικρότερα ε δεν δίνει ιδιαίτερα χρήσιμες πληροφορίες, αφού δημιουργούνται πολλές μπάλες (αρκετές από τις οποίες περιέχουν μόνο μία παρατήρηση) κάτι που αντίκειται στον αρχικό μας στόχο να αποκτήσουμε μία περίληψη του υποκείμενου τοπολογικού χώρου. Οι τρίτος λόγος είναι πρακτικός. Συγκεκριμένα, η επαναληπτική διαδικασία για τον καθορισμό του ε απαιτεί πολύ χρόνο. Μετά από κάθε κατασκευή, αποδόθηκαν πιθανότητες αθέτησης και με τους δύο τρόπους που περιγράφονται παραπάνω, βρέθηκαν τα AUCs και υπολογίστηκε η μέση, η ελάχιστη και η μέγιστη τιμή τους καθώς και η τυπική απόκλισή τους για κάθε τιμή του ε. Για την κατασκευή του Ball Mapper χρησιμοποιήθηκε το πακέτο "BallMapper" και για τον υπολογισμό του AUC το πακέτο "pROC" στην R. Τα αποτελέσματα παρουσιάζονται στα παρακάτω διαγράμματα:



Εικόνα 17: Μέσο, ελάχιστο και μέγιστο AUC για διάφορες τιμές του ε (μοντέλο Ι-τρόπος Α)



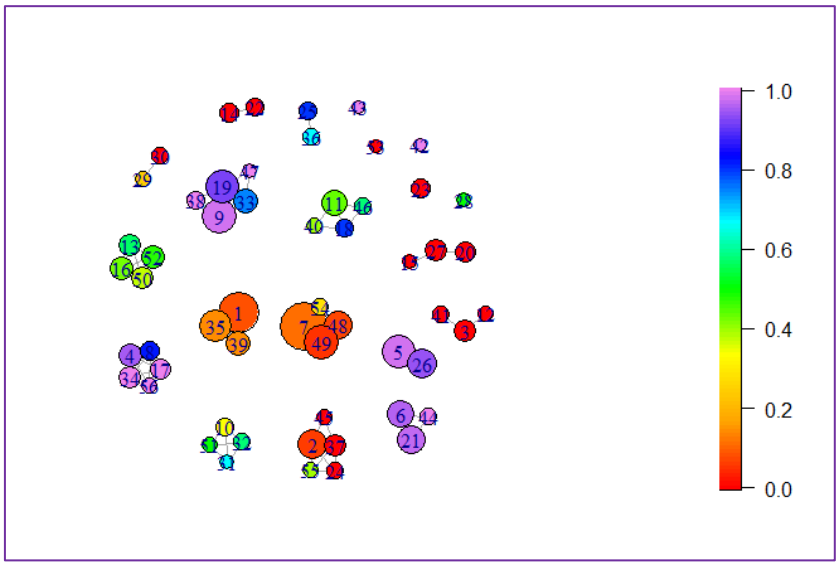
Εικόνα 18: Μέσο, ελάχιστο και μέγιστο AUC για διάφορες τιμές του ε (μοντέλο I-τρόπος B)

Παρατηρούμε ότι και με τις δύο μεθόδους οι μεγαλύτερες τιμές για το μέσο AUC προκύπτουν για μικρές τιμές του ε, από 0.80-1.05 χωρίς βέβαια να σημειώνονται αξιόλογες διαφορές. Κοντά στο 1.10 ξεκινά μία μικρή πτώση του AUC η οποία συνεχίζεται μέχρι το 1.50 χωρίς ωστόσο το μέσο AUC να επανέρχεται στις αρχικές τιμές του. Είναι αξιοσημείωτη η μεγάλη πτώση του ελάχιστου AUC για ε =1.79 και στις δύο περιπτώσεις.

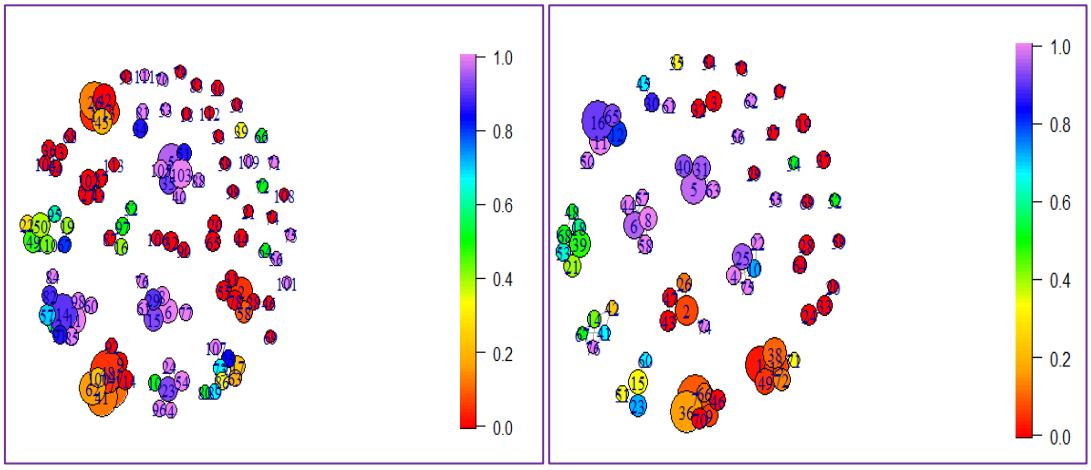
Στον Πίνακας 3 και στον Πίνακας 4 του παραρτήματος παρουσιάζονται αναλυτικά οι τιμές του μέσου, ελάχιστου και μέγιστου AUC καθώς και οι αντίστοιχες τυπικές αποκλίσεις για τις διάφορες τιμές του ε για τις μεθόδους A και B αντιστοίχως.

Το μεγαλύτερο μέσο AUC και στις δύο περιπτώσεις παρατηρήθηκε για ε = 0.82. Με τον τρόπο A, για ε = 0.82, το μέσο, ελάχιστο και μέγιστο AUC προέκυψαν 0.9176, 0.8657, και 0.9630 αντίστοιχα και η τυπική απόκλισή των AUCs βρέθηκε 0.02123. Με τον τρόπο B, για ε = 0.82, το μέσο, ελάχιστο και μέγιστο AUC προέκυψαν 0.9131, 0.8768, και 0.9534 αντίστοιχα και η τυπική απόκλισή τους 0.02.

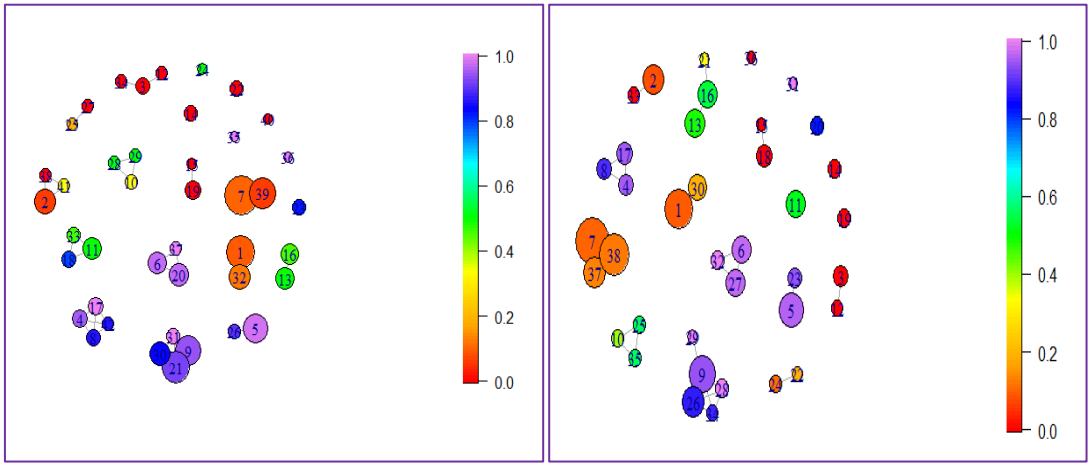
Η κατασκευή του Ball Mapper θα γίνει για διάφορες τιμές του ε, προκειμένου να αποκτήσουμε μια πιο ολοκληρωμένη εικόνα για την εξέλιξη του γραφήματος καθώς η ακτίνα μεγαλώνει. Σε κάθε περίπτωση, αποδίδονται οι πιθανότητες αθέτησης για τις παρατηρήσεις του training set χρησιμοποιώντας τις μεθόδους A&B, και αξιολογείται η απόδοση του μοντέλου με χρήση του AUC και του KS. Εκτός από αυτά τα μέτρα υπολογίζεται και ο αριθμός των νέων παρατηρήσεων που δεν τοποθετούνται σε κάποια από τις μπάλες.



