

Ανάλυση σε Κύριες Συνιστώσες.

Απόστολος Τάψας
Επιβλέπων Καθηγητής Σταύρος Αδάμ

Στον αδερφό μου και τους γονείς μου
που με στηρίζουν σε ότι και κάνω
στην ζωή μου πνευματικά
και οικονομικά.

Στον καθηγητή μου που όχι μόνο με
βοήθησε με της γνώσεις του αββιά
και με την όρεξή του για εργασία.

Στους συμφοιτητές μου και φίλους
μου για την υποστήριξη τους .

Περίληψη

Τάψας Απόστολος
Μαθητής ΤΕΙ Ηπείρου
aptapsas@hotmail.com

Επιβλέπων καθηγητής Αδάμ Σταυρός
Καθηγητής Εφαρμογών ΤΕΙ Ηπείρου
adamsp@teiep.gr

Ζούμε στην εποχή της πληροφορίας, καθημερινά όπου και αν σταθούμε βλέπουμε την πληθώρα πληροφοριών που υπάρχει εκεί έξω. Για παράδειγμα, πληροφορίες ζητάνε και από εμάς όπου και αν πάμε, στην τράπεζα, σε πληρωμές λογαριασμών, σε βενζινάδικα ακόμη και στα super market. Πως μπορούμε όμως να εκμεταλλευτούμε όλη αυτή την πληροφορία που μας δίνεται άφθονη καθημερινά; Πως μπορούμε να εξάγουμε γνώση από αυτές τις βάσεις δεδομένων που υπάρχουν εκεί έξω; Το 1950 με την ανακάλυψη των τεχνητών νευρωνικών δικτύων και περαιτέρω των γενετικών αλγορίθμων ένας νέος κλάδος ξεκίνησε να ανθίζει, η εξόρυξη δεδομένων, όπως επίσης και η αναγνώριση προτύπων. Ο κύριος στόχος των επιστημονικών αυτών κλάδων της πληροφορικής είναι η ανακάλυψη γνώσης από μια ή περισσότερες βάσεις δεδομένων και η εξαγωγή προτύπων από αυτές αναλύοντας δεδομένα. Έχοντας κάνει αυτή την ανάλυση προσπαθούμε να ανακαλύψουμε κάποιο ή κάποια πρότυπα μέσα από αυτά τα δεδομένα. Αν αναλογιστεί κανείς ότι αυτά τα δεδομένα μπορούν να εκφραστούν με αριθμούς τότε από αυτούς τους αριθμούς προκύπτουν πρότυπα, συνεπώς υπάρχουν πρότυπα παντού μέσα σε αυτά τα δεδομένα και εμείς θα πρέπει να τα αναγνωρίσουμε. Ακόμη τα δεδομένα αυτά αποτελούνται από χαρακτηριστικά τα οποία εκφράζονται και αυτά με αριθμούς. Άρα ένα δεδομένο αποτελείται από κάποια χαρακτηριστικά, δεδομένου ότι αυτά τα χαρακτηριστικά είναι αριθμοί το δεδομένο αυτό μπορεί να εκφραστεί ως ένα σημείο στο καρτεσιανό επίπεδο. Αυτό το δεδομένο μπορεί να ανήκει σε μία ομάδα δεδομένων και μπορεί να ομαδοποιηθεί- μετασχηματιστεί με βάση τα χαρακτηριστικά που το απαρτίζουν. Τι γίνεται όμως όταν αυτά τα χαρακτηριστικά υπερβαίνουν έναν συγκεκριμένο αριθμό και δεν μπορούμε να τα απεικονίσουμε είτε στο δυδιάστατο είτε στο τρισδιάστατο καρτεσιανό επίπεδο; Οι μαθηματικοί και οι επιστήμονες της πληροφορικής έχουν προτείνει διάφορες μεθόδους για την επίλυση αυτού του προβλήματος, μία μέθοδος από αυτές είναι η ανάλυσης κύριων συνιστωσών ή Principal Components Analysis (PCA) η οποία προτάθηκε για πρώτη φορά το 1933 από τον H.Hottelling. Η ανάλυση σε κύριες συνιστώσες μας επιλύει το πρόβλημα αυτό «μειώνοντας» την διάσταση των δεδομένων καθώς επίσης δίνεται η δυνατότητα οπτικοποίησης των δεδομένων. Στόχος λοιπόν της παρούσας πτυχιακής εργασίας είναι η παρουσίαση των προαναφερθέντων μεθόδων, υλοποιημένες σε πραγματικά προβλήματα καθώς επίσης και η οπτικοποίηση δεδομένων του εκάστοτε προβλήματος για ασφαλή εξαγωγή συμπερασμάτων.

Abstract

Tapsas Apostolos
Student TEI of Epirus
aptapsas@hotmail.com

Supervisor: S. P. Adam
Applications Professor TEI of Epirus
adamsp@teiep.gr

We live in the age of information, every day no matter where you stand you can see the wealth of information that's out there. Further information and ask from us wherever we go, bank, bill payments, gas stations, even in the super market. How can we exploit all this information abundant every day gives us out there? How we can extract knowledge from these databases out there? In 1950 with the discovery of artificial neural nets and a further genetic algorithms a new science began to flourish, data mining, and this and other such as pattern recognition. The main objective of these scientific computing disciplines is to discover knowledge from one or more databases and export standards from them. Having done this analysis we try to find someone or some standards through these data. Since these data can be expressed in numbers, then these numbers resulting standards therefore no standards anywhere within those data and we will have to recognize that. Even these data consist of features which properties are expressed by these numbers. Thus a data comprises a number of features, so as these characteristics are numbers this data can be expressed as a point on a Cartesian plane, so this data can belong to a data block and may be grouped, transformed into based features that compose it. But what if these features exceed a certain number, and we can not visualize either the two-dimensional Cartesian plane or three-dimensional? Mathematicians and computer scientists have proposed various methods to solve this problem, a method of them is principal components Principal Components Analysis (PCA) which was first proposed in 1933 by Hottelling. Resolution in our main components solves this problem by "downloading" dimension of our data and further allows the visualization. In this then the thesis will introduce you to this method on real problems in order of their visualization and further the conclusions.

Principal Components Analysis

by

Apostolos tapsas

Thesis Supervisor: Lecturer Professor S. P. Adam

Copyright by
Apostolos Tapsas
2016
All Rights Reserved

Περιεχόμενα

Ευχαριστίες.....σελ. 8	σελ. 8
Κεφάλαιο 1ο Εξόρυξη και προ επεξεργασία δεδομένων.....σελ. 9	σελ. 9
1.1 Δεδομένα.....σελ. 11	σελ. 11
1.2 Πληροφορία.....σελ.12	σελ.12
1.3 Συνάθροιση.....σελ. 13	σελ. 13
1.4 Μείωση πολλών διαστάσεων.....σελ. 14	σελ. 14
Κεφάλαιο 2ο Μεθοδολογίες μείωσης διαστάσεων των χαρακτηριστικών.....σελ. 16	σελ. 16
2.1 Κατηγοριοποίηση τεχνικών μείωσης διαστάσεων.....σελ. 18	σελ. 18
2.2 Γραμμικές μέθοδοι μείωσης διαστάσεων.....σελ. 18	σελ. 18
2.2.1 Ανάλυση σε κύριες συνιστώσες.....σελ. 18	σελ. 18
2.2.2 Ανάλυση πίνακα σε ιδιάζουσες τιμές.....σελ. 19	σελ. 19
Κεφάλαιο 3ο Ανάλυση Κυρίων Συνιστωσών.....σελ. 20	σελ. 20
3.1 Ιστορικό υπόβαθροσελ. 22	σελ. 22
3.2 Μαθηματικό υπόβαθρο.....σελ. 23	σελ. 23
3.2.1 Μέτρα θέσηςσελ. 23	σελ. 23
3.2.2 Μέτρα διασποράς.....σελ. 24	σελ. 24
3.3 Διατύπωση της μεθόδου.....σελ. 25	σελ. 25
Κεφάλαιο 4ο Το περιβάλλον R Studio και η γλώσσα προγραμματισμού R.....σελ. 28	σελ. 28
4.1 Το περιβάλλον R studio.....σελ. 30	σελ. 30
4.2 Εντολές εισόδου και εξόδου στην R.....σελ. 31	σελ. 31
4.3 Υπολογισμός μέτρων θέσης και διασποράς.....σελ. 33	σελ. 33
Κεφάλαιο 5 Παραδείγματα ΑΚΣ στην R.....σελ. 36	σελ. 36
5.1 Παράδειγμα κώδικα σε R της μεθόδου σε τυχαία δεδομένα....σελ. 38	σελ. 38
5.2 Παράδειγμα κώδικα σε R της μεθόδου σε πραγματικά δεδομένα.....σελ. 41	σελ. 41
5.3 Παράδειγμα PCA σε δεδομένα ασθενών της νόσου Parkinson.....σελ.43	σελ.43
Συμπεράσματα.....σελ. 42 και σελ. 44	σελ. 42 και σελ. 44
Λεξικό όρων.....σελ.46	σελ.46

Προοίμιο

Σκοπός μας είναι να σας παρουσιάσουμε την μέθοδο της Ανάλυσης Σε Κύριες Συνιστώσες (ΑΚΣ) σε διάφορους τύπους δεδομένων που μπορούμε να συναντήσουμε σε ένα πρόβλημα μέσω διαφόρων παραδειγμάτων που δεν θα έχουν μόνο σκοπό την μηχανική παρουσίαση της μεθόδου αλλά την εκμάθηση της . Καθώς προχωρούμε θα γνωρίζουμε και προγραμματιστικές έννοιες και την χρησιμότητα των εργαλείων που χρησιμοποιήθηκαν για την εκπόνηση των αλγορίθμων .

Προσωπικός μου στόχος είναι μετά το πέρας της ανάγνωσης της παρούσας πτυχιακής εργασίας να είναι σε θέση ο αναγνώστης να κατανοήσει πλήρως την έννοια της ΑΚΣ καθώς επίσης και να βρίσκεται σε θέση να εφαρμόσει όλες τις πρακτικές που επρόκειτο να αναφερθούν παρακάτω.

*Τάψας Απόστολος
Άρτα Ιανουάριος 2016*

Ευχαριστίες

Πρωτίστως θα ήθελα να ευχαριστήσω του γονείς μου για την οικονομική τους στήριξη καθώς επίσης και την ψυχολογική τους συμπαράσταση όσον αφορά τις σπουδές μου και ότι επρόκειτο να κάνω στην ζωή μου.

Θέρμες ευχαριστίες θα πρέπει να δοθούν στους συμφοιτητές μου και φίλους οι οποίοι με ωθήσαν στην υλοποίηση της παρούσας εργασίας.

Θα ήθελα επίσης να ευχαριστήσω των άνθρωπο που με δίδαξε προγραμματισμό και με ώθησε στο να δουλέψουμε μαζί για αυτή εδώ την πτυχιακή εργασία τον Κ.Σταύρο Αδάμ, καθηγητή του εκπαιδευτικού ιδρύματος Ηπείρου.

Τέλος θα ήθελα να ευχαριστήσω τον Κύριο Κωνσταντίνο Πατεράκη επίκουρο καθηγητή και νευροχειρουργό με έδρα στην Ιατρική σχολή του πανεπιστημίου Θεσσαλίας για την συνέντευξη που μας παραχώρησε.

*Τάψας Απόστολος
Άρτα Ιανουάριος 2016*

Κεφάλαιο 1^ο Εξόρυξη και προ επεξεργασία δεδομένων

Εισαγωγή

Στον τομέα της εξόρυξης δεδομένων χρησιμοποιείται μια σειρά τεχνικών για την ανάλυση των δεδομένων.

Ποιες όμως είναι αυτές οι τεχνικές; Πως ονομάζονται ; Με ποια σειρά εφαρμόζονται ;

Πως σε πρώτο στάδιο επεξεργαζόμαστε τα δεδομένα ;

Τι είναι δεδομένο και τι πληροφορία ;

Στο κεφάλαιο αυτό λοιπόν θα παρουσιασθούν κάποιες βασικές έννοιες όπως τι είναι πληροφορία και τι δεδομένο.

Σκοπός μας είναι να κατανοήσει ο αναγνώστης τι είναι η προεπεξεργασία των δεδομένων και τι εννοούμε με την έννοια μείωση διαστάσεων.

1.1 Δεδομένα.

Τα δεδομένα αποτελούν μια από της αρχικές μορφές της πληροφορίας. Αυτά αποτελούνται στην ουσία από εγγραφές δοσοληψιών ή γεγονότων οι οποίες χρησιμοποιούνται σε δοσοληψίες-συναλλαγές μεταξύ ανθρώπων ή και υπολογιστικών μηχανών (υπολογιστών). Σαν τέτοια δεδομένα δεν έχουν κάποιο νόημα εκτός αν κάποιος έχει γνώση από το περιβάλλον από το οποίο τα δεδομένα συγκεντρώθηκαν. Μια, λέξη ένας αριθμός ή ένα σύμβολο μπορούν να χρησιμοποιηθούν για να περιγράψουν ένα επιχειρησιακό αποτέλεσμα αλλά εκείνο που τους δίνει νόημα και τα κάνει ικανά να δώσουν πληροφορία είναι το περιβάλλον τους[2].

Σαν δεδομένα μπορεί να θεωρηθεί :

- Η πληροφορία, συχνά με τη μορφή γεγονότων (fact) τα οποία αποκτώνται από πειράματα ή έρευνες, τα οποία συνήθως αποτελούν την βάση των υπολογισμών μας και άλλοτε την σχεδίαση συμπερασμών.
- Η αριθμητική πληροφορία, η πληροφορία με την μορφή κειμένου, εικόνων, ήχου σε μορφή κατάλληλη για αποθήκευση και επεξεργασία από υπολογιστικές μηχανές (υπολογιστές).