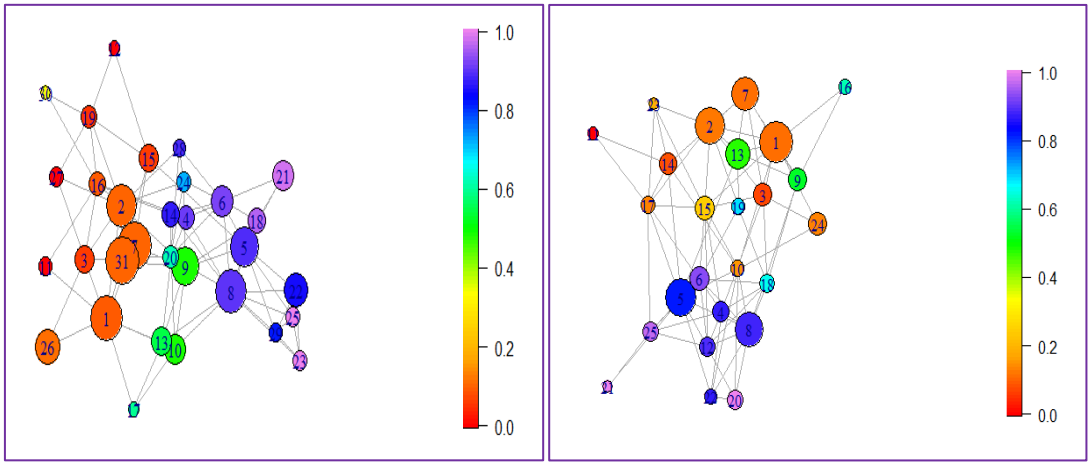
Εικόνα 19: Γράφημα BallMapper για ε=0.82



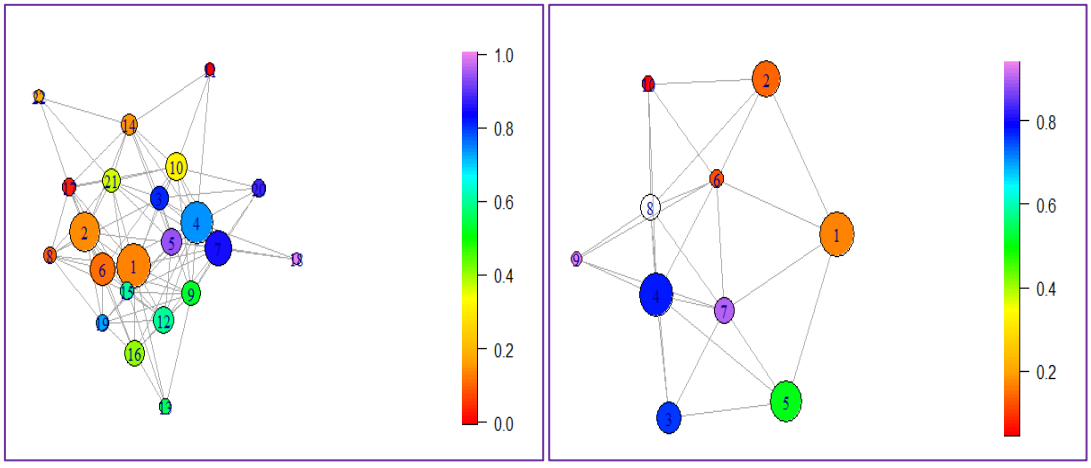
Εικόνα 20: Γραφήματα BallMapper για ε= 0.6 (αριστερά) και ε=0.7 (δεξιά)



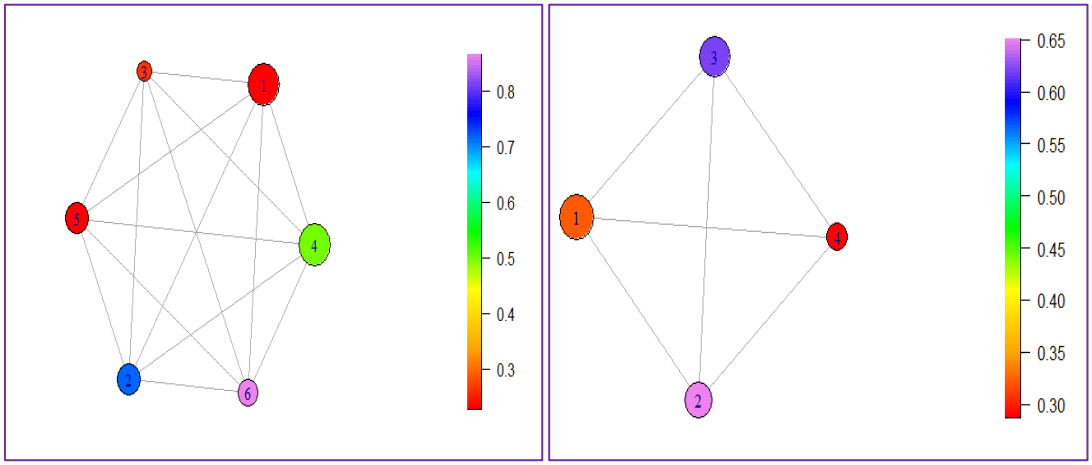
Εικόνα 21: Γραφήματα BallMapper για ε= 0.9 (αριστερά) και ε=0.97 (δεξιά)



Εικόνα 22: Γραφήματα BallMapper για ε=1.05 (αριστερά) και ε=1.12 (δεξιά)



Εικόνα 23: Γραφήματα BallMapper για ε=1.2 (αριστερά) και ε=1.4 (στο αριστερό γράφημα η μπάλα 8 έχει ποσοστό αθέτησης 0.9367)



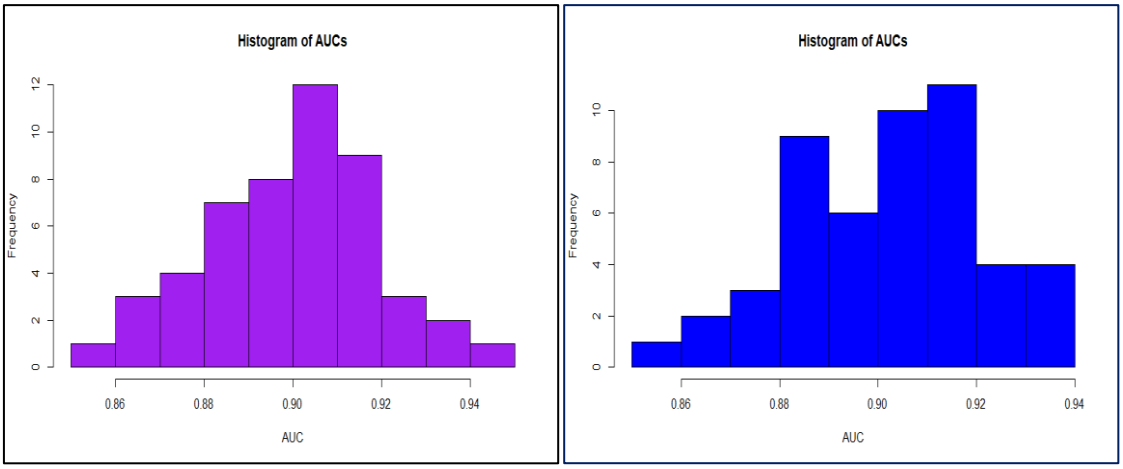
Εικόνα 24: Γραφήματα BallMapper για ε=1.60 (αριστερά) και ε=1.79 (δεξιά)

Πίνακας 2: Οι τιμές των AUC και KS για διάφορες τιμές του ε για το training set (μοντέλο Ι)

	Τρόπος Α		Τρόπος Β		
ε	AUC	KS	AUC	KS	εκτός μπάλας
0.60	0.8796	0.7592838 t= 0.3077	0.8998	0.6974233 t= 0.3077	30
0.70	0.8786	0.7310534 t= 0.33334	0.8953	0.7036756 t = 0.33334	21
0.80	0.8837	0.7616521 t= 0.43479	0.9017	0.7569155 t=0.44001	10
0.82	0.8821	0.7420424 t= 0.42106	0.9007	0.7459265 t= 0.44001	6
0.90	0.8851	0.7271694 t=0.36364	0.8995	0.7334218 t= 0.44001	8
0.97	0.8788	0.7185487 t=0.19231	0.8941	0.731906 t=0.19231	6
1.05	0.8803	0.6840659 t= 0.10938	0.9039	0.706044 t=0.10938	4
1.12	0.8851	0.6377416 t= 0.25491	0.8970	0.6597196 t= 0.25491	5
1.2	0.8891	0.6502463 t= 0.40678	0.8978	0.6722243 t= 0.40678	3
1.4	0.8208	0.5866806 t= 0.3358	0.8248	0.5976696 t= 0.3358	1
1.6	0.8679	0.662751 t= 0.4163	0.8679	0.662751 t= 0.4163	0
1.79	0.802	0.5797651 t= 0.44491	0.802	0.5797651 t= 0.44491	0

Cross Validation

Προκειμένου να γίνει η επικύρωση του μοντέλου I (Ball Mapper) με τους τρόπους Α και Β, πραγματοποιήσαμε 50 επαναλήψεις στις οποίες το αρχικό σύνολο δεδομένων χωρίστηκε σε training set I και testing set I σε αναλογία 70/30 με αναλογία κακών αυτή του αρχικού πληθυσμού (δηλαδή 0.44). Σε κάθε επανάληψη χρειάστηκε να καθορίσουμε την παράμετρο του μοντέλου, δηλαδή την ακτίνα ε. Για το σκοπό αυτό το κάθε training set I χωρίστηκε 10 φορές σε training set II και testing set II τηρώντας τις παραπάνω αναλογίες. Σε κάθε μία από αυτές τις επαναλήψεις, εφαρμόσαμε τον αλγόριθμο Ball Mapper για ε από 0.80-1.05 με βήμα 0.01, υπολογίσαμε τις πιθανότητες αθέτησης για το αντίστοιχο testing set II και βρήκαμε το AUC. Μόλις τελείωσαν οι 10 επαναλήψεις υπολογίσαμε το μέσο AUC και προσδιορίσαμε το ε στο οποίο αντιστοιχεί το μεγαλύτερο μέσο AUC. Χρησιμοποιώντας αυτό το ε εφαρμόσαμε τον αλγόριθμο Ball Mapper και καταγράψαμε το AUC χρησιμοποιώντας τα αντίστοιχα training I-testing I sets. Οι τιμές των AUCs καθώς και τα ε που επιλέχθηκαν και με τις δύο μεθόδους παρουσιάζονται στον Πίνακας 5 του παραρτήματος. Το μέσο AUC βρέθηκε 0.8994 με τον τρόπο Α και 0.9017 με τον τρόπο Β και οι τυπικές αποκλίσεις 0.0196 και 0.0189 αντίστοιχα. Τα αποτελέσματα παρουσιάζονται συνοπτικά στα ακόλουθα ιστογράμματα συχνοτήτων.

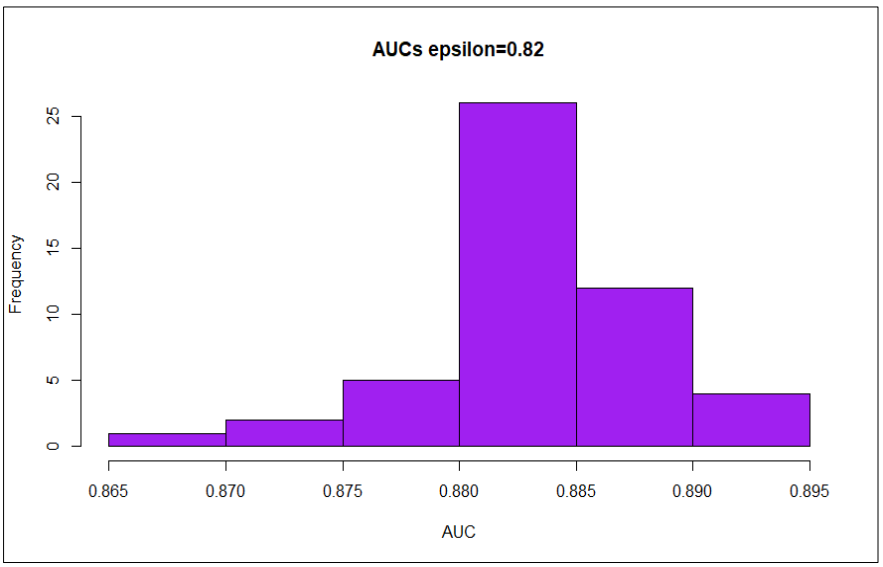


Εικόνα 25: Ιστογράμματα συχνοτήτων των AUCs που προέκυψαν κατά το cross validation του τοπολογικού μοντέλου I και με τους δύο τρόπους. Αριστερά χρησιμοποιώντας τον τρόπο Α και δεξιά τον τρόπο Β

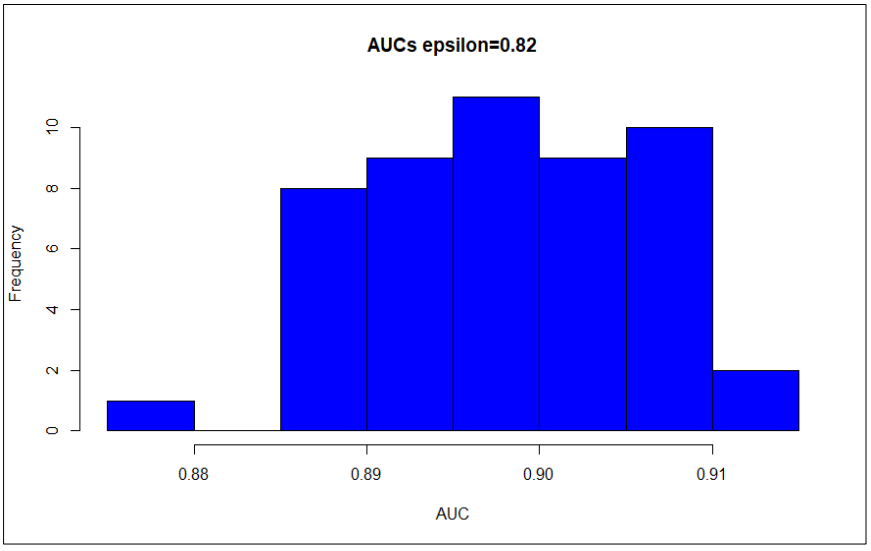
Ανακάτεμα των δεδομένων:

Στη συνέχεια, ανακατέψαμε το training set 50 φορές, και για κάθε ανακάτεμα κατασκευάσαμε το Ball Mapper με ε =0.82 (αφού αυτό μας έδωσε το μεγαλύτερο μέσο AUC για το συγκεκριμένο training set) και αποδώσαμε πιθανότητες αθέτησης με τους τρόπους Α και Β. Υπολογίσαμε τις τιμές του AUC, προκειμένου να αξιολογήσουμε κατά πόσο η αλλαγή στη σειρά των δεδομένων θα μεταβάλλει την απόδοση του μοντέλου. Χρησιμοποιώντας τον τρόπο Α το μέσο AUC βρέθηκε 0.8836 και η τυπική απόκλιση 0.0046. Για τον τρόπο Β προέκυψε μέσο AUC 0.8979 με και η τυπική

απόκλιση βρέθηκε 0.0075. Στον Πίνακας 6 του παραρτήματος παρουσιάζονται αναλυτικά οι τιμές των AUCsαπό τα 50 ανακατέματα τα του training set και με τους δύο τρόπους. Στις παρακάτω εικόνες παρουσιάζονται τα αντίστοιχα ιστογράμματα.



Εικόνα 26: Ιστόγραμμα συχνοτήτων των AUCs που προέκυψαν από το ανακάτεμα των δεδομένων (μοντέλο Ι-τρόπος Α)



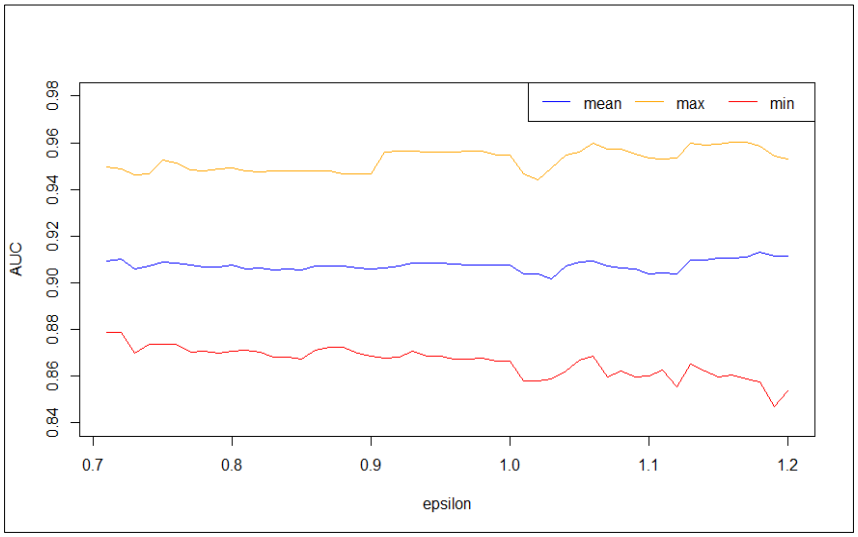
Εικόνα 27: Ιστόγραμμα συχνοτήτων των AUCs που προέκυψαν από το ανακάτεμα των δεδομένων (μοντέλο Ι-τρόπος Β)

Τοπολογικό μοντέλο ΙΙ

Καθορισμός της παραμέτρου ε:

Για να καθορίσουμε την παράμετρο ε του τοπολογικού μοντέλου ΙΙ, χωρίσαμε το training set 50 φορές σε βοηθητικά training/testing sets σε αναλογία 70/30 διατηρώντας την αρχική αναλογία «κακών». Σε κάθε επανάληψη, κατασκευάσαμε το κάλυμμα του υποκείμενου χώρου για διάφορες τιμές του ε (από 0.71-1.20, με βήμα 0.01), αποδώσαμε πιθανότητες αθέτησης και υπολογίσαμε τις τιμές του AUC. Το εύρος τιμών καθορίστηκε από το persistence diagram, σε μια προσπάθεια να αποτυπώσουμε τόσο τις συνεκτικές συνιστώσες όσο και τους κύκλους που φαίνεται να επιμένουν και άρα χαρακτηρίζουν τα δεδομένα. Επιπλέον, λήφθηκε υπόψη το πρακτικό πρόβλημα του ιδιαίτερα μεγάλου υπολογιστικού χρόνου που απαιτείται.

Το μεγαλύτερο μέσο AUC προέκυψε για ε = 1.18. Για αυτήν την τιμή το μέσο, μέγιστο και ελάχιστο AUC ήταν 0.9173670, 0.9672790 και 0.8720819 αντίστοιχα, και η τυπική απόκλιση τους 0.01957162. Ο Πίνακας 7 παρουσιάζει αναλυτικά τα αποτελέσματα για όλες τις τιμές του ε που χρησιμοποιήσαμε. Στην παρακάτω εικόνα, αναπαρίστανται γραφικά το μέσο, μέγιστο και ελάχιστο AUC.



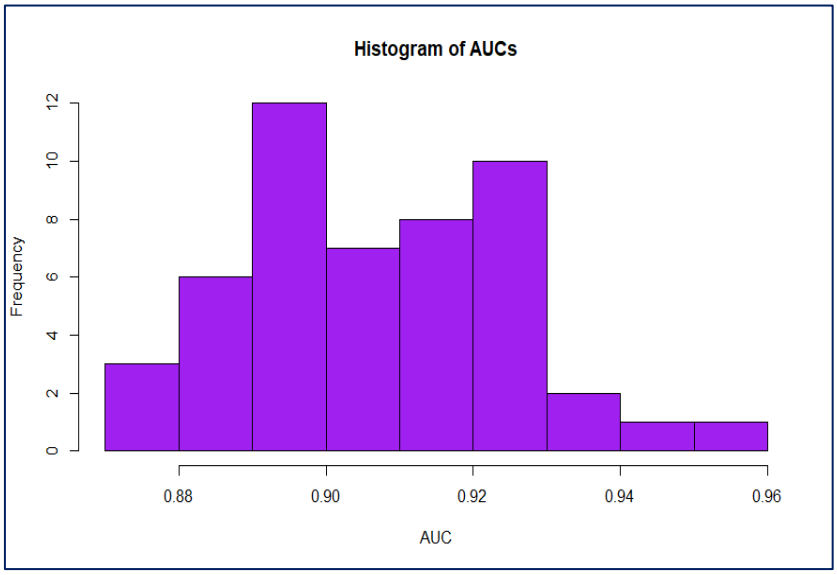
Εικόνα 28: Μέσο, ελάχιστο και μέγιστο AUC για διάφορες τιμές του ε για το μοντέλο ΙΙ

Και για αυτό το τοπολογικό μοντέλο παρατηρούμε ότι δε σημειώνονται αξιόλογες διαφορές στο μέσο AUC για τις διάφορες τιμές του ε.

Για ε =1.18 (δηλαδή, για την τιμή του ε που αντιστοιχεί στο μέγιστο μέσο AUC) θα εφαρμόσουμε την μέθοδο που περιγράψαμε και θα υπολογίσουμε πιθανότητες αθέτησης. Το AUC προέκυψε 0.898 και το στατιστικό KS = 0.7592838 (threshold = 0.46092).