Οπότε μπορούμε να ορίσουμε ότι :

Δεδομένα(Data) : είναι τα στοιχεία που εισάγουμε στην υπολογιστική μας μηχανή (υπολογιστή).

Όταν τα δεδομένα αυτά τύχουν επεξεργασίας τότε μετατρέπονται σε πληροφορία. Συμφωνά με τους Takeoff¹ (εικόνα 2) και Bellinger² (εικόνα 1), τα δεδομένα είναι σύμβολα και αναπαριστούν ένα γεγονός ή αναφορά σε ένα γεγονός χωρίς συσχέτιση με άλλα πράγματα. Τα δεδομένα δεν έχουν τύχει προηγούμενη επεξεργασία βρίσκονται σε μορφή χρησιμοποιήσιμη ή όχι, δεν περιέχουν κανένα νόημα και δεν έχουν καμία άλλη σημασία πέρα από την ύπαρξη τους. Έτσι, για παράδειγμα ένα φύλλο του excel μπορεί να περιέχει τα δεδομένα μιας έρευνας αγοράς τα οποία από μόνα τους δεν σημαίνουν τίποτε[2].



Εικόνα 1

Εικόνα 2

Τα δεδομένα έχουν ένα μεγάλο αριθμό ορισμών, έναν από τους οποίους δώσαμε παραπάνω. Ο ορισμός των δεδομένων από επιστημονικό πεδίο σε επιστημονικό πεδίο αλλάζει. Στο επιστημονικό πεδίο της πληροφορικής που ονομάζετε εξόρυξη δεδομένων τα δεδομένα ορίζονται ως είσοδοι μίας υπολογιστικής μηχανής (υπολογιστή). Έτσι λοιπόν στο χώρο της πληροφορικής μπορούμε να τα ονομάσουμε και ανεξάρτητη πληροφορία (πχ data stream), την οποία μπορούμε να αποθηκεύσουμε σε μία ή και περισσότερες βάσεις δεδομένων, για διάρκεια και διαχείριση[2].

¹Russell Lincoln Ackoff (12 February 1919 – 29 October 2009) : Αμερικάνος σύμβουλος και θεωρητικός οργανωτής Καθηγητής Διοικητικής Επιστήμης στο Wharton School, University of Pennsylvania. Πρωτοπόρος στον τομέα της επιχειρησιακής έρευνας, των συστημάτων σκέψης και της επιστήμης της διαχείρισης.

² Gene Bellinger (born March 9, 1948) : Αμερικανός σύμβουλος και θεωρητικός οργανωτής, γνωστός από την δουλειά του στα συστήματα σκέψης και στην διαχείριση της γνώσης.

1.2 Πληροφορία

Η πληροφορία επεκτείνει την έννοια των δεδομένων σε ένα ευρύτερο πλαίσιο. Η πληροφορία, όπως και τα δεδομένα παρουσιάζεται μέσω συμβόλων, τα οποία έχουν πολύπλοκη δομή και κανόνες. Η πληροφορία εμφανίζεται με μια ποικιλία μορφών όπως γραπτή, με μορφή διαγραμμάτων κανόνων κοκ[2].

Σαν πληροφορία μπορεί να θεωρηθεί :

- Η συγκεκριμένη σαφής γνώση αποκτηθείσα γύρω από κάτι ή κάποιον.
- Τα δεδομένα τα οποία έχουν συλλεχθεί γύρω από ένα συγκεκριμένο θέμα.
- Δεδομένα υπολογιστών τα οποία έχουν οργανωθεί και παρουσιάζονται με συστηματικό τρόπο για να κάνουν ξεκάθαρο το βασικό μήνυμα.

Συνεπώς μπορούμε να ορίσουμε ότι :

Πληροφορία(information): είναι το στοιχείο εκείνο που βοηθά κάποιον να γνωρίσει κάτι.

Συμφωνά με τους Takeoff (εικόνα 2) και Bellinger (εικόνα 1), η πληροφορία είναι δεδομένα τα οποία έχουν τύχει επεξεργασίας για να γίνουν χρήσιμα και να μπορούν να δώσουν απαντήσεις σε ερωτήσεις της μορφής, ποιος, τι, που, και πότε. Ως εκ τούτου, πληροφορία είναι δεδομένα στα οποία έχει δοθεί νόημα με κάποιο τρόπο σχεσιακής διασύνδεσης. Το νόημα μπορεί να είναι χρήσιμο αλλά δεν είναι απαραίτητο να είναι. Η πληροφορία ενσωματώνει την κατανόηση των σχέσεων. Έτσι παραδείγματος χάριν μια σχεσιακή βάση δεδομένων δημιουργεί πληροφορία από τα δεδομένα που είναι αποθηκευμένα σε αυτή όπως στον παρακάτω πίνακα[2].

Κωδικός συναλλαγής	Είδος	Τοποθεσία κατάστασης	Ημερομηνία	Τιμή	...
.	.	.	.	.	...
123456	Χαρτί	Λάρισα	11/9/2015	1,99€	...
5548787	Παπούτσια	Άρτα	22/9/2015	49,00€	...
00000001	Φαγητά	Αθήνα	22/9/2015	5.00€	...
.	.	.	.	.	...

Ένας τρόπος συνάθροισης των παραπάνω συναλλαγών είναι για αυτό το σύνολο δεδομένων είναι να αντικατασταθούν όλες οι συναλλαγές ενός καταστήματος με μια μόνο συναλλαγή, η οποία θα περιλαμβάνει όλα τα είδη του καταστήματος. Αυτό θα έχει ως αποτέλεσμα τη μείωση των εκατοντάδων συναλλαγών που γίνονται ημερησίως σε ένα συγκεκριμένο κατάστημα σε μια απλή ημερήσια συναλλαγή, και επομένως, το πλήθος των αντικειμένων δεδομένων μειώνονται και γίνεται ίσο με το πλήθος των καταστημάτων[2].

1.3 Συνάθροιση

Μερικές φορές ισχύει ότι «τα λιγότερα είναι πιο πολλά» και αυτό ακριβώς συμβαίνει με την συνάθροιση, τη συνένωση δηλαδή δύο ή πιο πολλών αντικειμένων σε ένα. Θεωρήστε ένα σύνολο δεδομένων, το οποίο αποτελείται από συναλλαγές που καταγράφουν τις ημερήσιες πωλήσεις προϊόντων σε διάφορες πόλεις (Άρτα, Λάρισα, Αθήνα...) και για διαφορετικές ημέρες στην πορεία του χρόνου. Δείτε τον πίνακα της προηγούμενης ενότητας. Ένας τρόπος συνάθροισης των συναλλαγών για αυτό το σύνολο δεδομένων είναι να αντικατασταθούν όλες οι συναλλαγές ενός καταστήματος με μία μόνο συναλλαγή, η οποία περιλαμβάνει όλα τα είδη του καταστήματος. Αυτό θα έχει ως αποτέλεσμα τη μείωση των εκατοντάδων ή χιλιάδων συναλλαγών των που γίνονται ημερησίως σε ένα συγκεκριμένο κατάστημα σε μία απλή ημερήσια συναλλαγή, και επομένως, το πλήθος των αντικειμένων δεδομένων μειώνεται και γίνεται ίσο με το πλήθος των καταστημάτων[1].

Αμέσως κάποιος μπορεί να διακρίνει την ανάγκη να «μειώσουμε» κάποια χαρακτηριστικά τα οποία μπορεί σε μία από τις συναλλαγές που είδαμε να μην μας είναι και τόσο χρήσιμα πράγμα το οποίο είναι αδύνατον εμείς να γνωρίζουμε εκ των προτέρων, και ίσως μία συνάθροιση αυτόν των χαρακτηριστικών να μας είναι πιο αρεστή στο να μπορέσουμε π.χ. να οπτικοποιήσουμε αυτές της συναλλαγές. Δηλαδή θα μας ήταν πιο αρεστών να μπορέσουμε με κάποιο τρόπο να παράξουμε έναν π.χ. μετασχηματισμό των γνωρισμάτων-χαρακτηριστικών αυτόν[1].

Η συνάθροιση λοιπόν των χαρακτηριστικών αυτόν μπορεί να θεωρηθεί και ως ο γραμμικός μετασχηματισμός των γνωρισμάτων-χαρακτηριστικών γενικότερα, και τον ημερήσιων πωλήσεων ειδικότερα όσων αναφορά το παράδειγμα που αναφέραμε νωρίτερα στην προηγούμενη ενότητα[1].

Τέλος οφείλουμε να αναφέρουμε ότι μία συνάθροιση δεν είναι αναγκαία ένας γραμμικός μετασχηματισμός των γνωρισμάτων-χαρακτηριστικών που καλούμαστε να αναλύσουμε κάθε φορά αλλά θα μπορούσε να είναι ο μέσος όρος των τιμών των καταστημάτων αναφερόμενη πάντα στον πίνακα του παραδείγματος της προηγούμενης ενότητας[1].

Συνεπώς μπορούμε να ορίσουμε ότι :

Συνάθροιση (aggregation): είναι η διαδικασία εξάλειψης χαρακτηριστικών.

1.4 Μείωση πολλών διαστάσεων

Τα σύνολα δεδομένων μπορούν να απαρτίζονται από ένα μεγάλο πλήθος γνωρισμάτων-χαρακτηριστικών. Μπορούμε να θεωρήσουμε ένα σύνολο εγγραφών, όπου κάθε έγγραφο μπορεί να αναπαρασταθεί από ένα διάνυσμα του οποίου τα στοιχεία είναι οι συχνότητες εμφάνισης κάθε λέξης στο έγγραφο. Σε τέτοιες περιπτώσεις υπάρχουν τυπικά χιλιάδες ή δεκάδες χιλιάδες γνωρισμάτων-χαρακτηριστικών, ένα για κάθε λέξη του λεξιλογίου. Ως ένα δεύτερο παράδειγμα μπορούμε να θεωρήσουμε τα δεδομένα μίας χρονοσειράς η αποτελείται από την καθημερινή τιμή κλεισίματος διαφόρων μετοχών σε μία περίοδο 30 χρόνων. Σε αυτή την περίπτωση τα γνωρίσματα-χαρακτηριστικά των οποίων οι τιμές σε συγκεκριμένες ημερομηνίες θα είναι χιλιάδες[1].

Τουτέστιν υπάρχει μία ποικιλία οφελών από την μείωση των πολλών διαστάσεων. Ένα βασικό όφελος που προέκυψε, είναι ότι πολλοί αλγόριθμοι εξόρυξης δεδομένων λειτουργούν καλύτερα όταν το πλήθος των διαστάσεων είναι μικρότερο. Αυτό συμβαίνει κατά ένα μέρος επειδή η μείωση των διαστάσεων μπορεί να εξαλείψει μη-σχετικά γνωρίσματα-χαρακτηριστικά και να μειώσει το «θόρυβο». Κατά ένα άλλο μέρος εξαιτίας αυτού που στην εξόρυξη δεδομένων ονομάζουμε κατάρα των πολλών διαστάσεων, η οποία θα περιγραφεί παρακάτω. Άλλο ένα όφελος είναι ότι η μείωση των διαστάσεων μπορεί να μας οδηγήσει σε πιο κατανοητά μοντέλα επειδή ακριβώς τα μοντέλα αυτά θα εμπεριέχουν λιγότερα γνωρίσματα-χαρακτηριστικά. Επίσης, όπως αναφέραμε και στην προηγούμενη ενότητα η μείωση των διαστάσεων μπορεί να καταστήσει την οπτικοποίηση των δεδομένων ευκολότερη[1].

Η κατάρα των πολλών διαστάσεων.

Η κατάρα των πολλών διαστάσεων αναφέρεται στο φαινόμενο πολλών μορφών ανάλυσης δεδομένων που δυσκολεύουν αρκετά όσο αυξάνεται το πλήθος των διαστάσεων των δεδομένων. Ειδικότερα, όσο αυξάνεται το πλήθος των διαστάσεων τα δεδομένα γίνονται πολύ αραιά στον χώρο που καταλαμβάνουν. Για την κατηγοριοποίηση τους αυτό μπορεί να σημαίνει ότι δεν υπάρχουν αρκετά αντικείμενα δεδομένων ώστε να επιτρέψουν την δημιουργία ενός μοντέλου που αποδίδει, με τρόπο αξιόπιστο, μια κατηγορία σε κάθε πιθανό αντικείμενο. Για τη δημιουργία συστάδων οι έννοιες της πυκνότητας και της απόστασης μεταξύ των σημείων γίνονται λιγότερο ουσιαστικές. Ως συνέπεια των παραπάνω, πολλοί αλγόριθμοι συσταδιοποίησης και κατηγοριοποίησης π.χ. ο αλγόριθμος k-means, παρουσιάζουν προβλήματα με δεδομένα πολλών διαστάσεων, όπως μικρότερη ακρίβεια κατηγοριοποίησης και φτωχές σε ποιότητα συστάδες[1].

Τεχνικές γραμμικής άλγεβρας για τη μείωση πολλαπλών διαστάσεων.

Μερικές από τις πιο κοινές προσεγγίσεις για την μείωση των πολλαπλών διαστάσεων, ειδικότερα για συνεχή δεδομένα, χρησιμοποιούν τεχνικές από την γραμμική άλγεβρα για να προβάλουν τα δεδομένα από ένα χώρο πολλαπλών διαστάσεων σε ένα χώρο με μικρότερο πλήθος διαστάσεων. Η ανάλυση κύριων συνιστωσών (principal components analysis-PCA) η οποία αποτελεί το κύριο θέμα της παρούσας πτυχιακής εργασίας, είναι μία τεχνική της γραμμικής άλγεβρας για συνεχή χαρακτηριστικά η οποία βρίσκει νέα χαρακτηριστικά (κύριες συνιστώσες) τα οποία :

Είναι γραμμικοί συνδυασμοί των αρχικών χαρακτηριστικών.