Cross validation:

Τέλος, για την επικύρωση του μοντέλου πραγματοποιήσαμε 50 επαναλήψεις στις οποίες το αρχικό σύνολο δεδομένων χωρίστηκε σε training set και testing set σε αναλογία 70/30 με αναλογία κακών αυτή του αρχικού πληθυσμού. Σε κάθε επανάληψη χρειάστηκε να καθορίσουμε την παράμετρο του μοντέλου, δηλαδή την ακτίνα ε. Για το σκοπό αυτό το training set χωρίστηκε 5 φορές σε training set II και testing set II τηρώντας τις παραπάνω αναλογίες. Σε κάθε μία από αυτές τις επαναλήψεις, κατασκευάσαμε το κάλυμμα του υποκείμενου χώρου για ε από 0.72-1.20 με βήμα 0.02, υπολογίσαμε τις πιθανότητες αθέτησης για το αντίστοιχο training set II και βρήκαμε το AUC. Μόλις τελείωσαν οι 5 επαναλήψεις υπολογίσαμε το μέσο AUC και προσδιορίσαμε το ε στο οποίο αντιστοιχεί το μεγαλύτερο μέσο AUC. Χρησιμοποιώντας αυτό το ε υπολογίσαμε τις πιθανότητες αθέτησης για κάθε ένα από τα αρχικά testing sets και καταγράψαμε το AUC. Το μέσο AUC βρέθηκε 0.9066 και η τυπική απόκλιση ίση με 0.0180. Ο Πίνακας 8 του παραρτήματος δείχνει αναλυτικά τα AUCs που προέκυψαν και τις τιμές του ε που επιλέχθηκαν σε κάθε επανάληψη. Τα αποτελέσματα απεικονίζονται και στο παρακάτω ιστόγραμμα συχνοτήτων:



Εικόνα 29: Ιστόγραμμα συχνοτήτων των AUCs που προέκυψαν κατά το cross validation του μοντέλου II

Λογιστική Παλινδρόμηση

Το μοντέλο της λογιστικής παλινδρόμησης εκπαιδεύτηκε πάνω στο ίδιο training set όπως και τα προηγούμενα μοντέλα. Στον Πίνακας 9 του παραρτήματος παρουσιάζονται οι τιμές των συντελεστών, το τυπικό σφάλμα τους και η τιμή του p-value του ελέγχου σημαντικότητας.

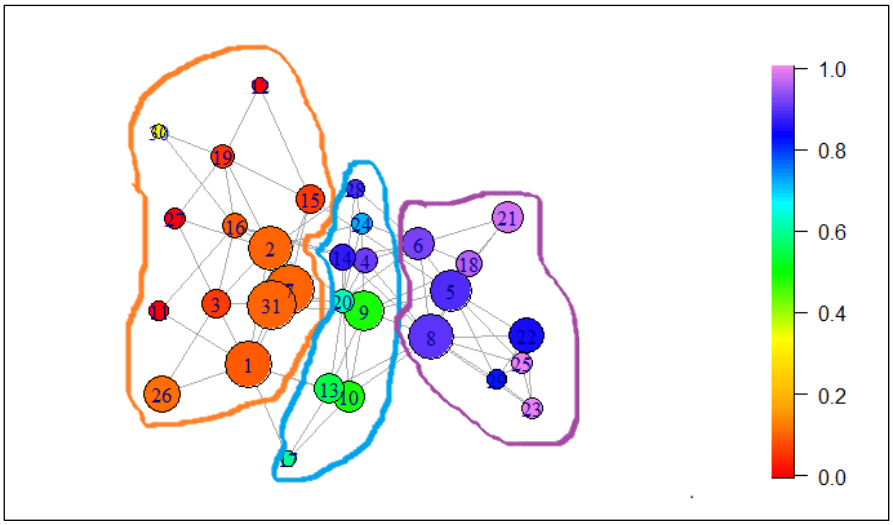
Στη συνέχεια υπολογίστηκαν οι πιθανότητες αθέτησης για το testing setκαι η απόδοση της λογιστικής παλινδρόμησης αξιολογήθηκε με χρήση του AUC, το οποίο βρέθηκε

94.02%. Το στατιστικό KS προέκυψε 0.7812618 (και η αντίστοιχη τιμή του t βρέθηκε 0.3732).

Τέλος, πραγματοποιήσαμε cross validation διαιρώντας το αρχικό dataset σε training και testing set 50 φορές, χρησιμοποιώντας την εντολή set.seed ακριβώς όπως και για τα δύο τοπολογικά μοντέλα. Το training set αποτελείτο από το 70% των παρατηρήσεων και η αναλογία κακών δανειζόμενων ήταν 0.44, όπως και στο αρχικό σύνολο δεδομένων. Η μέση τιμή των AUCs βρέθηκε 0.9295 και η τυπική απόκλισή τους 0.0163.

Ball Mapper και λογιστική παλινδρόμηση

Ως ένα τελευταίο πείραμα δοκιμάσαμε να συνδυάσουμε τη λογιστική παλινδρόμηση με το Ball Mapper. Πιο συγκεκριμένα, σκοπός ήταν να χωρίσουμε τα δεδομένα σε ομάδες με παρόμοια συμπεριφορά (ως προς την αθέτηση) και για κάθε μία από αυτές να εκπαιδεύσουμε ένα μοντέλο λογιστικής παλινδρόμησης.
Χρησιμοποιήσαμε ε = 1.05 και χωρίσαμε τα δεδομένα στις ομάδες του παρακάτω σχήματος.



Η Ομάδα 1 (πορτοκαλί) αποτελείται από 240 παρατηρήσεις και η αναλογία «κακών» είναι 0.09583333. Η Ομάδα 2 (μπλε) περιλαμβάνει 127 παρατηρήσεις με αναλογία «κακών» 0.5984252 και η Ομάδα 3 (μωβ) έχει 149 παρατηρήσεις και αναλογία 0.8993289. Επιπλέον, οι ομάδες έχουν ανά δύο μη κενές τομές ενώ δεν υπάρχουν παρατηρήσεις που να ανήκουν ταυτόχρονα και στις τρεις.

Επομένως, θα εκπαιδεύσουμε συνολικά 6 μοντέλα λογιστικής παλινδρόμησης: ένα για κάθε ομάδα και ένα για κάθε ένωση δύο ομάδων. Όταν έρχεται μία καινούρια παρατήρηση εντοπίζουμε τον κοντινότερο γείτονά της και προσδιορίζουμε την ομάδα στην οποία ανήκει. Αν ανήκει σε μία μόνο ομάδα εφαρμόζουμε το αντίστοιχο μοντέλο

λογιστικής παλινδρόμησης που έχει φτιαχτεί πάνω στο συγκεκριμένο σύνολο και αποδίδουμε την πιθανότητα αθέτησης βάσει αυτού. Εάν ανήκει στην τομή δύο ομάδων, υπολογίζουμε την πιθανότητα αθέτησης χρησιμοποιώντας το μοντέλο που έχει εκπαιδευτεί στην ένωση τους.

Στους πίνακες 10-15 του παραρτήματος παρουσιάζονται αναλυτικά οι εκτιμήσεις των συντελεστών, τα τυπικά σφάλματά τους, η τιμή του στατιστικού z και το p-value του ελέγχου σημαντικότητας και για τα 6 μοντέλα λογιστικής παλινδρόμησης.

Το AUC βρέθηκε 94.21%.

7. Συμπεράσματα και προβληματισμοί

Η απόδοση και των δύο τοπολογικών προσεγγίσεων για την εκτίμηση πιθανοτήτων αθέτησης κρίνεται ικανοποιητική. Παρόλα αυτά η απόδοση της λογιστικής παλινδρόμησης ήταν λίγο καλύτερη, τουλάχιστον για το συγκεκριμένο σύνολο δεδομένων και τα συγκεκριμένα μέτρα αξιολόγησης. Από την άλλη, το γράφημα Ball Mapper παρείχε χρήσιμες πληροφορίες για τον τρόπο με τον οποίο οι επεξηγηματικές μεταβλητές χωρίζουν τους δανειολήπτες σε γειτονιές «κακών» και «καλών» και πως συνδέονται αυτές οι γειτονιές μεταξύ τους αλλά και για τις περιπτώσεις στις οποίες άτομα με παρόμοια χαρακτηριστικά είχαν τελικά διαφορετική έκβαση. Ειδικά οι γειτονιές που χαρακτηρίζονται από αμφισημία έχουν ιδιαίτερο ενδιαφέρον και η ανακάλυψή τους βοηθά στην περαιτέρω μελέτη και έρευνα για το πως και αν μπορούμε τελικά να διαχωρίσουμε άτομα με παρόμοια χαρακτηριστικά. Ενδεχομένως, μία πιο κοντινή ματιά σε αυτές τις περιπτώσεις, θα μπορούσε να βελτιώσει την προβλεπτική ικανότητα των μοντέλων. Επιπλέον, η δυνατότητα οπτικοποίησης που παρέχει το Ball Mapper επιτρέπει να προταθούν στον δανειολήπτη πιθανές κατευθύνσεις βελτίωσης κάποιων από τις μεταβλητές του προκειμένου να μετακινηθεί σε μία πιο ευνοϊκή περιοχή του χώρου των χαρακτηριστικών.

Επιπρόσθετα, η κατασκευή διαφορετικών μοντέλων λογιστικής παλινδρόμησης για τις διάφορες περιοχές που παρατηρήθηκαν στο γράφημα Ball Mapper βελτίωσε την απόδοσή της – έστω και λίγο. Αυτό είναι ενθαρρυντικό και επιβεβαιώνει τη χρησιμότητα της μελέτης του σχήματος των δεδομένων. Φυσικά, η προσέγγισή μας ήταν σε πρωταρχικό στάδιο επομένως, δεν μπορούμε να εξάγουμε ασφαλή συμπεράσματα προτού μελετηθεί περισσότερο. Και τα δύο τοπολογικά μοντέλα πρέπει να δοκιμαστούν και σε άλλα σύνολα δεδομένων, να συνδυαστούν με επιλογή χαρακτηριστικών και τελικά, να συγκριθούν με περισσότερα μοντέλα προκειμένου να αποκτήσουμε μια πιο ολοκληρωμένη εικόνα για τις δυνατότητες και τις αδυναμίες τους. Ακόμη, αν και το AUC αποτελεί ένα ισχυρό μέτρο απόδοσης έχουν διατυπωθεί επιφυλάξεις και έχουν προταθεί εναλλακτικά μέτρα απόδοσης όπως το H-Measure το οποία σύμφωνα με τον δημιουργό του θεραπεύει αδυναμίες του AUC. Συνεπώς, θα ήταν ενδιαφέρον να δούμε τα παραπάνω χρησιμοποιώντας ως μέτρο απόδοσης και το H-Measure.

Είναι σίγουρα θετικό ότι το ανακάτεμα των δεδομένων δεν επηρέασε (σημαντικά) την απόδοση του Ball Mapper (όπως αυτή αξιολογήθηκε με μέτρο το AUC). Όσον αφορά την επιλογή της παραμέτρου ε δεν παρατηρήθηκαν μεγάλες διαφορές στη διαχωριστική ικανότητα και για τα δύο τοπολογικά μοντέλα. Αυτό δημιουργεί μια σχετική ευελιξία και σιγουριά για την επιλογή της παραμέτρου ε και στις δύο περιπτώσεις, αφού κοντινές τιμές δεν επιφέρουν σημαντικές αλλαγές στο AUC. Εντούτοις, θα ήταν σημαντική και σίγουρα πιο αξιόπιστη μία θεωρητικά πιο θεμελιωμένη απάντηση στο

ζήτημα της επιλογής του ε. Ειδικά για το δεύτερο τοπολογικό μοντέλο, η παράμετρος του οποίου έχει άμεση σχέση με το persistence diagram. Περαιτέρω έρευνας χρήζουν και οι κύκλοι που παρατηρήθηκαν στο persistence diagram. Αυτό που αποτελεί μεγάλο πρόβλημα και για τις δύο τοπολογικές προσεγγίσεις είναι ο υπολογιστικός χρόνος που απαιτείται.

Τέλος, θα είχε ιδιαίτερο ενδιαφέρον να μελετηθεί το ίδιο θέμα χρησιμοποιώντας όμως τον κλασσικό αλγόριθμο Mapper ο οποίος δίνει τη δυνατότητα οπτικοποίησης των δεδομένων μέσα από διαφορετικές οπτικές, δηλαδή συναρτήσεις φίλτρου οι οποίες μάλιστα μπορούν να απεικονίζονται σε μετρικούς χώρους πέραν των πραγματικών αριθμών.

Βιβλιογραφία

A. Markov, Z. Seleznyova , V. Lapshin. «Credit scoring methods: Latest trends and points to consider.» *The Journal of Finance and Data Science* , 2022.

Carlsson, Gunnar. «Topology and Data.» *Bulletin of the American Mathematical Society*, 2009: 255-308.

Dlotko, Pawel. «Ball mapper: a shape summary for topological data analysis.» 2019.

Frédéric Chazal, Bertrand Michel. «An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists.» 2021.

Gero, Szepannek. « On the practical relevance of modern machine learning algorithms for credit scoring applications.» *WIAS Report Series*, 2017: 88-96.

Gurjeet Singh, Facundo Mémoli,Gunnar Carlsson. «Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition.» *The Eurographics Association* , 2007.

Hatcher, Allen. *Algebraic Topology.* Cambridge University Press, 2001.

Herbert Edelsbrunner, David Letscher,and Afra Zomorodian. «Topological Persistence and Simplification.» *Discrete and Computational Geometry* , 2002: 511-533.

Kantardzic, Mehmed. «Data Mining: Concepts, Models, Methods and Algorithms.» 184-185. 2020.

Lean Yu, Shouyang Wang, Kin Keung Lai, Ligang Zhou. *Bio-Inspired Credit Risk Analysis: Computational Intelligence with Support Vector Machines.* Springer, 2008.

Maria Rocha Sousa, João Gama, Elísio Brandão. «Introducing Time-Changing Economics into Credit Scoring.» *FEP Working Papers*, 2013.

Michael Bücker, Gero Szepannek, Alicja Gosiewska, Przemyslaw Biecek. «Transparency, auditability, and explainability of machine learning in credit scoring.» *Journal of the Operational Research Society*, 2022: 70-90.

Michel, Bertrand. «A Statistical Approach to Topological Data Analysis.» *UPMC Université Paris VI*, 2015.

Mohammed J. Zaki, Wagner Meira, Jr. «Data Mining and Machine Learning: Fundamental Concepts and Algorithms.» 554-561, 568-570, 623-630. 2020.

Munch, Elizabeth. «A user's guide to topological data analysis.» *Journal of Learning Analytics*, 2017: 47-61.

Munkres, James R. *Elements of Algebraic Topology.* Addison-Wesley Publishing Company, 1984.

Nina Otter, Mason A Porter, Ulrike Tillmann,Peter Grindrod,Heather A Harrington. «A roadmap for the computation of persistent homology.» *EPJ Data Science*, 2017.

P. Y. Lum, G. Singh,A. Lehman,T. Ishkanov, M. Vejdemo-Johansson,M. Alagappan,J. Carlsson,G. Carlsson. «Extracting insights from the shape of complex data using topology.» *Scientific Reports*, 2013.

Pawel Dlotko, Wanling Qiu, Simon Rudkin. «Topological Data Analysis Ball Mapper for Finance.» 2022.

Siddiqi, Naeem. *Intelligent Credit Scoring.* John Wiley & Sons Inc., 2017.

Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, Lyn C. Thomas. «Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research.» *European Journal of Operational Research*, 2013.

Tamal Krishna Dey, Yusu Wang. *Computational Topology for Data Analysis.* Cambridge University Press, 2022.

Wanling Qiu, Simon Rudkin, Pawel Dlotko. «Refining understanding of corporate failure through a topological data analysis mapping of Altman's Z-score model.» *Expert Systems with Applications*, 2020.

Wei Li, Florentina Paraschiv, Georgios Sermpinis. «A data-driven explainable case-based reasoning approach for financial risk detection.» *Quantitative Finance*, 2022.

X. Dastile, T. Celik and M. Potsane. «Statistical and machine learning models in credit scoring: A systematic literature survey.» *Applied Soft Computing Journal*, 2020.

Zomorodian, Afra J. «Topological Data Analysis.» *Advances in Applied and Computational Topology*, 2012.

—. *Topology for Computing.* 2005.

Παράρτημα

Πίνακας 3: Μέσο, ελάχιστο και μέγιστο AUC για τον καθορισμό της παραμέτρου του μοντέλου I (τρόπος A)

ε	Μέσο AUC	ελάχιστο	μέγιστο	Τυπική απόκλιση
0.80	0.9164007	0.8665327	0.9634520	0.02094288
0.81	0.9176196	0.8656716	0.9606774	0.02043434
0.82	0.9178033	0.8638538	0.9629736	0.02122875
0.83	0.9172847	0.8740911	0.9649828	0.02171710
0.84	0.9152813	0.8699770	0.9618255	0.02147547
0.85	0.9143819	0.8685419	0.9604860	0.02201237
0.86	0.9145752	0.8757176	0.9604860	0.02050694
0.87	0.9149196	0.8773440	0.9603904	0.02058033
0.88	0.9148565	0.8730387	0.9601033	0.02054978
0.89	0.9138481	0.8738041	0.9596250	0.02086850
0.90	0.9129315	0.8740911	0.9590509	0.02138629
0.91	0.9149445	0.8811711	0.9607731	0.02080684
0.92	0.9140662	0.8805013	0.9604860	0.02033670
0.93	0.9142250	0.8781095	0.9599120	0.02069616
0.94	0.9148029	0.8753349	0.9598163	0.02016703
0.95	0.9136663	0.8716992	0.9661309	0.02020960
0.96	0.9142212	0.8778224	0.9658439	0.02030352
0.97	0.9139514	0.8780138	0.9670876	0.02007395
0.98	0.9145197	0.8778224	0.9645044	0.01911913
0.99	0.9143628	0.8781095	0.9645044	0.01910830
1.00	0.9140681	0.8746651	0.9623995	0.01895713
1.01	0.9141007	0.8691160	0.9623039	0.02053824
1.02	0.9131267	0.8615576	0.9623039	0.02017165
1.03	0.9091944	0.8532338	0.9579028	0.02130665
1.04	0.9080941	0.8592614	0.9543628	0.01978492
1.05	0.9061998	0.8623230	0.9502488	0.01915162
1.06	0.9044948	0.8554344	0.9560850	0.02127413
1.07	0.9010926	0.8539036	0.9466131	0.02093970
1.08	0.8985572	0.8454841	0.9550325	0.02167419
1.09	0.8976043	0.8343858	0.9575201	0.02330908
1.10	0.8924072	0.8451014	0.9511098	0.02147158
1.11	0.8915844	0.8475890	0.9380980	0.02076759
1.12	0.8877937	0.8490241	0.9308266	0.02065084
1.13	0.8845637	0.8394566	0.9260429	0.02126198
1.14	0.8854133	0.8399349	0.9301569	0.02086982