Είναι κάθετα μεταξύ τους.

Λαμβάνουν τη μέγιστη απόκλιση στα δεδομένα :

Για παράδειγμα οι δύο πρώτες κύριες συνιστώσες λαμβάνουν όσο το δυνατόν μεγαλύτερη απόκλιση στα δεδομένα με δύο ορθογώνια χαρακτηριστικά, τα οποία είναι γραμμικοί συνδυασμοί των αρχικών χαρακτηριστικών. Η μέθοδος αυτή θα περιγραφεί αναλυτικότερα στο κεφάλαιο 3.

Η διάσπαση μοναδιαίων τιμών (singular value decomposition-SVD) είναι μία άλλη τεχνική της γραμμικής άλγεβρας που σχετίζεται με την ανάλυση κύριων συνιστωσών και χρησιμοποιείται συχνά για τη μείωση των πολλαπλών διαστάσεων. Σκοπός του επόμενου κεφαλαίου προτού φτάσουμε να κάνουμε μια ενδελεχή ανάλυση της PCA είναι η αναφορά και η περιγραφή μερικών άλλων τέτοιων μεθόδων που βρίσκουν εφαρμογή στον χώρο της πληροφορικής και ποιο συγκεκριμένα στην εξόρυξη δεδομένων (Data Mining) και αναγνώρισης προτύπων (Pattern Recognition).

Βιβλιογραφία και Αναφορές πρώτου Κεφαλαίου

[1] Pang-Ning Tan & Michael-Steinbach & Vipin-Kumar, 2010, *Εισαγωγή Στην Εξόρυξη Δεδομένων*, 2016, Εκδόσεις ΤΖΙΟΛΑ, Αθήνα.

[2] Νικόλαος Ματσαστίνης, 2010, *Συστήματα Υποστήριξης Αποφάσεων*, 2014, Εκδόσεις νέων τεχνολογιών, Αθήνα.

Κεφάλαιο 2^ο Μεθοδολογίες μείωσης διαστάσεων των χαρακτηριστικών

Εισαγωγή

Στο προηγούμενο κεφάλαιο αναφέραμε μερικές από τις μεθόδους που έχουν προταθεί για την μείωση των διαστάσεων. Είναι ξεκάθαρο ότι εμείς θα ασχοληθούμε με την Ανάλυση Κύριων Συνιστωσών. Προτιμήσαμε όμως να παρουσιάσουμε επικουρικά αλλά και ακαδημαϊκά το εύρος όλης της γκάμας που ονομάζεται «μεθοδολογίες μείωσης διαστάσεων». Για τον λόγο αυτό σκοπός αυτού του κεφαλαίου είναι να παρουσιασθούν στον αναγνώστη γραμμικές αλλά και μη-γραμμικές μέθοδοι μείωσης διαστάσεων. Έτσι ώστε να είναι σε θέση να κατανοεί την κατηγοριοποίηση των μεθόδων αυτών και έπειτα να εμβαθύνει σε μία από αυτές όπως εμείς θα εμβαθύνουμε στην Ανάλυση κύριων Συνιστωσών στο επόμενο κεφάλαιο.

2.1 Κατηγοριοποίηση τεχνικών μείωσης διαστάσεων

Οι τεχνικές μείωσης των διαστάσεων κατηγοριοποιούνται σε δύο μεγάλες κατηγορίες, τις γραμμικές και μη-γραμμικές μεθοδολογίες.

2.2 Γραμμικές μέθοδοι μείωσης διαστάσεων

Με τον όρο γραμμικές εννοούμε τις μεθόδους που χρησιμοποιούν τεχνικές της γραμμικής άλγεβρας, μερικές από αυτές τις μεθόδους είναι η μελέτη διανυσμάτων, διανυσματικών χώρων, γραμμικών απεικονίσεων και συστημάτων γραμμικών εξισώσεων. Ακόμη σε μερικές από αυτές τις μεθόδους γίνεται και έντονη χρήση στατιστικών εργαλείων όπως τα μέτρα θέσης και διασποράς που θα αναλύσουμε αργότερα στο κεφάλαιο 3.

Οι δύο πιο δημοφιλείς μέθοδοι είναι :

- Η ανάλυση κυρίων συνιστωσών
- Η πολυδιάστατη κλιμάκωση

Για τις παραπάνω μεθόδους θα κάνουμε κουβέντα στις επόμενες δύο ενότητες.

2.2.1 Ανάλυση σε κύριες συνιστώσες

Η ιδιοσύνθεση ενός συμμετρικού θετικά ημιορισμένου πίνακα αποδίδει μία ορθογώνια βάση ιδιοδιανυσμάτων, καθένα εκ των οποίων έχει μια μη-αρνητική ιδιοτιμή[2]. Η ορθογώνια αποσύνθεση ενός πίνακα χρησιμοποιείται στην ανάλυση πολλαπλών μεταβλητών, όπου το δείγμα με πίνακα συνδιασποράς είναι ένας πίνακας ο οποίος προκύπτει από την Ανάλυση Κυρίων Συνιστωσών(ΑΚΣ). Η ΑΚΣ μελετά γραμμικούς συσχετισμούς μεταξύ μεταβλητών, εφαρμόζεται σε πίνακες συνδιασποράς ή σε πίνακες συσχετισμού (στους οποίους κάθε μεταβλητή είναι κλιμακωτή έτσι ώστε η διασπορά του δείγματος να ισούται με ένα). Για τους πίνακες συνδιασποράς και συσχετισμού τα ιδιοδιανύσματα αντιστοιχούν στις κύριες συνιστώσες και οι ιδιοτιμές στις εξηγημένες διασπορές των κύριων συνιστωσών. Η ανάλυση κύριων συνιστωσών του πίνακα συσχετισμού παρέχει μία ορθογώνια ιδιοβάση για τον χώρο των παρατηρούμενων δεδομένων. Σε αυτή την βάση οι μεγαλύτερες ιδιοτιμές αντιστοιχούν στις κύριες συνιστώσες που σχετίζονται περισσότερο με την συμμεταβλητότητα ανάμεσα σε έναν αριθμό παρατηρούμενων δεδομένων[2].

Η ΑΚΣ χρησιμοποιείται για την μελέτη μεγάλων συνόλων δεδομένων, όπως αυτά που βλέπουμε στην εξόρυξη δεδομένων, στη χημική έρευνα, στη ψυχολογία καθώς επίσης και στον τομέα του μάρκετινγκ[2].

Η ΑΚΣ είναι δημοφιλής κυρίως στην ψυχολογία, στον τομέα της ψυχομετρικής. Στην Q μεθοδολογία (ερευνητική μέθοδος που χρησιμοποιείται στην ψυχολογία και στις κοινωνικές επιστήμες για τη μελέτη «υποκειμενικότητας» των ανθρώπων) οι ιδιοτιμές του πίνακα συσχετισμού καθορίζουν την κρίση του Q-μεθοδολόγου στην πρακτική σημαντικότητα (που διαφέρει από την στατιστική σημαντικότητα του ελέγχου της υπόθεσης). Οι παράγοντες με ιδιοτιμές μεγαλύτερες του 1 θεωρούνται πρακτικά σημαντικές το οποίο οδηγεί στην εξήγηση σημαντικού ποσού των δεδομένων

μεταβλητότητας. Ενώ οι ιδιοτιμές που είναι μικρότερες της μονάδας (1) θεωρούνται πρακτικά ασήμαντες, καθώς εξηγούν μόνο ένα αμελητέο ποσοστό των δεδομένων μεταβλητότητας. Γενικότερα, η ανάλυση κύριων συνιστωσών μπορεί να χρησιμοποιηθεί ως μέθοδος ανάλυσης παραγόντων σε μοντέλα διαρθρωτικής εξίσωσης[2].

2.2.2 Ανάλυση πίνακα σε ιδιάζουσες τιμές

Στη γραμμική άλγεβρα η ανάλυση σε ιδιάζουσες τιμές είναι μία παραγοντοποίηση ενός πίνακα με πραγματικά ή μιγαδικά στοιχεία με πολλές χρήσιμες εφαρμογές στη θεωρία σημάτων και τη στατιστική.

Η ανάλυση ενός $m \times n$ πίνακα M , με πραγματικά ή μιγαδικά στοιχεία, σε ιδιάζουσες τιμές είναι μια παραγοντοποίηση της μορφής

$$M = U \Sigma V^*$$

όπου U είναι ένας $m \times m$ πραγματικός ή μιγαδικός ορθομοναδιαίος πίνακας, Σ ένας $m \times n$ ορθογώνιος διαγώνιος πίνακας με μη αρνητικές τιμές στην διαγώνιο και V^* (ο συζυγής ανάστροφος του V , ή απλά ο ανάστροφος του V αν ο V είναι πραγματικός) ένας $n \times n$ πραγματικός ή μιγαδικός ορθομοναδιαίος πίνακας. Τα διαγώνια στοιχεία $\Sigma_{i,i}$ του Σ είναι γνωστά ως ιδιάζουσες τιμές του M . Οι m στήλες του U και οι n στήλες του V ονομάζονται αριστερά-ιδιάζοντα διανύσματα και δεξιά-ιδιάζοντα διανύσματα του M , αντίστοιχα.

Η ανάλυση σε ιδιάζουσες τιμές και η ανάλυση σε ιδιοτιμές είναι στενά συνδεδεμένες.

Δηλαδή:

Τα αριστερά-ιδιάζοντα διανύσματα του M είναι τα ιδιοδιανύσματα του MM^* .

Τα δεξιά-ιδιάζοντα διανύσματα του M είναι τα ιδιοδιανύσματα του M^*M .

Οι μη-μηδενικές ιδιάζουσες τιμές του M (που εμφανίζονται στις διαγώνιες θέσεις του Σ) είναι οι τετραγωνικές ρίζες των μη μηδενικών ιδιοτιμών του M^*M και του MM^* .

Εφαρμογές που χρησιμοποιούν την ανάλυση σε ιδιάζουσες τιμές περιλαμβάνουν τον υπολογισμό του ψευδοαντιστρόφου τη προσαρμογή δεδομένων με την μέθοδο των ελαχίστων τετραγώνων, τις προσεγγίσεις πινάκων και την εύρεση της βαθμίδας του χώρου των στηλών και του μηδενοχώρου ενός πίνακα[3].

Βιβλιογραφία και Αναφορές δευτέρου Κεφαλαίου

[1] Pang-Ning Tan & Michael-Steinbach & Vipin-Kumar, 2010, *Εισαγωγή Στην Εξόρυξη Δεδομένων*, 2016, Εκδόσεις ΤΖΙΟΛΑ, Αθήνα.

[2] Νικόλαος Μαρτσαστίνης, 2010, *Συστήματα Υποστήριξης Αποφάσεων*, 2014, Εκδόσεις νέων τεχνολογιών, Αθήνα.

[3] http://wikipedia.qwika.com/en2el/Nonlinear_dimensionality_reduction

Κεφάλαιο 3^ο Ανάλυση Κυρίων Συνιστωσών

Εισαγωγή

Έχοντας αναφέρει τις διάφορες μεθόδους για την μείωση των διαστάσεων στο κεφάλαιο 2 έχουμε μία πρώτη εξοικείωση με το θέμα το οποίο θα εξετάσουμε και θα αναλύσουμε στο κεφάλαιο αυτό. Σκοπός είναι να παρουσιαστεί η Ανάλυση Κυρίων Συνιστωσών (ΑΚΣ) ως έννοια σε μία πρώτη μας γνωριμία με την μέθοδο αυτή, καθώς και το θεωρητικό, ιστορικό και μαθηματικό υπόβαθρο της μεθόδου της ΑΚΣ.

Είναι ήδη γνωστό ότι η ΑΚΣ είναι μια μέθοδος που χρησιμοποιεί μεθόδους και έννοιες της γραμμικής άλγεβρας άλλα μετά το πέρας της ανάγνωσης αυτού του κεφαλαίου δεν θα αντιληφθείτε ότι η μέθοδος αυτή είναι το ίδιο στενά συνδεδεμένη με την στατιστική καθώς και με την γραμμική άλγεβρα.

Για το λόγο αυτό αναφέρουμε και κάποιες στατιστικές έννοιες όπως τα μέτρα θέσης και διασποράς ως μαθηματικό υπόβαθρο της μεθόδου.

3.1 Ιστορικό υπόβαθρο

Η μέθοδος της ΑΚΣ ξεκίνησε τον προηγούμενο αιώνα και πιο συγκεκριμένα το 1901 από τον Pearson ο οποίος πρότεινε την μέθοδο μιας γραμμικής παλινδρόμησης ελαχίστων τετραγώνων σε n διαστάσεις. «Πατέρας» της μεθόδου θεωρείται ο H. Hotelling ο οποίος το 1933 πρότεινε τη μέθοδο αυτή για την ανάλυση της διασποράς πολυδιάστατων τυχαίων μεταβλητών.