1.15	0.8862706	0.8441447	0.9358974	0.02220849
1.16	0.8888021	0.8468236	0.9385763	0.02219242
1.17	0.8890165	0.8406047	0.9382893	0.02362191
1.18	0.8830635	0.8245312	0.9393418	0.02769617
1.19	0.8799330	0.8267317	0.9446039	0.02818494
1.20	0.8781209	0.8264447	0.9399158	0.02783910
1.21	0.8811041	0.8356295	0.9393418	0.02678845
1.22	0.8739170	0.8078837	0.9350364	0.03021066
1.23	0.8752641	0.8098928	0.9332185	0.02930815
1.24	0.8739246	0.8043437	0.9303483	0.02899399
1.25	0.8731822	0.8066399	0.9235553	0.03054219
1.26	0.8655549	0.7895140	0.9387677	0.03640374
1.27	0.8684252	0.8053004	0.9393418	0.03544240
1.28	0.8691427	0.7988902	0.9393418	0.03553804
1.29	0.8678550	0.7833907	0.9390547	0.03296206
1.30	0.8621642	0.7877918	0.9451780	0.03766760
1.31	0.8630004	0.7942021	0.9469958	0.03735388
1.32	0.8611003	0.7758324	0.9441255	0.04012552
1.33	0.8659147	0.7920972	0.9495790	0.04023349
1.34	0.8638749	0.7910448	0.9413509	0.03860405
1.35	0.8624110	0.7851129	0.9373326	0.03929514
1.36	0.8602909	0.7829124	0.9402028	0.03960698
1.37	0.8580827	0.7721010	0.9380980	0.04373142
1.38	0.8556793	0.7749713	0.9443169	0.04437511
1.39	0.8465901	0.7692308	0.9419250	0.04723440
1.40	0.8443226	0.7656908	0.9409682	0.04629985
1.41	0.8439533	0.7566973	0.9433601	0.04803067
1.42	0.8425775	0.7600459	0.9426904	0.04866712
1.43	0.8445293	0.7544011	0.9398201	0.04851927
1.44	0.8548374	0.7717183	0.9438385	0.04277080
1.45	0.8501454	0.7642556	0.9483352	0.04661145
1.46	0.8444087	0.7647340	0.9370455	0.04756429
1.47	0.8459740	0.7634902	0.9244164	0.04478893
1.48	0.8417930	0.7641600	0.9162840	0.04309802
1.49	0.8429832	0.7649254	0.9207807	0.04496553
1.50	0.8414007	0.7666475	0.9215461	0.04103228
1.51	0.8410486	0.7519135	0.9173364	0.04175411
1.52	0.8689457	0.7838691	0.9244164	0.03028119
1.53	0.8735170	0.8167815	0.9216418	0.02832179
1.54	0.8723440	0.7868351	0.9330272	0.03130402
1.55	0.8694546	0.7870264	0.9207807	0.02919396
1.56	0.8726770	0.7937237	0.9341753	0.03038909
1.57	0.8776904	0.8089361	0.9412553	0.03111997
1.58	0.8792269	0.8109453	0.9368542	0.02940186
1.59	0.8790796	0.8100842	0.9497704	0.02932820
1.60	0.8801110	0.7817643	0.9464217	0.03128070
1.61	0.8762821	0.7706659	0.9387677	0.03571136

1.62	0.8772254	0.8114237	0.9338883	0.02891651
1.63	0.8721546	0.8003253	0.9159013	0.02681703
1.64	0.8699464	0.7988902	0.9142748	0.02812461
1.65	0.8645848	0.8026215	0.9112132	0.02870412
1.66	0.8669020	0.7998469	0.9230769	0.02911588
1.67	0.8611596	0.8003253	0.9150402	0.02980800
1.68	0.8586376	0.7876961	0.9089170	0.02918338
1.69	0.8608611	0.7865480	0.9499617	0.03239226
1.70	0.8521757	0.7923842	0.9495790	0.03457223
1.71	0.8584118	0.7902794	0.9413509	0.03298556
1.72	0.8535610	0.7732491	0.9310180	0.03462821
1.73	0.8609721	0.7819556	0.9241294	0.03270270
1.74	0.8451378	0.7737275	0.9268083	0.03280135
1.75	0.8419537	0.7654038	0.9192499	0.03596729
1.76	0.8496537	0.7853999	0.9276693	0.03377172
1.77	0.8466380	0.7809989	0.9248948	0.03512692
1.78	0.8276942	0.6524110	0.9278607	0.04220195
1.79	0.8309874	0.6396862	0.9225985	0.04432812

Πίνακας 4:Μέσο, ελάχιστο και μέγιστο AUC για τον καθορισμό της παραμέτρου του μοντέλου I (τρόπος Β)

ε	μέσο	ελάχιστο	μέγιστο	τυπική απόκλιση
0.80	0.9097072	0.8679679	0.9536931	0.02075092
0.81	0.9118427	0.8750478	0.9514925	0.01979004
0.82	0.9131037	0.8767700	0.9534060	0.01996746
0.83	0.9112687	0.8772484	0.9596250	0.02130292
0.84	0.9084788	0.8616533	0.9583812	0.02266848
0.85	0.9076330	0.8573479	0.9574244	0.02343424
0.86	0.9083008	0.8723689	0.9566590	0.02170138
0.87	0.9108860	0.8732300	0.9590509	0.02157347
0.88	0.9117719	0.8713165	0.9614428	0.02174644
0.89	0.9103349	0.8673938	0.9611558	0.02187468
0.90	0.9105281	0.8724646	0.9592423	0.02182615
0.91	0.9104191	0.8734214	0.9542671	0.02193859
0.92	0.9096383	0.8749522	0.9551282	0.02163883
0.93	0.9078569	0.8740911	0.9547455	0.02221474
0.94	0.9076005	0.8636625	0.9545542	0.02178105
0.95	0.9058706	0.8609835	0.9606774	0.02142814
0.96	0.9064485	0.8690203	0.9602947	0.02159652
0.97	0.9062706	0.8709338	0.9615385	0.02156979

0.98	0.9079698	0.8735170	0.9589552	0.02049494
0.99	0.9089284	0.8738041	0.9589552	0.02009824
1.00	0.9087926	0.8703597	0.9560850	0.02006981
1.01	0.9094527	0.8653846	0.9565633	0.02190471
1.02	0.9081382	0.8584003	0.9571374	0.02157251
1.03	0.9057271	0.8556257	0.9528320	0.02266432
1.04	0.9038385	0.8556257	0.9440299	0.02097249
1.05	0.9012897	0.8500765	0.9393418	0.02036025
1.06	0.9012514	0.8473020	0.9482396	0.02190206
1.07	0.8977095	0.8587830	0.9391504	0.02094430
1.08	0.8951206	0.8525641	0.9617298	0.02210663
1.09	0.8955434	0.8454841	0.9567547	0.02227248
1.10	0.8898948	0.8486414	0.9489093	0.02127308
1.11	0.8887715	0.8525641	0.9349407	0.02083784
1.12	0.8859663	0.8439533	0.9324531	0.02097871
1.13	0.8828951	0.8394566	0.9272866	0.02199890
1.14	0.8834883	0.8400306	0.9275737	0.02110350
1.15	0.8833295	0.8376387	0.9358974	0.02285646
1.16	0.8862132	0.8391695	0.9385763	0.02304258
1.17	0.8871450	0.8366820	0.9363758	0.02424209
1.18	0.8816628	0.8161117	0.9353234	0.02766247
1.19	0.8782989	0.8267317	0.9422120	0.02797807
1.20	0.8765940	0.8276885	0.9370455	0.02761642
1.21	0.8795829	0.8385955	0.9369499	0.02666364
1.22	0.8722790	0.8021431	0.9330272	0.02995735
1.23	0.8749062	0.8074053	0.9300612	0.02851920
1.24	0.8737103	0.8026215	0.9290088	0.02770368
1.25	0.8725316	0.8066399	0.9234596	0.02953330
1.26	0.8650019	0.7895140	0.9362801	0.03505559
1.27	0.8685553	0.8028129	0.9405855	0.03496307
1.28	0.8698641	0.8008993	0.9396288	0.03507543
1.29	0.8685859	0.7833907	0.9357061	0.03296864
1.30	0.8627057	0.7877918	0.9410639	0.03784162
1.31	0.8635075	0.7942021	0.9423077	0.03794813
1.32	0.8625182	0.7746843	0.9429774	0.04042126
1.33	0.8655224	0.7859740	0.9402028	0.04094474
1.34	0.8635228	0.7870264	0.9333142	0.03994116
1.35	0.8623957	0.7896096	0.9373326	0.04007971
1.36	0.8596843	0.7743016	0.9366628	0.04033892
1.37	0.8564351	0.7685610	0.9371412	0.04433280
1.38	0.8539744	0.7713356	0.9449866	0.04467731
1.39	0.8452162	0.7664562	0.9425947	0.04686609
1.40	0.8435055	0.7635859	0.9416380	0.04659310
1.41	0.8432434	0.7547838	0.9399158	0.04785527
1.42	0.8421374	0.7567930	0.9392461	0.04771892
1.43	0.8438863	0.7515308	0.9398201	0.04784697
1.44	0.8540471	0.7696135	0.9438385	0.04224203

1.45	0.8511175	0.7642556	0.9483352	0.04582613
1.46	0.8455205	0.7711443	0.9370455	0.04642477
1.47	0.8469575	0.7700918	0.9244164	0.04350080
1.48	0.8427574	0.7598546	0.9162840	0.04210662
1.49	0.8437562	0.7626292	0.9207807	0.04433070
1.50	0.8422388	0.7646383	0.9215461	0.04017972
1.51	0.8417222	0.7562189	0.9173364	0.04105218
1.52	0.8681037	0.7810945	0.9212591	0.03024302
1.53	0.8728856	0.8133372	0.9208764	0.02839651
1.54	0.8720991	0.7942977	0.9313050	0.03122596
1.55	0.8687811	0.7870264	0.9192499	0.02952143
1.56	0.8724244	0.7937237	0.9328358	0.02993358
1.57	0.8776081	0.8089361	0.9412553	0.03084373
1.58	0.8794680	0.8179296	0.9368542	0.02939889
1.59	0.8793494	0.8176426	0.9497704	0.02931891
1.60	0.8801397	0.7869307	0.9464217	0.03104095
1.61	0.8762055	0.7769805	0.9387677	0.03507256
1.62	0.8769881	0.8080750	0.9338883	0.02842763
1.63	0.8716724	0.7978377	0.9141791	0.02658281
1.64	0.8696269	0.7965940	0.9140834	0.02780318
1.65	0.8643819	0.8005166	0.9091083	0.02849191
1.66	0.8669173	0.7975507	0.9230769	0.02932885
1.67	0.8610352	0.8003253	0.9150402	0.02997680
1.68	0.8585228	0.7879832	0.9089170	0.02919943
1.69	0.8612954	0.7892269	0.9499617	0.03264930
1.70	0.8528626	0.7981248	0.9495790	0.03429298
1.71	0.8597704	0.7899923	0.9413509	0.03263031
1.72	0.8538117	0.7688481	0.9310180	0.03538182
1.73	0.8609682	0.7809032	0.9263299	0.03326621
1.74	0.8446192	0.7737275	0.9268083	0.03336612
1.75	0.8422005	0.7654038	0.9192499	0.03601734
1.76	0.8504956	0.7853999	0.9264256	0.03381429
1.77	0.8473708	0.7809989	0.9248948	0.03507309
1.78	0.8288136	0.6551856	0.9278607	0.04106393
1.79	0.8320494	0.6425564	0.9225985	0.04315962