Η ιδέα της μεθόδου ξεκινάει από την παρατήρηση πολλών συστημάτων που περιγράφονται από πολλές n τυχαίες μεταβλητές. Οι βαθμοί ελευθερίας m είναι λιγότεροι (δηλ. $m < n$). έτσι ενώ η διάσταση του διανύσματος παρατήρησης είναι n , το σύστημα μπορεί να περιγράψει από m ασυσχέτιστες αλλά κρυφές τυχαίες μεταβλητές. Συχνά οι m κρυφές αυτές μεταβλητές ονομάζονται παράγοντες ή χαρακτηριστικά του διανύσματος παρατήρησης. Στη στατιστική ορολογία θα ονομάσουμε τον αριθμό n ως την επιφανειακή διάσταση και τον αριθμό m ως την ουσιαστική διάσταση. Προφανώς αν οι n τυχαίες μεταβλητές είναι ασυσχέτιστες τότε ισχύει ότι $m=n$ ενώ όσο πιο συσχετισμένες είναι μεταξύ τους τόσο πιο λίγα χαρακτηριστικά αρκούν για να περιγράψουν τις τυχαίες παρατηρήσεις.

Ακόμη παρατηρούμε ότι έχουμε δώσει έμφαση σε δύο έννοιες, την γραμμική παλινδρόμηση και τη διασπορά.

Σκοπός μας είναι να αφήσουμε για λίγο στην άκρη την ΑΚΣ και στις επόμενες ενότητες να δείξουμε το μαθηματικό υπόβαθρο της μεθόδου και τέλος να παρουσιάσουμε το θεωρητικό υπόβαθρο και την ίδια την μέθοδο σε βάθος όπως αναφέραμε πρωτίστος στο κεφάλαιο αυτό.

3.2 Μαθηματικό υπόβαθρο

Στην προηγούμενη ενότητα αναφέραμε μερικές από τις έννοιες που θα αναλύσουμε στις επόμενες ενότητες. Στο σημείο αυτό σας αναφέρουμε τις έννοιες που θα παρουσιαστούν στις επόμενες ενότητες.

Σαν μέτρα θέσης στα πλαίσια αυτής της πτυχιακής εργασίας αναφερόμαστε στην έννοια της μέσης τιμής.

Σαν μέτρα διασποράς στα πλαίσια αυτής της πτυχιακής εργασίας αναφερόμαστε στις έννοιες της κύμανσης, διακύμανσης, τυπικής απόκλισης και του συντελεστής μεταβολής.

Οι παραπάνω έννοιες αποτελούν μερικά από τα μαθηματικά εργαλεία της ΑΚΣ. Το γεγονός ότι πρέπει κάποιος να κατανοήσει την μέθοδο σε βάθος, μας ώθησε στο να θεωρήσουμε απαραίτητη την περιγραφή τους καθώς από την αρχή θεωρήσαμε ότι ο αναγνώστης μπορεί να μην έχει ασχοληθεί ξανά με κάτι παρόμοιο και προφανώς μπορεί να μπερδευτεί. Σκοπός μας δεν είναι να μπερδέψουμε τον αναγνώστη αλλά να του δώσουμε όλα τα απαραίτητα εφόδια για την κατανόηση της μεθόδου αλλά και για την εισαγωγή τους στον χώρο, με σκοπό να του κεντρίσουμε το ενδιαφέρον να ερευνήσει και παραπέρα από αυτή την πτυχιακή εργασία χτίζοντάς του τις απαραίτητες βάσεις[3].

3.2.1 Μέτρα θέσης

Μέτρα θέσης μίας κατανομής λοιπόν ονομάζουμε : *τα μέτρα εκείνα τα οποία μας δίνουν τη θέση του «κέντρου» των παρατηρήσεων.*

Ένα μέτρο θέσης είναι και η **μέση τιμή** των παρατηρήσεων μας η οποία ορίζεται ως:

$$\bar{X} = (1/n) \sum_{i=1}^n X_i \quad (1)$$

Ακόμη ως μέση τιμή μπορεί να ορισθεί και το άθροισμα των παρατηρήσεων X_i δια του πλήθους των παρατηρήσεων n .

Σταθμικός μέσος είναι ένα ακόμη μέτρο θέσης το οποίο μας είναι χρήσιμο στις περιπτώσεις στις οποίες δίνουμε διαφορετική βαρύτητα (έμφαση) στις τιμές ενός συνόλου δεδομένων, τότε αντί για την μέση τιμή χρησιμοποιούμε τον σταθμικό μέσο ο οποίος μας δίνεται από τον τύπο :

$$\bar{X} = \sum_{i=1}^n x_i w_i / \sum_{i=1}^n w_i \quad (2)$$

Διάμεσος

Τέλος η διάμεσος δεν έχει κάποιο συγκεκριμένο τύπο υπολογισμού καθώς μπορεί να υπολογιστεί διαισθητικά, μπορούμε όμως να ορίσουμε την διάμεσο ως εξής :

Διάμεσος ενός δείγματος n παρατηρήσεων οι οποίες έχουν διαταχθεί σε αύξουσα σειρά ορίζεται: η μεσαία παρατήρηση, όταν το n είναι περιττός αριθμός ή ο μέσος όρος των δύο μεσαίων παρατηρήσεων όταν το n είναι άρτιος αριθμός.

3.2.2 Μέτρα διασποράς

Μέτρα διασποράς ονομάζουμε : τα μέτρα εκείνα τα οποία μας δίνουν τη διασπορά των παρατηρήσεων, δηλαδή την έκτασης των παρατηρήσεων γύρω από το «κέντρο» τους.

Ένα μέτρο διασποράς είναι η **διακύμανση ή διασπορά** των παρατηρήσεων $X_1, X_2, \dots, X_n$ μίας μεταβλητής X η οποία ορίζεται ως:

$$\sigma_X^2 = (1/n) \sum_{i=1}^n (X_i - \bar{X})^2 \quad (3)$$

Ακόμη ως μέση διασπορά μπορεί να ορισθεί και ο μέσος όρος των τετραγώνων των αποστάσεων X_i από την μέση τιμή τους $\bar{X}$.

Τυπική απόκλιση : ονομάζουμε την θετική τετραγωνική ρίζα της διακύμανσης σ^2 η οποία μας δίνεται από την σχέση :

$$\sigma_X = \sqrt{\sigma_X^2} \quad (4)$$

Ως έννοιας η τυπική απόκλιση μας δείχνει την κατανομή που ακολουθούν τα δεδομένα ή χαρακτηριστικά μας.

Παραδείγματος χάριν αν η καμπύλη συχνοτήτων ενός χαρακτηριστικού x καθώς και η μέση τιμή του $\bar{X}$ είναι κανονικές τότε η τυπική απόκλιση σ είναι κανονική ή περίπου κανονική, τότε :

- Το 68%, περίπου των παρατηρήσεων βρίσκεται στο διάστημα $(\bar{X} - \sigma_X, \bar{X} + \sigma_X)$
- Το 95%, περίπου των παρατηρήσεων βρίσκεται στο διάστημα $(2\bar{X} - 2\sigma_X, 2\bar{X} + 2\sigma_X)$
- Το 97%, περίπου των παρατηρήσεων βρίσκεται στο διάστημα $(3\bar{X} - 3\sigma_X, 3\bar{X} + 3\sigma_X)$
- Το εύρος R ισούται περίπου με έξι τυπικές αποκλίσεις $R \cong 6\sigma$

Συντελεστής μεταβολής CV

Ο συντελεστής μεταβολής CV μας δείχνει την ομοιογένεια ενός δείγματος σε σχέση με κάποιο άλλο, ο συντελεστής μεταβολής υπολογίζεται ως η διαίρεση της τυπικής απόκλισης σ ενός χαρακτηριστικού από την μέση τιμή του ως εξής :

$$CV = \sigma_x / \bar{X} \quad (5)$$

Εάν αυτός ο συντελεστής είναι μικρότερος του 10 τότε λέμε ότι το δείγμα μας είναι ομοιογενές ειδάλλως λέμε ότι είναι ανομοιογενές.

3.3 Διατύπωση της μεθόδου

Έχοντας αναφέρει το μαθηματικό υπόβαθρο της μεθόδου είμαστε έτοιμοι να παρουσιάσουμε το θεωρητικό υπόβαθρο της μεθόδου της ανάλυσης σε κύριες συνιστώσες, όπως αναφέραμε η μέθοδος αυτή προτάθηκε από τον Hotelling το 1933 για την ανάλυση της διασποράς πολυδιάστατων τυχαίων μεταβλητών.

Η ιδέα της μεθόδου είναι απλή δηλαδή η μείωση των διαστάσεων των δεδομένων μας.

Ας θεωρήσουμε έναν πίνακα D με διαστάσεις $m \times n$ ο οποίος πίνακας είναι ο πίνακας των δεδομένων ή των παρατηρήσεων μας οι οποίες θεωρητικά είναι τυχαίες, ακόμη μπορούμε να θεωρήσουμε ότι οι τιμές αυτές οι μεταβλητές βρίσκονται στο n -διαστατό χώρο. Στόχος μας είναι να εφαρμόσουμε ένα γραμμικό μετασχηματισμό προκειμένου να προκύψει ένα νέο σύνολο χαρακτηριστικών.

Στο σημείο αυτό θα δώσουμε τον αλγόριθμο της μεθόδου και στην συνέχεια στο κεφάλαιο 5 θα δούμε μεταφέρουμε αυτόν τον αλγόριθμο στην γλώσσα προγραμματίσμου R.

Δοθέντων μίας μήτρας D τυχαίων χαρακτηριστικών διαστάσεων $m \times n$ και έστω ότι n ο αριθμός των χαρακτηριστικών των παρατηρήσεων m . Ακόμη έστω ότι το n είναι ίσο με N , όπου N ο αριθμός του χώρου των διαστάσεων άπου θα βρίσκονται οι παρατηρήσεις μας.

Τότε για να «μειώσουμε» από το N -διαστατό χώρο ($n=N$) στον δισδιάστατο ($N=2$) ή τρισδιάστατο ($N=3$) τότε η μέθοδός της ανάλυσης σε κύριες συνιστώσες προτείνει τα εξής παρακάτω βήματα:

1. Τον υπολογισμό των συνδιακυμάνσεων της μήτρας D σε μία νέα μήτρα S .
2. Τον υπολογισμό των ιδιοτημών και ιδιοδιανυσμάτων του πίνακα S σε δύο νέους πίνακες με όνομα $eigenval$ και $eigenvec$.

Τα δεδομένα του πίνακα $eigenvec$ είναι οι κύριες συνιστώσες μας και τα δεδομένα του πίνακα $eigenval$ είναι οι τιμές συμμετοχής των συνιστωσών αυτών. Όσο πιο μεγάλη η συμμετοχή μίας συνιστώσας τόσο πιο μεγαλύτερος είναι και ο ρόλος του χαρακτηριστικού που εκφράζει.

Για παράδειγμα αν είχαμε το πίνακα $D = \begin{pmatrix} 3 & 4 \\ 1 & 4 \end{pmatrix}$ η μήτρα D μπορεί να χαρισθεί σε δύο

υπομήτρες x και y με $x=(3 \ 1)$ και $y=(4 \ 4)$ οπότε κατανοούμε ότι τα κάθε μία μήτρα D μπορεί να χαρισθεί σε υπομήτρες τόσες όσες και τα χαρακτηριστικά του D .

Ο λόγος που το κάνουμε σε αυτό το παράδειγμα είναι καθαρά επικουρικός καθώς για να υπολογίσουμε τον πίνακα των συνδιακυμάνσεων S θα πρέπει να υπολογίσουμε των μέσο όρο του κάθε χαρακτηριστικού δηλαδή τα $\bar{X}$ και $\bar{Y}$

Με βάση την σχέση (1) την οποία σας θυμίζουμε έχουμε ότι :

$$\bar{X} = (1/n) \sum_{i=1}^n X_i \text{ και } \bar{Y} = (1/n) \sum_{i=1}^n Y_i$$

Άρα το $\bar{X} = 1/2 (3+1)=2$ και $\bar{Y} = 1/2(4+4)=4$

Αμέσως ελέγχουμε αν τα $\bar{X}$ και $\bar{Y}$ είναι ίσα με 0 και εάν είναι τότε ο πίνακας S είναι ίσος με το γινόμενο του πίνακα D με τον ανάστροφο του D^T ειδικά υπολογίζεται από τον τύπο :

$$\text{cov}(x, y) = \sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y}) \quad (6)$$

Στο παράδειγμα μας έχουμε ότι $\bar{X} \neq 0$ και $\bar{Y} \neq 0$ άρα υπολογίζοντας με βάση την σχέση (6) έχουμε ότι :

$$S = \begin{pmatrix} 0,5 & 1,5 \\ 1,5 & 4,5 \end{pmatrix}$$

Τέλος υπολογίζουμε του δύο πίνακες eigenval και eigenvec ως

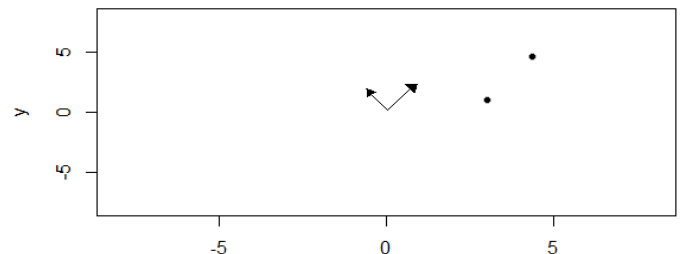
Θεωρούμε δύο πίνακες eigenval και eigenvec έτσι ώστε

$S * \text{eigenvec} = \text{eigenval} * \text{eigenvec}$ όπου eigenval και eigenvec μη μηδενικοί πίνακες.

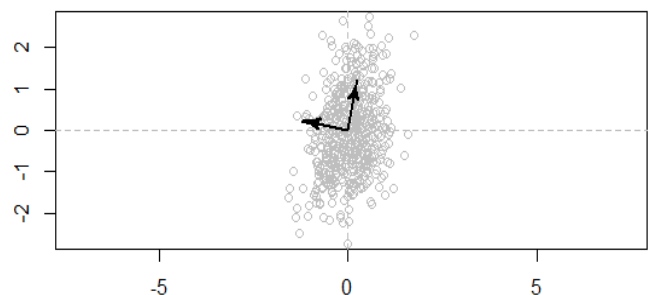
Στο παράδειγμά μας έχουμε ότι $\text{eigenval} = (5 \ -2,71 * 10^{-10})$ $\text{eigenvec} = \begin{pmatrix} 0,31 & -0,94 \\ 0,94 & 0,31 \end{pmatrix}$

Αυτό που παρατηρούμε είναι ότι οι δύο κύριες συνιστώσες μας είναι κάθετες μεταξύ τους καθώς $0,31 * (-0,94) + 0,31 * 0,94 = 0$

Τέλος στο διπλανό σχήμα βλέπουμε μια απεικόνιση της μεθόδου του παραδείγματος μας.



Το παραπάνω παράδειγμα είναι καθαρά επικουρικό και έχει ως στόχο να δώσει στον αναγνώστη την ιδέα της μεθόδου, ένα ποιο αναλυτικό παράδειγμα μαζί με τον κώδικα του θα παρουσιαστεί στο κεφάλαιο 5. Αξίζει όμως να σας δείξουμε από τώρα την εικόνα που περιμένουμε να δούμε από ένα αντίστοιχο παράδειγμα τυχαίων μεταβλητών έτσι ώστε να σας δώσουμε μία εικόνα του τι έπεται. Στην διπλανή εικόνα λοιπόν βλέπεται το διάγραμμα διασποράς ενός εκ



των παραδειγμάτων που θα δούμε στην ενότητα 5 αφότου πρώτιστος φροντίσουμε να σας παρουσιάσουμε το περιβάλλον στο οποίο δουλέψαμε.

Αφότου λοιπόν εξετάσαμε το μαθηματικό υπόβαθρο της μεθόδου, στην συνέχεια εξετάζοντας και το θεωρητικό υπόβαθρο ήμαστε σε ένα πολύ καλό σημείο να κάνουμε μία εισαγωγή στο περιβάλω ή στα περιβάλλοντα που μπορούμε να υλοποιήσουμε τέτοιου είδους μεθόδους ένα τέτοιο λοιπόν περιβάλω θα δούμε και θα περιγράψουμε στο επόμενο κεφάλαιο.

Βιβλιογραφία και Αναφορές τρίτου Κεφαλαίου

- [1] Pang-Ning Tan & Michael-Steinbach & Vipin-Kumar, 2010, *Εισαγωγή Στην Εξόρυξη Δεδομένων*, 2016, Εκδόσεις ΤΖΙΟΛΑ, Αθήνα.
- [2] Θεοδωρίδης Σέργιος, 2011, *Εισαγωγή στην αναγνώριση προτύπων με Matlab*, 2013, Ιατρικές Εκδόσεις Π. Χ. Πασχαλίδης, Αθήνα.

Κεφάλαιο 4° Το περιβάλλον R Studio και η γλώσσα προγραμματισμού R

Εισαγωγή

Έως τώρα εξετάσαμε διάφορες έννοιες είτε μαθηματικές είτε όχι, θυμίζουμε τις έννοιες “δεδομένο και πληροφορία” τις οποίες είδαμε στις πρώτες δύο ενότητες αυτής της πτυχιακής εργασίας. Στην συνέχεια γνωρίσαμε κάποιες περαιτέρω μαθηματικές έννοιες οι οποίες σε μερικούς αναγνώστες ίσως είναι ήδη γνωστές. Μην ξεχνάμε όμως ότι στα πλαίσια της παρούσας πτυχιακής εργασίας δεν είναι μόνο να σας παρουσιάσουμε μία μέθοδο μόνον από την μαθηματική οπτική γωνία αλλά και από τη σκοπιά της πληροφορικής. Παρουσιάζοντας και περιβάλλοντα τα οποία κάνουν την δουλεία μας πιο εύκολη και γρήγορη. Ένα λοιπόν από αυτά τα περιβάλλοντα είναι και η γλώσσα προγραμματισμού R, η οποία έχει αναπτυχθεί με σκοπό να βοηθήσει τον κλάδο της στατιστικής ανάλυσης δεδομένων και όχι μόνο.

Σε αυτό το κεφάλαιο θα εστιάσουμε περισσότερο στην παρουσίαση του περιβάλλοντος της γλώσσας προγραμματισμού R που ονομάζουμε R studio μέσω διαφόρων υπολογισμών μέτρων θέσης και διασποράς που παρουσιάσαμε στο κεφάλαιο 3 έτσι ώστε στο κεφάλαιο 5 να έχετε όλα τα εφόδια από μαθηματικής και τεχνικής πλευράς ώστε να παρουσιαστεί πλήρως ένα πρόγραμμα το οποίο θα πραγματοποιεί την μέθοδο ΑΚΣ στο παράδειγμα που υποσχθήκαμε πως θα παρουσιάσουμε στα τέλη τις ενότητας 3.3 και ενός ακόμη με δεδομένα τα οποία έχουν να κάνουμε με ασθενείς οι οποίοι έχουν κάποια καρδιολογική πάθηση.

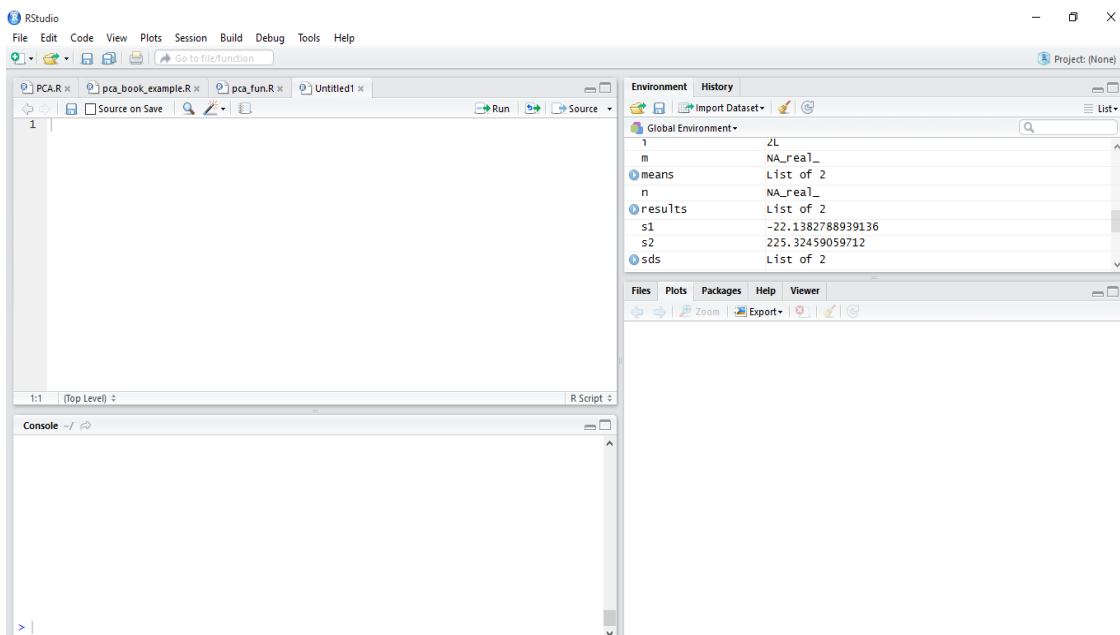
4.1 Το περιβάλλον R studio

Το περιβάλλον λοιπόν το οποίο θα χρησιμοποιήσουμε πρωτίστως για των υπολογισμό διαφόρων μέτρων θέσης και διασποράς είναι το R studio, καλό θα ήταν ο αναγνώστες να έχει μία εξοικείωση με τον προγραμματισμό αν και δεν είναι απαραίτητο προαπαιτούμενο. Όπως λοιπόν στο κεφάλαιο 3 έτσι και εδώ θα θεωρήσουμε ότι ο αναγνώστης δεν έχει κάποια προηγούμενη επαφή με το περιβάλλον οπότε πέραν από το περιβάλλον της γλώσσας θα πρέπει να δώσουμε και μερικές βασικές προγραμματιστικές έννοιες όπως το πώς να παίρνουμε είσοδο από το πληκτρολόγιο της υπολογιστικής μηχανής μας και πώς να εμφανίζουμε στην οθόνη αυτής.

Ας δούμε λοιπόν αρχικά το μια εικόνα του R studio το οποίο μπορείτε να το κατεβάσετε δωρεάν από το διαδίκτυο.

Προσοχή για να εγκαταστήσετε το R studio θα πρέπει να έχετε εγκατεστημένη είδη την R την οποία μπορείτε να κατεβάσετε και αυτή δωρεάν από το διαδίκτυο.

Μια εικόνα λοιπόν το περιβάλλοντος δίνεται παρακάτω, στην συνέχεια θα σας δείξουμε πώς να παίρνεται είσοδο και πως θα εμφανίσετε πράγματα στην οθόνη.



Αμέσως στην άνωθεν εικόνα παρατηρούμε τέσσερις περιοχές, η πρώτη περιοχή είναι η περιοχή η οποία γράφεται ο κώδικας μας και μένει εκεί με σκοπό να εκτελεσθεί, στο πεδίο λοιπόν αυτό γράφουμε τα λεγόμενα scripts (κομμάτια κώδικα) τα οποία θα εκτελούνται και **μπορούν να αποθηκεύουν στον δίσκο μας**. Ακριβώς παρακάτω βλέπουμε την περιοχή της κονσόλας στην οποία μπορούμε να γράψουμε και εκεί κώδικα όπως στην άνωθεν της περιοχή αλλά ο κώδικας αυτός δεν μένει σταθερός, αν γράψουμε ένα κομμάτι κώδικα εκεί το εκτελέσουμε και έπειτα κλείσουμε το περιβάλλον τότε ο κώδικας μας πάει χαμένος για τον λόγο αυτόν προτιμούμε να γράφουμε των κώδικα σε scripts τα οποία και αυτά αν δεν τα αποθηκεύσουμε θα χαθούν αλλά σε αντίθεση με την κονσόλα αυτά μπορούν να αποθηκεύουν. Στην

συνέχεια πάνω δεξιά έχουμε την περιοχή στην οποία αποθηκεύονται η μεταβλητές μας και από κάτω της την περιοχή την οποία γίνονται διάφορες απεικονίσεις των δεδομένων μας όπως π.χ. η απεικόνιση του διαγράμματος διασποράς των δεδομένων μας.

Ας δούμε λοιπόν πως θα μπορούμε να εισάγουμε δεδομένα από το πληκτρολόγιο και να εμφανίσουμε δεδομένα στην οθόνη.

4.2 Εντολές εισόδου και εξόδου στην R

Υπάρχουν δύο τρόποι να εκτελέσουμε εντολές στην R, ο ένας είναι να γράψουμε τις εντολές μας σε ένα script και ο δεύτερος είναι να τις γράψουμε στην περιοχή της κονσόλας (console). Εμείς ξεκινώντας σε αυτή την ενότητα θα εκτελέσουμε τις πρώτες μας εντολές στην περιοχή της κονσόλας και έπειτα θα γράψουμε το πρώτο μας script.

Εντολή εισόδου `readline()`

Η εντολή εισόδου `readline` είναι μια εντολή η οποία διαβάζει ένα αλφαριθμητικό χαρακτήρα από το πληκτρολόγιο. Με τον όρο αλφαριθμητικό εννοούμε μία ακολουθία από αριθμούς ή χαρακτήρες ή τον συνδυασμό και των δύο, σας πληροφορούμε ότι αν έχουμε σε μία μεταβλητή αλφαριθμητικό χαρακτήρα ο οποίος θα μπορούσε να είναι το αλφαριθμητικό 1 και μια δεύτερη με το αλφαριθμητικό 2 π.χ. τότε αυτές οι δύο μεταβλητές αν προσθέθουν, αφαιρεθούν κ.τ.λ. π δεν θα δώσουν κάποιο λογικό αποτέλεσμα οπότε εμείς θα πρέπει να τις μετατρέπουμε σε ένα τέτοιο τύπο που να μας επιτρέπει να κάνουμε πράξεις με αυτές.

Η διαδικασία αυτή στον κλάδο του προγραμματισμού ονομάζεται `type casting` και αυτό που κάνει είναι να μετατρέπει έναν τύπο δεδομένων σε ένα άλλο. Παραδείγματος χάριν εμείς στην κονσόλα μας γράψαμε το κατηγορήμα :

```
a=readline()
```

Τότε η μεταβλητή `a` θα είχε μία τιμή που θα της δινόταν από εμάς μέσω του πληκτρολογίου, στην περίπτωση που θα δίναμε ένα 2 και θέλαμε να το μετατρέψουμε σε αριθμό τότε θα έπρεπε να κάνουμε στο `a` `type casting` ως εξής :

```
a= as.numeric(a)
```

Για να μην μπούμε σε αυτή την διαδικασία μπορούμε να κάνουμε όλη την δουλειά γράφοντας κατευθείαν το παρακάτω κατηγορήμα.

```
a=as.numeric(readline())
```

Το οποίο μας παράγει το ίδιο ακριβώς αποτέλεσμα.