Πίνακας 5: Cross validation για το τοπολογικό μοντέλο I

Μέθοδος Α		**Μέθοδος Β**	
AUC	ε	AUC	ε
0.9046	1.01	0.9027	1.01
0.9114103	0.84	0.9030785	0.80
0.9084447	0.80	0.9022783	0.92
0.9117869	0.80	0.9169648	0.92
0.8887215	0.82	0.8886744	0.82
0.8901808	1.04	0.9093862	0.92
0.9172472	0.86	0.9155056	0.92

0.8627377	0.80	0.8684805	0.80
0.9105630	0.84	0.8992186	0.88
0.8854735	0.80	0.8874976	0.80
0.8727641	1.01	0.8818019	1.01
0.9086330	0.84	0.9106571	0.88
0.9098098	1.02	0.9103747	1.01
0.8938995	0.82	0.9142346	1.04
0.8814724	1.01	0.8809546	1.01
0.8752118	0.80	0.8747882	0.80
0.9093391	0.85	0.9147995	0.82
0.9250612	1.02	0.9222839	1.02
0.9038787	0.86	0.9070326	0.94
0.8994540	0.87	0.8947468	0.89
0.8643852	0.80	0.8684805	1.01
0.9435135	0.81	0.9324515	0.89
0.8809546	1.03	0.8800603	1.03
0.8884862	1.02	0.9001130	1.02
0.8932875	0.92	0.8978535	0.92
0.8758708	0.82	0.8804368	0.81
0.9032668	0.81	0.8997364	1.01
0.8609960	0.93	0.8718697	0.92
0.8825551	1.04	0.8777066	1.04
0.8974299	0.87	0.8890510	0.82
0.9002071	0.91	0.9158351	0.93
0.9024666	0.82	0.9005837	1.01
0.9329693	1.04	0.9325927	1.02
0.9324044	1.04	0.9321220	1.03
0.9142346	0.82	0.9052909	0.82
0.9114103	0.88	0.9134344	0.88
0.8987008	0.94	0.8944643	1.03
0.9010544	0.82	0.8987479	0.80
0.9170119	0.82	0.9277443	0.81
0.8954058	0.82	0.8847204	1.01
0.8775184	1.04	0.8822726	1.04
0.9038317	1.02	0.9204011	1.02
0.9155526	0.90	0.9179062	0.90
0.9203540	0.81	0.9220957	0.81
0.9091037	0.80	0.9188947	0.80
0.8981359	1.03	0.9072679	1.04
0.8878272	0.87	0.9058558	0.82
0.9155997	0.94	0.9169177	0.82
0.8504519	0.81	0.8536528	0.81
0.9236961	1.03	0.9308511	1.03

Πίνακας 6: Τα AUCs που προέκυψαν από το ανακάτεμα του training set για ε=0.82 (μοντέλο Ι)

	AUC (τρόπος Α)	AUC (τρόπος Β)
1	0.8848522	0.8892099
2	0.8844259	0.8875995
3	0.8844733	0.8990621
4	0.8697897	0.8771315
5	0.8738158	0.8880258
6	0.8898257	0.9085828
7	0.8875047	0.8911993
8	0.8822471	0.8873153
9	0.8858943	0.8941834
10	0.8848522	0.8921466
11	0.8812524	0.9037041
12	0.8938045	0.9064987
13	0.8808261	0.8968359
14	0.8800682	0.8970254
15	0.8872205	0.9017147
16	0.8862258	0.8866048
17	0.8849469	0.9004358
18	0.8788367	0.8974517
19	0.8908678	0.9067829
20	0.8814892	0.8933782
21	0.8854206	0.8855627
22	0.8848522	0.9014305
23	0.8846628	0.9001989
24	0.8830997	0.8939939
25	0.8780788	0.9019041
26	0.8850890	0.8970254
27	0.8854206	0.9001042
28	0.8794051	0.9016199
29	0.8840470	0.8972149
30	0.8817734	0.8947992
31	0.8825786	0.9062145
32	0.8825313	0.9045567
33	0.8822471	0.8964570
34	0.8861785	0.8977359
35	0.8795945	0.9053145
36	0.8759473	0.8929519
37	0.8880731	0.9066881
38	0.8851364	0.9069723
39	0.8804945	0.9051724

40	0.8866995	0.8922887
41	0.8806366	0.9056461
42	0.8836207	0.8999147
43	0.8842838	0.8992990
44	0.8932361	0.9106669
45	0.8841891	0.9067355
46	0.8834312	0.8873626
47	0.8929992	0.9105722
48	0.8840944	0.8937097
49	0.8848522	0.8983990
50	0.8744790	0.8878363

Πίνακας 7: Μέσο, ελάχιστο, μέγιστο και τυπική απόκλιση των AUCs που υπολογίστηκαν για τον καθορισμό της παραμέτρου ε του τοπολογικού μοντέλου ΙΙ

ε	μέσο	μέγιστο	ελάχιστο	τυπική απόκλιση
0.71	0.9111902	0.9478569	0.8742824	0.01949199
0.72	0.9125105	0.9480482	0.8713165	0.02022244
0.73	0.9091466	0.9440299	0.8628971	0.02059240
0.74	0.9097589	0.9455607	0.8671068	0.02059474
0.75	0.9098507	0.9468044	0.8716035	0.02146001
0.76	0.9098144	0.9522579	0.8717949	0.02185381
0.77	0.9095101	0.9510142	0.8672981	0.02097061
0.78	0.9089131	0.9501531	0.8652889	0.02210781
0.79	0.9091198	0.9505358	0.8658630	0.02178598
0.80	0.9094910	0.9509185	0.8638538	0.02156405
0.81	0.9080387	0.9475698	0.8630884	0.02166397
0.82	0.9091140	0.9476655	0.8714122	0.02119236
0.83	0.9087294	0.9590509	0.8697857	0.02173631
0.84	0.9088289	0.9588595	0.8691160	0.02176042
0.85	0.9085668	0.9601033	0.8675852	0.02184301
0.86	0.9099292	0.9609644	0.8675852	0.02118988
0.87	0.9100823	0.9616341	0.8676808	0.02129481
0.88	0.9095886	0.9612514	0.8673938	0.02129640
0.89	0.9090719	0.9601033	0.8645235	0.02126771
0.90	0.9088021	0.9601033	0.8628971	0.02143340
0.91	0.9108094	0.9586682	0.8733257	0.02043288
0.92	0.9114581	0.9592423	0.8739954	0.02039208
0.93	0.9125010	0.9611558	0.8752392	0.02014905
0.94	0.9123613	0.9610601	0.8743781	0.02004955
0.95	0.9122675	0.9593379	0.8743781	0.01999566
0.96	0.9122235	0.9591466	0.8748565	0.02005855

0.97	0.9119996	0.9595293	0.8748565	0.02021943
0.98	0.9116246	0.9595293	0.8737084	0.02027288
0.99	0.9114160	0.9597206	0.8737084	0.02041407
1.00	0.9114045	0.9597206	0.8741868	0.02040606
1.01	0.9102794	0.9549369	0.8651933	0.02153215
1.02	0.9100842	0.9589552	0.8596441	0.02188174
1.03	0.9095331	0.9606774	0.8702641	0.02049534
1.04	0.9142040	0.9636433	0.8676808	0.02082709
1.05	0.9151282	0.9614428	0.8651933	0.02127887
1.06	0.9156487	0.9600077	0.8649062	0.02115688
1.07	0.9139265	0.9580941	0.8634711	0.02049546
1.08	0.9129717	0.9565633	0.8601225	0.02057918
1.09	0.9123976	0.9536931	0.8593571	0.02068834
1.10	0.9109568	0.9507271	0.8630884	0.02055114
1.11	0.9109568	0.9522579	0.8637581	0.02070183
1.12	0.9098374	0.9528320	0.8593571	0.02076210
1.13	0.9152564	0.9601033	0.8672981	0.02080895
1.14	0.9164581	0.9644087	0.8755262	0.01992063
1.15	0.9158706	0.9619212	0.8697857	0.02007198
1.16	0.9157463	0.9646958	0.8673938	0.02031378
1.17	0.9158783	0.9654612	0.8688289	0.02046715
1.18	0.9173670	0.9672790	0.8720819	0.01957162
1.19	0.9159816	0.9675660	0.8700727	0.01984628
1.20	0.9166418	0.9672790	0.8717949	0.01967664

Πίνακας 8: Τα AUCs που προέκυψαν κατά το cross validation του μοντέλου ΙΙ

	AUC	ε που επιλέχθηκε
1	0.9115515	1.18
2	0.9281680	1.16
3	0.8951704	1.04
4	0.9243080	1.02
5	0.8974299	1.20
6	0.9218132	1.18
7	0.9202128	0.76
8	0.8930051	1.20
9	0.9203540	1.18
10	0.8830729	0.76
11	0.8846262	1.14
12	0.8981359	0.88
13	0.9248729	1.04
14	0.9078328	1.02

15	0.8950763	1.06
16	0.9012898	1.20
17	0.9197891	1.20
18	0.9348522	1.06
19	0.8938995	1.08
20	0.8999718	1.18
21	0.8717285	1.04
22	0.9295801	1.16
23	0.8939465	1.06
24	0.8975711	1.16
25	0.9041141	1.20
26	0.8813783	0.72
27	0.9026549	0.74
28	0.8721051	0.72
29	0.9077857	1.04
30	0.8942290	1.04
31	0.8954528	1.18
32	0.8962531	0.92
33	0.9510921	1.20
34	0.9165411	1.20
35	0.9066089	0.72
36	0.9400772	1.20
37	0.9214366	1.16
38	0.8855206	1.20
39	0.9203069	1.14
40	0.8755413	1.16
41	0.9134815	1.04
42	0.9045848	1.04
43	0.9132932	1.18
44	0.9261439	1.18
45	0.8898042	1.18
46	0.9181416	1.04
47	0.9145641	1.18
48	0.9146583	1.16
49	0.8856618	1.20
50	0.9304745	1.18

Πίνακας 9: Συντελεστές, τυπικό σφάλμα και p-value για τη λογιστική παλινδρόμηση

β_i	Τυπικό σφάλμα	p-value
-4.849201	0.848159	1.08*10^(-8)
0.002103	0.326320	0.994859
-0.318368	0.864181	0.712572
-0.643560	0.935680	0.491579
1.189278	0.749877	0.112748
2.286938	0.631756	0.000295
-0.385466	0.709653	0.587009
2.608412	1.599829	0.103010
3.368258	0.362071	< 2e-16
0.637377	0.407383	0.117685
6.430260	4.328794	0.137421
-0.381430	0.315242	0.226295
2.057305	1.061314	0.052568
-2.998834	2.099075	0.153106
37.539045	17.582002	0.032754

Πίνακας 10: Λογιστική παλινδρόμηση για την ομάδα 1

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.4263958	1.4772320	-3.6733538	2.393877e-04
training_set[group1,]$V1	0.3761864	0.6266037	0.6003577	5.482679e-01
training_set[group1,]$V2	1.7796213	1.6992985	1.0472682	2.949759e-01
training_set[group1,]$V3	-8.1977157	3.1198775	-2.6275762	8.599559e-03
training_set[group1,]$V4	0.9665614	1.2553748	0.7699385	4.413364e-01
training_set[group1,]$V5	3.5523532	1.2333221	2.8803127	3.972810e-03
training_set[group1,]$V6	0.4111698	1.6066683	0.2559145	7.980168e-01
training_set[group1,]$V7	4.2274911	4.1079622	1.0290969	3.034342e-01
training_set[group1,]$V8	5.0539209	1.1373537	4.4435790	8.847466e-06
training_set[group1,]$V9	-1.3633459	1.2145505	-1.1225107	2.616454e-01
training_set[group1,]$V10	-3.8373048	31.8391238	-0.1205217	9.040699e-01
training_set[group1,]$V11	-0.2248586	0.6219799	-0.3615206	7.177103e-01
training_set[group1,]$V12	1.2953739	1.4720026	0.8800079	3.788550e-01
training_set[group1,	0.5853018	3.4120057	0.1715419	8.637977e-01

]$V13				
training_set[group1,]$V14	40.9045607	38.6669509	1.0578688	2.901153e-01

Πίνακας 11: Λογιστική παλινδρόμηση για την ομάδα 2

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-3.40327249	1.7385509	-1.95753403	0.05028471
training_set[group2,]$V1	0.02067412	0.5995218	0.03448435	0.97249092
training_set[group2,]$V2	-1.56726785	1.2965986	-1.20875333	0.22675763
training_set[group2,]$V3	1.28312032	1.4429317	0.88924539	0.37387122
training_set[group2,]$V4	2.07197643	1.2116530	1.71004112	0.08725827
training_set[group2,]$V5	1.82112144	1.0081633	1.80637541	0.07085972
training_set[group2,]$V6	0.89846323	1.1569580	0.77657380	0.43741026
training_set[group2,]$V7	2.06809833	2.1360230	0.96820042	0.33294429
training_set[group2,]$V8	1.82863205	1.1680044	1.56560374	0.11744141
training_set[group2,]$V9	1.08839323	1.4110091	0.77135806	0.44049473
training_set[group2,]$V10	15.60690003	22.8812731	0.68208181	0.49518723
training_set[group2,]$V11	-0.32376294	0.5001339	-0.64735252	0.51740382
training_set[group2,]$V12	0.47708451	1.5022623	0.31757737	0.75080554
training_set[group2,]$V13	-7.60740911	3.3746276	-2.25429591	0.02417756
training_set[group2,]$V14	125.09384849	56.9292553	2.19735614	0.02799502

Πίνακας 12: Λογιστική παλινδρόμηση για την ομάδα 3

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-25.2860181	1735.771439	-0.01456760	0.98837715
training_set[group3,]$V1	-0.6917431	1.141145	-0.60618355	0.54439287
training_set[group3,]$V2	-1.4940148	2.182093	-0.68467049	0.49355187
training_set[group3,]$V3	-1.4795529	2.315232	-0.63905162	0.52278935

training_set[group3,]$V4	2.4361007	1.879523	1.29612743	0.19493159
training_set[group3,]$V5	1.5422637	1.443977	1.06806696	0.28549031
training_set[group3,]$V6	-3.0482608	1.755127	-1.73677501	0.08242689
training_set[group3,]$V7	13.0594608	6.449413	2.02490687	0.04287693
training_set[group3,]$V8	24.9755392	1735.769769	0.01438874	0.98851984
training_set[group3,]$V9	2.7338050	1.170773	2.33504322	0.01954116
training_set[group3,]$V10	1.7349975	5.255570	0.33012548	0.74130515
training_set[group3,]$V11	-0.8862070	1.042154	-0.85036061	0.39512463
training_set[group3,]$V13	1.3931299	4.958733	0.28094472	0.77875280
training_set[group3,]$V14	129.4931484	100.062606	1.29412129	0.19562353

Πίνακας 13: Λογιστική παλινδρόμηση- Ομάδες 1&2

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-4.37011696	0.9159850	-4.7709483	1.833606e-06
training_set[group1_2,]$V1	-0.08210751	0.3749571	-0.2189784	8.266669e-01
training_set[group1_2,]$V2	-0.10138184	0.9854165	-0.1028822	9.180564e-01
training_set[group1_2,]$V3	-0.74506963	1.0986946	-0.6781408	4.976824e-01
training_set[group1_2,]$V4	0.99941129	0.8322823	1.2008080	2.298257e-01
training_set[group1_2,]$V5	2.43143815	0.7209581	3.3725097	7.448645e-04
training_set[group1_2,]$V6	0.09098168	0.8022225	0.1134120	9.097039e-01
training_set[group1_2,]$V7	1.57047979	1.7672452	0.8886598	3.741859e-01
training_set[group1_2,]$V8	2.96685394	0.3861962	7.6822446	1.563250e-14
training_set[group1_2,]$V9	-0.23787245	0.5664056	-0.4199684	6.745086e-01
training_set[group1_2,]$V10	12.34520188	7.0110522	1.7608201	7.826885e-02
training_set[group1_2,]$V11	-0.24763200	0.3565546	-0.6945136	4.873602e-01
training_set[group1_2,]$V12	1.56252383	1.0447967	1.4955290	1.347764e-01

training_set[group1_2,]$V13	-4.72396495	2.4478756	-1.9298223	5.362885e-02
training_set[group1_2,]$V14	32.63667852	19.5820526	1.6666628	9.558148e-02

Πίνακας 14: Λογιστική παλινδρόμηση-Ομάδες 2&3

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.4772755	1.5013938	-2.9820793	0.002862978
training_set[group2_3,]$V1	-0.1106534	0.4318742	-0.2562168	0.797783471
training_set[group2_3,]$V2	-0.9468456	1.0216529	-0.9267782	0.354041711
training_set[group2_3,]$V3	0.1836844	1.1435622	0.1606248	0.872388937
training_set[group2_3,]$V4	1.7713440	0.9189209	1.9276348	0.053900572
training_set[group2_3,]$V5	1.4871794	0.7723435	1.9255413	0.054161671
training_set[group2_3,]$V6	-0.5171601	0.8685284	-0.5954441	0.551546669
training_set[group2_3,]$V7	3.0581839	1.7917696	1.7067953	0.087860103
training_set[group2_3,]$V8	3.3645279	1.1258993	2.9883027	0.002805316
training_set[group2_3,]$V9	1.6674971	0.5467237	3.0499815	0.002288555
training_set[group2_3,]$V10	1.9704343	4.7981347	0.4106667	0.681316914
training_set[group2_3,]$V11	-0.4174127	0.4038852	-1.0334934	0.301373065
training_set[group2_3,]$V12	1.4322664	1.4203723	1.0083739	0.313274996
training_set[group2_3,]$V13	-3.8830689	2.4359645	-1.5940581	0.110923029
training_set[group2_3,]$V14	58.8289506	31.6226601	1.8603416	0.062837210

Πίνακας 15: Λογιστική παλινδρόμηση- Ομάδες 1&3

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.61805799	1.1395477	-4.0525360	5.066544e-05
training_set[group1_3,]$V1	0.09835513	0.4452829	0.2208823	8.251840e-01
training_set[group1_3,]$V2	1.17570779	1.2479093	0.9421420	3.461199e-01
training_set[group1_3,	-3.23769884	1.3574817	-2.3850773	1.707553e-02

]$V3				
training_set[group1_3,]$V4	1.19633367	1.0091013	1.1855437	2.358025e-01
training_set[group1_3,]$V5	2.76446353	0.8630249	3.2032258	1.358974e-03
training_set[group1_3,]$V6	-1.13263889	1.0479448	-1.0808192	2.797775e-01
training_set[group1_3,]$V7	6.75030578	2.8636671	2.3572244	1.841212e-02
training_set[group1_3,]$V8	4.51274838	0.5786833	7.7983037	6.274487e-15
training_set[group1_3,]$V9	-0.30755977	0.5663308	-0.5430744	5.870786e-01
training_set[group1_3,]$V10	4.00386410	4.8202225	0.8306389	4.061777e-01
training_set[group1_3,]$V11	-0.56706017	0.4572923	-1.2400389	2.149610e-01
training_set[group1_3,]$V12	1.41366867	1.5135685	0.9339971	3.503054e-01
training_set[group1_3,]$V13	-0.63501787	2.7042913	-0.2348186	8.143495e-01
training_set[group1_3,]$V14	41.01878825	22.9604872	1.7864947	7.401920e-02