Τέλος, αν θέλουμε να εμφανίσουμε και ένα μήνυμα στο χρήστη ώστε να τον προειδοποιήσουμε για τον τυπο μεταβλητής που θέλουμε να μας δώσει τότε χρησιμοποιούμε το κατηγορήμα :

```
readline(prompt="μήνυμα προς τον χρήστη")
```

Όποτε εάν θέλουμε ο χρήστης να μας δώσει ένα αριθμό δείχνοντας του το σχετικό μήνυμα και σε δεύτερο χρόνο θέλουμε να κάνουμε πράξη με αυτό τον αριθμό τότε γράφουμε το εξής κατηγορημα το οποίο είναι συνδυασμός όλων των παραπάνω ως :

```
a=as.numeric(readline(prompt="Enter a number: "))
```

Εντολή εξόδου print()

Κάθε γλώσσα προγραμματισμού έχει μία ή περισσότερες εντολές εξόδου ομοίως και η R δεν έχει μία μόνο εντολή εξόδου αλλά περισσότερες όπως π.χ. την plot που θα σας παρουσιάσουμε παρακάτω.

Η διαδικασία εμφάνισης μηνυμάτων στην οθόνη είναι πιο εύκολη από την αντίθετή της που περιγράψαμε παραπάνω, το μόνο που έχουμε να κάνουμε είναι να πληκτρολογήσουμε το παρακάτω κατηγορημα.

print("μήνυμα ή μεταβλητή")

εάν για παράδειγμα θέλαμε να γράψουμε ένα μήνυμα στην οθόνη το ποιο θα θέλαμε να γράφει hello world τότε θα το γράφαμε ως εξής :

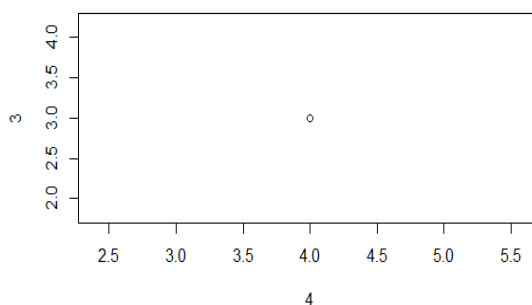
print("hello world")

Το παραπάνω μήνυμα και όλα τα μηνύματα τα οποία γράφουμε με την εντολή print εμφανίζονται στην περιοχή της κονσόλας, εκεί δηλαδή που υποθέσαμε πως θα γράψουμε τον κώδικα μας από την αρχή αυτής της ενότητας.

Τέλος η εντολή εξόδου plot είναι μία πιο περίπλοκη εντολή η οποία μας βοηθά στην απεικόνιση διαφόρων στοιχείων ή και σημείων, εάν για παράδειγμα θέλαμε να απεικονίσουμε το σημείο a(4,3) στο καρτεσιανό επίπεδο τότε θα πληκτρολογούσαμε το κατηγορημα

plot(4,3)

Του οποίου το αποτέλεσμα το φαίνεται στην διπλανή εικόνα, η οποία εικόνα μας είναι πλέον γνωστή από το προηγούμενο κεφάλαιο. Η εντολή plot όπως βλέπεται έχει δύο χαρακτηριστικά που θα πρέπει να της δοθούν το πρώτο είναι η τιμή του σημείου πάνω στο άξονα x'x και η δεύτερη πάνω στο άξονα y'y.



4.3 Υπολογισμός μέτρων θέσης και διασποράς

Έχοντας παρουσιάσει κάποια βασικά εργαλεία της R τώρα θα γράψουμε το πρώτο μας script το οποίο θα υπολογίζει διάφορα μέτρα θέσης και διασποράς.

Για να δημιουργήσουμε ένα script στην R θα πρέπει να πάμε πάνω αριστερά εκεί όπου έχει το όνομα file, new file, R script.

Τώρα θα έχει γίνει ορατό και το πεδίο του script που σας αναφέραμε παραπάνω. Στο script γράφουμε τις εντολές που έχουμε μάθει έως τώρα, εμείς αυτό που θα χρειασθούμε είναι μερικές μεταβλητές τις οποίες δεν έχει νόημα να της ζητήσουμε από τον χρήστη αλλά θα τις παράξουμε τυχαία εμείς μέσα από το κατηγορήμα

```
x <- rnorm(n,mean=0)
```

Το παραπάνω κατηγορήμα παράγει n τυχαίες μεταβλητές διασκορπισμένα κοντά στο 0. Το n στην περίπτωση μας προτιμήσαμε να το ζητήσουμε από τον χρήστη ως εξής :

```
n <- as.numeric(readline(prompt="Enter the number of data: "))
```

Στην συνέχεια μπορούμε να κάνουμε δούμε το διάγραμμα διασποράς κάνοντας plot(x)

Στην συνέχεια ήμαστε έτοιμη να βρούμε τα μέτρα θέσεις και διασποράς του συνόλου δεδομένων x.

Υπολογισμός μέσης τιμής

Ας ξεκινήσουμε με τον μέσο όρο των δεδομένων μας υπάρχουν δύο τρόποι για τον υπολογισμό του μέσου όρου των δεδομένων μας. Ο πρώτος είναι να αθροίσουμε τα δεδομένα όλου του συνόλου x και έπειτα να τα διαιρέσουμε με το μέγεθός του και ο δεύτερος είναι να χρησιμοποιήσουμε συναρτήσεις της γλώσσας προγραμματίσμου R. Αν αναλύαμε όλα τα προλεγόμενα τότε θα ξεφεύγαμε κατά πολύ αφενός από το θέμα αυτού του κεφαλαίου και αφετέρου από το θέμα της πτυχιακής αυτής. Σκοπός μας είναι να σας δείξουμε ένα γρήγορο τρόπο υπολογισμού αυτών των στατιστικών εννοιών με κανένα σκοπό να εμβαθύνουμε στις προγραμματιστικές έννοιες. Ο υπολογισμός της μέσης τιμής ενός συνόλου x υπολογίζεται από την συνάρτηση mean της γλώσσας προγραμματισμού R ως εξής :

```
xmean=mean(x)
```

Υπολογισμός συνδιακύμανσης

Ομοίως με τον υπολογισμό της μέσης τιμής ο υπολογισμός των συνδιακύμανσης γίνεται και αυτός με μία ιδική συνάρτηση της γλώσσας προγραμματισμού R με όνομα cov.

Ο υπολογισμός της συνδιακύμανσης ενό συνόλου x υπολογίζεται ως εξής :

```
xcov=cov(x)
```

Υπολογισμός τυπικής απόκλισης

Έχοντας υπολογίσει την συνδιακύμανση των δεδομένων μας ευκολά μπορούμε να υπολογίσουμε την τυπική απόκλιση ως εξής

```
sds=sqrt(xcov*xcov)
```

Η συνάρτηση sqrt υπολογίζει την τετραγωνική ρίζα οποιουδήποτε αριθμού ή τις τετραγωνικές ρίζες ενός συνόλου.

Συντελεστής μεταβολής

Τέλος για τον συντελεστή μεταβολής δεν θα χρησιμοποιήσουμε κάποια συνάρτηση αλλά θα τον υπολογίσουμε επαναληπτικά προκειμένου να σας δείξουμε την έννοια της επανάληψης στην R. Ο υπολογισμός της μεταβολής των δεδομένων γίνεται ως εξής:

```
CVs<-c((sds[1,]*0))
for (i in 1:length(sds)){
  if (xmean[i]>0)
    CVs[i]<-data.frame(s[i]/xmean[i])
  else
    CVs[i]<-data.frame(s[i]/(xmeans[i]*(-1)))
}
```

Τέλος θα σας δώσουμε ένα ολοκληρωμένα script το οποίο υπάρχει ένα πίνακας D με n χαρακτηριστικά ο πίνακας, στην συνέχεια κάνουμε της παραπάνω πράξεις στον D^T

```
D<-t(D)
means<-c((D[1,]*0))
# υπολογισμός των μέσων τιμών
for (i in 1:length(D[1,]))
  means[i]<-data.frame(mean(D[,i]))
# υπολογισμός των τυπικών αποκλίσεων
sds<-c((D[1,]*0))
for (i in 1:length(means))
  sds[i]<-data.frame(sd(D[,i]))
#Υπολογισμός συντελεστών μεταβολής
#Αν ο συντελεστής μεταβολής ενός στοιχείου CVs[i] είναι μεγαλύτερο
#του μηδενός τότε το συγκεκριμένο χαρακτηριστικό δεν είναι ομοιογενές.
CVs<-c((D[1,]*0))
for (i in 1:length(sds)){
  if (means[i]$V1>0)
    CVs[i]<-data.frame(sds[i]$V1/means[i]$V1)
  else
    CVs[i]<-data.frame(sds[i]$V1/(means[i]$V1)*(-1))
}
```

Πως λειτουργεί

Το παραπάνω script όπως αναφέραμε υπολογίζει τα μέτρα θέσης και διασποράς του αντιστρόφου ενός πίνακα D. Στην αρχή αντιστρέφουμε τον πίνακα D, στην συνέχεια ορίζουμε έναν πίνακα means μεγέθους ίσο με την πρώτη γραμμή του D. Υπολογίζουμε τις μέσες τιμές. Έχοντας όλες τις μέσες τιμές στον πίνακα means μπορούμε ευκολά να υπολογίσουμε τις τυπικές αποκλίσεις. Στη συνέχεια μπορούμε να υπολογίσουμε και τους συντελεστές μεταβολής των δεδομένων μας.

Οι παραπάνω υπολογισμοί δεν είναι καθόλου τυχαίοι καθώς μερικοί από αυτούς τους υπολογισμούς μας είναι απαραίτητοι στην διαδικασία της ΑΚΣ.

Μέσω λοιπόν αυτού του κεφαλαίου σας είναι ήδη γνωστός ο τρόπος υπολογισμού των μέτρων θέσης και διασποράς που θα χρειαστούμε στην επόμενη ενότητα κατά της υλοποίησης του προγράμματος στην R.

Τέλος, εκεί θα δούμε και μερικές ακόμη έννοιες σχετικά με τους πίνακες των δεδομένων μας όπως η έννοια means[i]\$V1

Η έννοια \$V1 θα μας γίνει γνωστή στο επόμενο κεφάλαιο. Κλείνοντας λοιπόν αυτό το κεφάλαιο θεωρούμε ότι πλέον είστε σε θέση να κατανοήσετε τους υπολογισμούς και τον κώδικα που θα παρουσιαστεί στην επόμενη ενότητα, έχοντας μελετήσει και το παράρτημα Β, διότι όπως αναφέραμε πολλάκις, υποθέτουμε ότι ο αναγνώστης δεν έχει προηγούμενη εμπειρία με το αντικείμενο.

Βιβλιογραφία και Αναφορές τετάρτου Κεφαλαίου

[1] Σταύρος Π. Αδάμ, *Σημειώσεις προγραμματισμού I και II*, Άρτα, 2006

[2] Jesse Liberty, *Εγχειρίδιο της C++*, Εκδόσεις Μ. Γκιούρδα, Αθήνα, 2007.

Κεφάλαιο 5^ο Παραδείγματα ΑΚΣ στην R

Εισαγωγή

Έως τώρα έχουμε παρουσιάσει όλα τα προαπαιτούμενα της μεθόδου σε μαθηματικό, θεωρητικό, προγραμματιστικό ακόμη και ιστορικό υπόβαθρο.

Οπότε στο σημείο αυτό είμαστε έτοιμοι να δώσουμε κάποια παραδείγματα μέσω της μετατροπής του αλγορίθμου που δώσαμε στο κεφάλαιο 3 καθώς έχουμε παρουσιάσει όλα τα βασικά προαπαιτούμενα που θα πρέπει να γνωρίζει ο αναγνώστης. Σκοπός λοιπόν αυτού του κεφαλαίου είναι να σας παρουσιάσει κάποια παραδείγματα διαφόρων τύπων δεδομένων όπως αναφέραμε από την αρχή.

Με τον όρο διαφόρων τύπων δεδομένων εννοούμε δεδομένα τα οποία είτε μας έχουν δοθεί από κάποιον ειδικά για ανάλυση όπως ένας ιατρός, μηχανολόγος, οικονομολόγος κ.ο.κ. .

Μετά το τέλος την ανάγνωσης του κεφαλαίου αυτού ο αναγνώστης θα είναι σε θέση να μπορεί να εμβahύνει στην ανάλυση άλλων τύπων δεδομένων όπως παραδείγματος χάριν τα διαστηματικά δεδομένα, καθώς σε συνδυασμό με το κεφάλαιο 3 θα του έχει παρουσιαστεί η μέθοδος αυτή σε βάθος όπως είχαμε υποσχεθεί από την αρχή της πτυχιακής εργασίας.

5.1 Παράδειγμα κώδικα σε R της μεθόδου σε τυχαία δεδομένων

Ξεκινούμε λοιπόν την παρουσίαση της μεθόδου πάνω σε τυχαία δεδομένα τα οποία θα παράξουμε εμείς έχοντας γνωστή την μήτρα συνδιακύμανσης S η οποία είναι :

$$S = \begin{pmatrix} 0,3 & 0,2 \\ 0,2 & 1 \end{pmatrix}$$

Πριν όμως ξεκινήσουμε θα πρέπει να σας θυμίσουμε την μέθοδο με λίγο διαφορετικό τρόπο από αυτόν που είδαμε στο κεφάλαιο 3.

Η ΑΚΣ λοιπόν είναι μια από τις πιο δημοφιλείς τεχνικές μείωσης των αριθμών των διαστάσεων. Ξεκινώντας από ένα αρχικό σύνολο n χαρακτηριστικών, που συνιστούν τα στοιχεία ενός διανύσματος $x \in \mathbb{R}^n$, ο στόχος είναι να εφαρμόσουμε ένα γραμμικό μετασχηματισμό προκειμένου να προκύψει ένα νέο σύνολο χαρακτηριστικών τέτοιο ώστε :

$$y = D^T x \quad (7)$$

Τα νέα δεδομένα της μήτρας y θα πρέπει να είναι ασυσχέτιστα μεταξύ τους. Σε δεύτερο στάδιο, θα επιλεγούν τα περισσότερα σημαντικά από τα στοιχεία αυτά όταν το n ξεπερνά της 2 ή 3 διαστάσεις[1].

Τα βήματα που ακολουθούμε είναι τα εξής :

1. Υπολογισμός του πίνακα συν-διασποράς ή συνδιακύμανσης. Είθισται να υποθέτουμε ότι η μέση τιμή των δεδομένων είναι το 0, δηλαδή ότι $E[x]=0$. Στην περίπτωση αυτή οι μήτρες συνδιασποράς και αυτοσυσχέτισης συμπίπτουν. Αν αυτό δεν συμβαίνει τότε αφαιρούμε την μέση τιμή. Θυμίζουμε ότι, για n διάστατα χαρακτηριστικά $x_i \in \mathbb{R}^n, i=1,2,...,n$, η μήτρα αυτοσυσχέτισης δίνεται από την σχέση :

$$R = (1/n) \sum_{i=1}^n x_i x_i^T \quad (8)$$

2. Υπολογισμός των ιδιοδιανυσμάτων του S . Υπολογίζονται οι n ιδιοτιμές λ_i και ο ιδιοδιανύσματα d_i (με $\lambda_i, d_i \in \mathbb{R}^n$)
3. Ταξινόμηση των ιδιοτιμών σε φθίνουσα τάξη
4. Επιλογή των m μεγαλύτερων ιδιοτιμών όπου m ο αριθμός των διαστάσεων που θέλουμε να απεικονίσουμε τα δεδομένα μας.
5. Χρησιμοποίηση των αντίστοιχων ιδιοδιανυσμάτων $d_i, i=0,1,2,...,m-1$ για το σχηματισμό της μήτρας μετασχηματισμού.
 $D = [d_0 \ d_1 \ d_2 \ \dots \ d_{m-1}]$
6. Μετασχηματισμός κάθε n -διάστατου διανύσματος x του αρχικού χώρου σε ένα m -διάστατο διάνυσμα y μέσω της σχέσης (7)

Έχοντας λοιπόν ως μπούσουλα τα παραπάνω 6 βήματα του αλγορίθμου και την μήτρα συνδιασποράς ή μαςτε έτοιμη να παράξουμε τα δεδομένα μας.

Στο παράδειγμά μας θα παράξουμε 500 δεδομένα στον δυσδιάστατο χώρο ($n=2$) τα οποία ακολουθούν Gaussian κατανομή μηδενικής μέσης τιμής ($E[x]=0$).

Έχοντας λοιπόν πράξει την μίτρα των δεδομένων μας D ήμαστε σε θέση γράψουμε το πρόγραμμα που υλοποιεί την μέθοδο έχοντας ως μούσουλα τα παραπάνω 6 βήματα ως εξής :

```
library(shape)
library(mvtnorm)
library(MASS)
#Αρχή αλγορίθμου
D <- rbind(x,y)
# D Matrix Transpose
D<-t(D)
#Υπολογισμός Του Πίνακα Συνδιακύμανσης S
S=NULL
S=cov(D)
# Υπολογισμός των ιδιοτιμών και ιδιοδιανυσμάτων του S
U<-eigen(S)
# Τοποθέτηση των ιδιοτιμών
eigenvals<-U$values
eigenvectors<- U$vectors
# Απεικόνιση των δεδομένων με διάγραμμα διασποράς
plot(x,y,col='grey',xlab='x', ylab='y',xlim=c(-3, 3), ylim=c(-3, 3))
eqscplot(D, y,col='grey', tol = 0.04)
# Τοποθέτηση αξόνων χ'χ και y'y
abline(h = 0, v = 0, lty = 2, col = 8)
# Απεικόνιση των δύο πρώτων κυρίων συνιστωσών
Arrows(0,0,U$vectors[1,1],U$vectors[2,1],lwd=2,col='black')
Arrows(0,0,U$vectors[1,2],U$vectors[2,2],lwd=2,col='black')
# Υπολογισμός νέων μετασχηματισμένων δεδομένων.
Y=D*U$vectors
# Τέλος αλγορίθμου
```

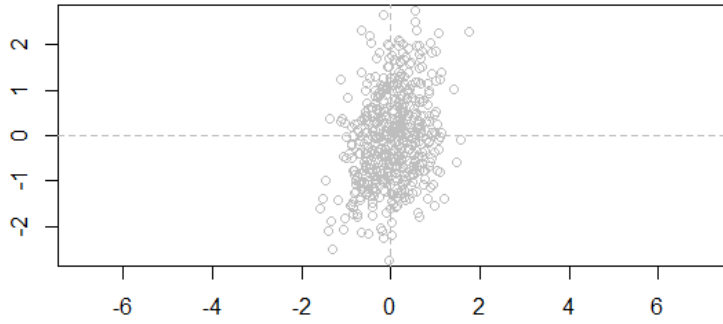
Πως λειτουργεί

Ο παραπάνω αλγόριθμος υλοποιεί την μέθοδο της ΑΚΣ στη γλώσσα προγραμματισμού R. Πριν ξεκινήσουμε την επεξήγηση θα πρέπει να σας ενημερώσουμε ότι το παράδειγμα υπάρχει υλοποιημένο στην αναφορά [1] στο περιβάλλον matlab, ένα άλλο ευρέως γνωστό περιβάλλον του χώρου της ανάλυσης και εξόρυξης δεδομένων και της μηχανικής μάθησης και αναγνώρισης προτύπων. Υπολογίζοντας λοιπόν σε έναν πίνακα U τις ιδιοτιμές και ιδιοδιανύσματα του S θα πρέπει να τα τοποθετήσουμε σε ένα νέο πίνακα καθώς η R έχει την ιδιομορφία να μην τα ξεχωρίζει από μόνη της σε αντίθεση με το περιβάλλον matlab. Ως αποτέλεσμα αυτού αν θέλουμε να πάρουμε τις κύριες συνιστώσες μας, δηλαδή τα ιδιοδιανύσματα θα πρέπει να το κάνουμε με το κατηγορημα U\$vectors και ομοίως U\$values για τις ιδιοτιμές. Στην επόμενη ενότητα θα σας παραθέσουμε ένα πλήρες παράδειγμα με πραγματικά δεδομένα από ασθενείς με καρδιοπάθεια όπου εκεί τα βήματα 5 και 6 δεν μπορούν να παραληφθούν.

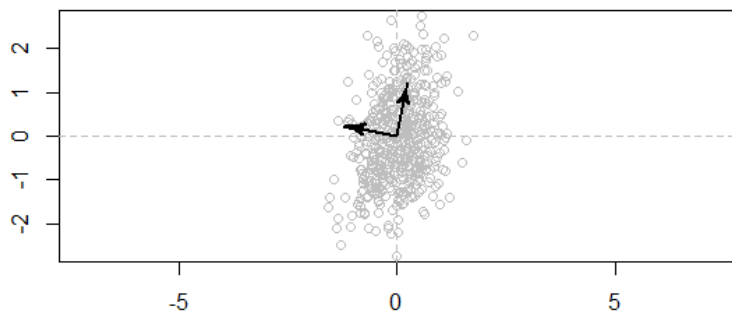
Παρακάτω σας δίνουμε την απεικόνιση των δεδομένων χωρίς τις κύριες συνιστώσες τους στο σχήμα 1 και με τις αυτές στο σχήμα 2. Στο σχήμα 3 Βλέπουμε τα νέα

μετασχηματισμένα δεδομένα τα οποία μετά την ΑΚΣ πλέον είναι κάθε μεταξύ τους.

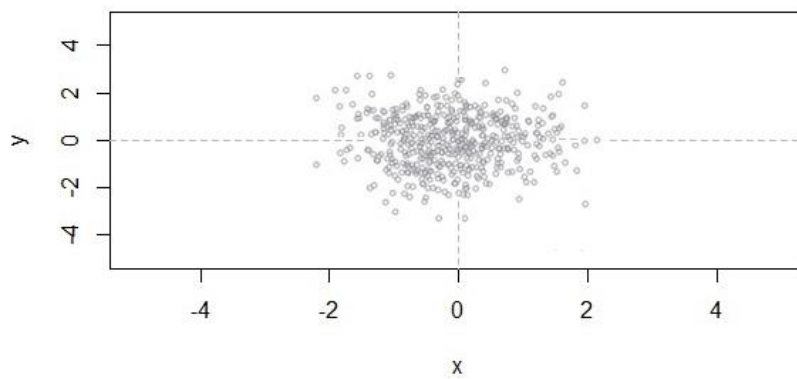
Σχήμα 1



Σχήμα 2



Σχήμα 3



5.2 Παράδειγμα κώδικα σε R της μεθόδου σε πραγματικά δεδομένα.

Ας εξετάσουμε στο σημείο αυτό ένα πραγματικό πρόβλημα με δεδομένα τα οποίο μας έχουν δοθεί από έναν καρδιολόγο να τα επεξεργασία. Τα δεδομένα αυτά έχουν τα εξής ακόλουθα χαρακτηριστικά.

Το σύνολο δεδομένων (dataset) αυτό μπορεί να το βρει κανείς στο [link \(https://www.dropbox.com/s/fkjf33q16ze1hi0/CardiologyNumerical2.xls?dl=0\)](https://www.dropbox.com/s/fkjf33q16ze1hi0/CardiologyNumerical2.xls?dl=0).

V1="age" (ηλικία)
 V2="sex" (φύλο)
 V3="chest pain type" (τύπος πόνου στήθους) πολλή, λίγο, καθόλου
 V4="blood pressure" (πίεση του αίματος)
 V5="cholesterol" (χοληστερίνη)
 V6="Fasting blood sugar <120" (σάκχαρα του αίματος)
 V7="resting ecg" (ξεκούραση)
 V8="maximum heart rate" (μέγιστη καρδιακή συχνότητα)
 V9="angina" (κυνανχή, φλεγμονή του φάρυγγα που προκαλεί δυσκολία στην αναπνοή και την κατάποση)
 V10="peak" (μέγιστη συγκέντρωση στο αίμα)
 V11="slope"(κλίση)
 V12="#colored vessels" (μαύρα αγγεία)
 V13="thal"
 V14="class" (κλάση, κατηγορία)

Με πίνακα ιδιοτιμών

U\$values

[V1] 2.695898e+03
 [V2] 5.426472e+02
 [V3] 3.051127e+02
 [V4] 5.849868e+01
 [V5] 3.751540e+00
 [V6] 1.206509e+00
 [V7] 9.110992e-01
 [V8] 8.107249e-01
 [V9] 6.309064e-01
 [V10] 2.145027e-01
 [V11] 1.875359e-01
 [V12] 1.462269e-01
 [V13] 1.216963e-01
 [V14] 9.453609e-02

Συμπεράσματα

Παρατηρούμε ότι οι παραπάνω ιδιοτιμές είναι διατεταγμένες από την μεγαλύτερη ως την μικρότερη. Η κάθε μία από αυτές τις ιδιοτιμές οι οποίες έχουν υπολογισθεί με την μέθοδο ΑΚΣ στην γλώσσα προγραμματισμού R μας δείχνει την συμμετοχή του κάθε χαρακτηριστικού στα δεδομένα μας.

Εάν αυτή η συμμετοχή είναι μεγαλύτερη από το ένα(1) ή κοντά σε αυτό τότε το κάθε χαρακτηριστικό μπορεί να θεωρηθεί σημαντικό και να κρατηθεί, αλλιώς θεωρούμε ότι το χαρακτηριστικό αυτό μας προκαλεί «θόρυβο» στα δεδομένα μας. Από τον παραπάνω λοιπόν πίνακα προκύπτει ότι την μεγαλύτερη συμμετοχή έχει το χαρακτηριστικό “ηλικία” και μετά το “φύλο”, τρίτο έρχεται το χαρακτηριστικό “πόνος στο στήθος” κ.ο.κ.. Ακόμη παρατηρούμε ότι από το V7 χαρακτηριστικό μέχρι το V14 οι ιδιοτιμές είναι κάτω από το 0 οπότε μπορούμε να υποθέσουμε ότι μας προκαλούν “θόρυβο” στα δεδομένα μας και μπορούν να παραληφθούν.

Φανταστείτε τώρα να είχαμε μία βάση με 1.000.000.000 εγγραφές και 200 χαρακτηρίσματα οπότε θα είχαμε ένα πίνακα 1.000.000.000x200.

Αν από την παραπάνω μέθοδο προέκυπτε ότι τα 100 χαρακτηριστικά είναι θόρυβος τότε θα είχαμε ένα νέο πίνακα 1.000.000.000x100 πολύ πιο μικρό σε σχέση με τον αρχικό μας.

Τέτοια λοιπόν δεδομένα ονομάζονται big data και βρίσκονται παντού μέσα σε βάσεις δεδομένων διαστημικών σταθμών, νοσοκομείων, εργαστήριων κλπ.

Η Ανάλυση Κυρίων Συνιστωσών μας βοηθά να μειώσουμε τις διαστάσεις των δεδομένων μας πριν τα επεξεργαστούμε καθώς η μέθοδος αυτή ανήκει στην προεπεξεργασία των δεδομένων.

Αφού πραγματοποιηθεί η προεπεξεργασία των δεδομένων, γίνεται ευκολότερη η απεικόνιση των δεδομένων αυτών και εν συνεχεία έχουμε τη δυνατότητα εισαγωγής αυτών των δεδομένων σε ένα τεχνητό νευρωνικό δίκτυο. Εισάγοντας τα δεδομένα σε αυτό το τεχνητό νευρωνικό δίκτυο, αυτό εκπαιδεύεται με βάση τα δεδομένα αυτά και είναι πλέον εφικτή η κατηγοριοποίηση οποιωνδήποτε νέων δεδομένων έρθουν.

Η μέθοδος αυτή έχει υπολογιστικό κόστος άλλα κερδίζουμε σε υπολογιστικό κόστος στην μετέπειτα επεξεργασία των δεδομένων μας όπως προαναφέραμε.

Τέλος, η μέθοδος αυτή δεν βρίσκει εφαρμογή μόνο σε δεδομένα τα οποία εξετάσαμε αλλά και σε άλλου τύπου δεδομένα όπως π.χ. τα διαστηματικά δεδομένα και άλλα.

5.3 Παράδειγμα PCA σε δεδομένα ασθενών της νόσου Parkinson

Το παράδειγμα που θα εξετάσουμε περιέχει ένα set δεδομένων (dataset) με μετρήσεις που έγιναν σε ασθενείς πριν διαγνωστούν με τη νόσο του Πάρκινσον.

Το σύνολο δεδομένων (dataset) αυτό μπορεί να το βρει κανείς στο [link \(http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring\)](http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring).

Στο παραπάνω dataset έχουν γίνει οι μετρήσεις των εξής χαρακτηριστικών.

V1="age" (ηλικία)

V2="sex" (φίλο)

V3="test time" (ο χρόνος από την έναρξη της παρακολούθησης σε ημέρες)

V4="motor UPDRS" (κλίμακα αξιολόγησης της νόσου (UPDRS) χρησιμοποιείται για να ακολουθήσει το διαμήκη πορεία της νόσου του Πάρκινσον, εδώ μετρούμε την κίνηση της νόσου)

V5="total UPDRS" (εδώ μετριέται η συνολική πορεία της νόσου)

V6="Jitter %" (τρέμουλο σε ποσοστό %)

V7="Jitter (abs)" (τρέμουλο σε τιμή)

V8="Jitter RAP" (τρέμουλο κουβεντιάσματος)

V9="Jitter PPQ5"

V10="Jitter DDP" (: διάφορα ήδη τρέμουλου)

V11="Shimmer" (τρεμόπαιγμα)

V12="Shimmer db" (τρεμόπαιγμα σε db)

V13="Shimmer APQ3"

V14="Shimmer APQ5"

V15="Shimmer APQ11"

V16="Shimmer DDA" (:διάφορα ήδη τρέμουλου)

V17="NHP" (δείκτης θορύβου τονικών χαρακτηριστικών φωνής)

V18="HNP" (δείκτης θορύβου τονικών χαρακτηριστικών φωνής)

V19="RPDE"

V20="DFA"

V21="PPE"

Από τα παραπάνω εικοσιένα χαρακτηριστικά προκύπτουν οι παρακάτω εικοσιένα ιδιότητες.

U\$values

[V1] 2.857508e+03
 [V2] 1.870588e+02
 [V3] 6.656389e+01
 [V4] 1.780441e+01
 [V5] 4.369510e+00
 [V6] 2.073941e-01
 [V7] 2.036893e-02
 [V8] 5.532623e-03
 [V9] 4.920669e-03
 [V10] 2.618922e-03
 [V11] 9.952229e-04
 [V12] 1.028085e-04
 [V13] 4.516101e-05
 [V14] 3.208330e-05
 [V15] 6.777316e-06
 [V16] 3.392175e-06
 [V17] 6.903599e-07
 [V18] 2.735362e-07
 [V19] 1.622420e-10
 [V20] 6.687459e-12
 [V21] 6.519241e-12

Συμπεράσματα

Από την μέθοδο όπως έχει παρουσιασθεί έως τώρα προκύπτει ότι θα πρέπει να αφαιρεθούν τα χαρακτηριστικά εκείνα με ιδιοτιμές μικρότερες του 1. Οπότε θα πρέπει να μειώσουμε τις 16 τελευταίες διαστάσεις. Ένα ενδιαφέρον ερώτημα είναι αν όντως είναι σωστό να γίνει αυτή η αφαίρεση.

Εδώ λοιπόν είναι που βρίσκει πάτημα το ερευνητικό κομμάτι αυτής της πτυχιακής εργασίας. Έχοντας ασχοληθεί και στο παρελθόν με την νόσο Parkinson ήρθαμε σε επαφή με τον κύριο Κωνσταντίνο Πατεράκη Επίκουρο καθηγητή και νευροχειρουργό με έδρα στην Ιατρική σχολή του πανεπιστημίου Θεσσαλίας.

Ο Κ.Πατεράκης μέσω συνέντευξης μας ανέφερε ότι μερικές από της παραπάνω μετρήσεις γίνονται προκειμένου να βοηθήσουν το γιατρό αφενός να σχηματίσει μια σφαιρική άποψη για τον εκάστοτε ασθενή με βάση το πλήθος των μετρήσεων των παραπάνω χαρακτηριστικών, αφετέρου να μπορέσει να συγκρίνει τα χαρακτηριστικά αυτά με άλλες μέτρησης παρόμοιων περιστατικών. Καταλήγοντας στο συμπέρασμα ότι σκοπός του ιατρού είναι να αποκτήσει μια μικρή εξειδίκευση σε κάθε ένα από τους ασθενής του για την σωστή ερμηνεία της πάθησης που μπορεί να εμφανίσει καθώς η νόσος του Παρακινήσουν είναι πολύ κοντά με άλλες εγκεφαλικές δυσλειτουργίες και είναι ιδιαίτερα δύσκολο να διακρίνουμε τις διαφορές τους. Ακόμη ο Κ.Πατεράκης μας ανέφερε ότι το να μειώσουμε εμείς τις διαστάσεις απαλείφοντας εκείνες με ιδιοτιμές μικρότερες του 1 δεν είναι λάθος καθώς είναι στην κρίση του ιατρού αφοτου έχει τα

αποτελέσματα μας στα χέρια του να επιλέξει εκείνα τα χαρακτηριστικά που θεωρεί αυτός σημαντικά.

Εν κατακλείδι σκοπός της ΑΚΣ είναι να μας βοηθήσει να εκπαιδύσουμε ένα τεχνητό νευρωνικό δίκτυο το οποίο δεν θα εξειδικεύεται σε ένα πλήθος ασθενών αλλά σε ένα μεγαλύτερο καθώς όπως αναφέραμε σε πολλές εγκεφαλικές δυσλειτουργίες οι ασθενείς εμφανίζουν παρόμοια συμπτώματα και η εξειδίκευση του τεχνητού νευρωνικού δικτύου σε μια από αυτές καθιστά δυσκολότερη την εξαγωγή ασφαλών συμπερασμάτων, μειώνοντας της διαστάσεις εκείνες που προκαλούν «θόρυβο». Για τον παραπάνω λοιπόν λόγο καταλήξαμε με τον Κ.Πατεράκη στο συμπέρασμα ότι μπορούμε να παραλείψουμε τα χαρακτηριστικά εκείνα με ιδιότητες μικρότερες του 1 και να επιτύχουμε την μείωση των διαστάσεων και την αποτροπή της εξειδίκευσης του τεχνητού νευρωνικού δικτύου έτσι ώστε να οδηγηθούμε στην κατασκευή ενός ασφαλέστερου συστήματος υποστήριξης απόφασης για τον ιατρό.

Βιβλιογραφία και Αναφορές Πέμπτου Κεφαλαίου

[1] Θεοδωρίδης Σέργιος, 2011, *Εισαγωγή στην αναγνώριση προτύπων με Matlab*, 2013, Ιατρικές Εκδόσεις Π. Χ. Πασχαλίδης, Αθήνα.

Λεξικό όρων

Απόκλιση: στατιστικός όρος που προσδιορίζει διάφορες αριθμητικές στατιστικές αποκλίσεις.

Γραμμική Παλινδρόμηση: είναι μια ευρέως χρησιμοποιημένη στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Χρησιμοποιείται με σκοπό την εκχώρηση δεδομένων σε μία πραγματική μεταβλητή πρόβλεψης, όπως ισχύει και στην περίπτωση της κατηγοριοποίησης όταν είναι διακριτή, αλλιώς καλείται παλινδρόμηση αν η μεταβλητή είναι συνεχής. Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα που έχουν δοθεί. Αποτέλεσμα της παλινδρόμησης όταν χρησιμοποιείται ως τεχνική εξόρυξης δεδομένων, αποτελεί ένα μοντέλο που χρησιμοποιείται αργότερα για να προβλέψει τις τιμές της κατηγορίας για τα νέα δεδομένα. Τέτοια παραδείγματα εφαρμογής της παλινδρόμησης αποτελεί η πρόβλεψη της ζήτησης για ένα νέο προϊόν ή υπηρεσία συναρτήσει των δαπανών διαφήμισης ή ο υπολογισμός της ταχύτητας του ανέμου σε σχέση με την θερμοκρασία, την υγρασία και την ατμοσφαιρική πίεση του περιβάλλοντος.

Εξόρυξη δεδομένων: Οι τεχνικές της εξόρυξης δεδομένων επιτρέπουν την επεξεργασία δεδομένων, συνήθως αποθηκευμένων σε βάσεις δεδομένων OLAP, με σκοπό την ανεύρεση ενός ή περισσότερων μοτίβων με τα οποία διαφορετικά δεδομένα μπορούν να συσχετιστούν. Με αυτό τον τρόπο μπορούν να εξαχθούν συμπεράσματα ή πληροφορίες που πρωτίστως φαίνονταν να είναι άσχετες με τα χρησιμοποιούμενα δεδομένα. Με μια τεχνική εξόρυξης δεδομένων αποδείχτηκε ότι η κατανάλωση ενός πελάτη μιας εταιρίας τηλεπικοινωνιών σχετίζεται με το αν θα πληρώσει τον επόμενο λογαριασμό του.

Ιδιοβάση: η βάση η οποία συντίθεται από τα ιδιοδιανύσματα μια γραμμικής απεικόνισης.

Ιδιοδιάνυσμα: ενός τετραγωνικού πίνακα A είναι ένα μη μηδενικό διάνυσμα v που, όταν πολλαπλασιαστεί με τον A , ισούται με το αρχικό διάνυσμα, πολλαπλασιασμένο με έναν αριθμό A , έτσι ώστε:

$$Av = \lambda v$$

Ιδιοτιμή: Ο αριθμός λ ονομάζεται ιδιοτιμή του A που αντιστοιχεί στο v . Οποίος μας δίνει την συμμετοχή των ιδιοδιανυσμάτων κάθε χαρακτηριστικού.

Μηδενοχώρος: χορός μηδενικών διαστάσεων.

Ορθομοναδιαίος: πίνακας με πραγματικά στοιχεία του οποίου οι γραμμές και οι στήλες είναι ορθογώνια μοναδιαία διανύσματα.

Ορθογώνια χαρακτηριστικά: τα στοιχεία εκείνα που βρίσκονται στην κύρια διαγώνιο ενός τετραγωνικού πίνακα.

Παραγοντοποίηση: η διαδικασία με την οποία μια παράσταση, που είναι άθροισμα, μετατρέπεται σε γινόμενο παραγόντων.

Συσταδοποίηση: η διαδικασία εκείνη κατά την οποία ένα σύνολο από «αντικείμενα», διαχωρίζονται σε ένα σύνολο από λογικές ομάδες. Η καταχώρηση αντικειμένων σε ίδια ομάδα μεταφράζεται ως ομοιότητα των αντικειμένων αυτών και αντίστροφα (αντικείμενα που ανήκουν σε διαφορετικές ομάδες είναι ανόμοια). Η ομοιότητα ή μη, μεταξύ των αντικειμένων, ουσιαστικά εξαρτάται από το συγκεκριμένο πρόβλημα και τη μορφή των «αντικειμένων». Τα αντικείμενα μπορούν να αναφερθούν και αυτά με διαφορετικούς όρους: πρότυπα, διανύσματα.

Συνδιασπορά ή Συνδιακύμανση ή Συμεταβλητότητα: το ποσό της ταυτόχρονης μεταβλητότητας δύο ή και περισσότερων μεταβλητών από τις μέσες τιμές τους.

Ψευδοαντίστροφος ή Γενικευμένη αντίστροφος: ενός πίνακα A είναι μια γενίκευση της αντίστροφης μήτρας. Ο πιο ευρέως γνωστός τύπος γενικευμένης αντίστροφης μήτρας είναι η ψευδοαντίστροφος Moore-Penrose, η οποία ανεξάρτητα περιγράφεται από τον EH Moore το 1920, Arne Bjerhammar το 1951 και Roger Penrose το 1955. Νωρίτερα ο Fredholm είχε εισαγάγει την έννοια του ψευδοαντιστρόφου το 1903. Όταν αναφερόμαστε σε μια μήτρα, ο όρος ψευδοαντίστροφος συχνά χρησιμοποιείται για να δείξει την ψευδοαντίστροφο Moore-Penrose. Ο όρος γενικευμένη αντίστροφος μερικές φορές χρησιμοποιείται ως συνώνυμο της ψευδοαντιστρόφου. Μια κοινή χρήση της ψευδοαντιστρόφου είναι να υπολογίσει μια «καλύτερη προσαρμοσμένη» λύση σε ένα σύστημα γραμμικών εξισώσεων που στερείται μια μοναδική λύση. Η ψευδοαντίστροφος διευκολύνει τη δήλωση και την απόδειξη των αποτελεσμάτων στο πεδίο της γραμμικής άλγεβρας.

Ψυχομετρία ή Ψυχομετρική: εφαρμοσμένη επιστήμη που μελετά τη μέτρηση ψυχολογικών χαρακτηριστικών.