Πανεπιστήμιο Πατρών
Τμήμα Οικονομικών Επιστημών
Πρόγραμμα Μεταπτυχιακών Σπουδών «Εφαρμοσμένη Οικονομική και Ανάλυση Δεδομένων»

Ακαδημαϊκό έτος 2023- 2024

1η Εργασία μαθήματος «Διαχείριση Μεγάλων Δεδομένων»

"A human being should be able to change a diaper, plan an invasion, butcher a hog, conn a ship, design a building, write a sonnet, balance accounts, build a wall, set a bone, comfort the dying, take orders, give orders, cooperate, act alone, solve equations, analyze a new problem, pitch manure, program a computer, cook a tasty meal, fight efficiently, die gallantly. Specialization is for insects."

- Robert Heinlein, Time Enough for Love

https://www.youtube.com/watch?v=GJDNkVDGM_s

- "There is no need to be upset"

Εισαγωγή

Σκοπός της εργασίας είναι να γίνει μία πρώτη εξοικείωση με το περιβάλλον R και Python για να εκτελέσετε τη μέθοδο Ανάλυση Κύρων Συνιστωστών (PCA) καθώς επίσης και να υλοποιήσετε μέτρα αποστάσεων και ομοιότητας χρήσιμα στην επεξεργασία δεδομένων. Ο κώδικας που σας ζητείται να γράψετε θα πρέπει να είναι στη γλώσσα προγραμματισμού R και Python.

Προκειμένου να εξοικειωθείτε με τη χρήση της R και της Python για την ανάγνωση αρχείων, επεξεργασία δεδομένων με τις ειδικές βιβλιοθήκες που αναφέρονται στην εκφώνηση καθώς επίσης για την δημιουργία συναρτήσεων (functions) στα περιβάλλοντα αυτά, **συστήνεται να μελετήσετε οπωσδήποτε** τα αρχεία που αναφέρονται στον παρακάτω πίνακα, που θα τα βρείτε στο ιστότοπο του μαθήματος στο eclass (στις ενότητες του μαθήματος LAB - R: Εγκατάσταση και παραδείγματα χρήσης και LAB - Python: Εγκατάσταση και παραδείγματα χρήσης) και τα οποία δίνουν απλά και συνοπτικά σχετικά παραδείγματα επεξεργασίας των δεδομένων με τον ζητούμενο τρόπο:

	R	Python
Ανάγνωση αρχείων csv	R-ReadingCSVFiles.R.rar και R-example.R	workingWithPandas.rar
Δημιουργία και χειρισμός διανυσμάτων	R-IntroVarVecListsFunctions.R.txt	workingWithNumpyArrays.py
Δημιουργία και χρήση συναρτήσεων	R-IntroVarVecListsFunctions.R.txt	workingWithNumpyArrays.py

Χειρισμός δεδομένων με χρήση data frame	subsetting-data.R	workingWithPandas.rar και python-examples.rar (αρχείο subsetting-data.py)

Τα αρχεία του παραπάνω πίνακα περιέχουν πηγαίο κώδικα σε R και Python που επιδεικνύουν τις εντολές που θα σας είναι χρήσιμες για την εκπόνηση της εργασίας. Περιέχουν σχόλια εργασία εξηγώντας τι κάνουν και πως λειτουργούν.
Επιπλέον, στις παραπάνω ενότητες στο eclass θα βρείτε επίσης εισαγωγικές σημειώσεις και βιβλία για τη μελέτη και εξοικείωση με τις γλώσσες R και Python. Συστήνεται να τις μελετήσετε καθότι θα σας βοηθήσουν στην κατανόηση των εννοιών που θα συναντήσετε.

Θέμα 1

Απαντήστε, ως ομάδα, στις online ερωτήσεις της άσκησης που έχει δημοσιευτεί στην ιστοσελίδα του μαθήματος στο eclass. Η άσκηση έχει τίτλο *«Ερωτήσεις θέματος 1) της 1ης εργασίας ακ. Έτους 2023-2024»* και μπορεί να βρεθεί ακολουθώντας τον εξής σύνδεσμο:

https://eclass.upatras.gr/modules/exercise/exercise_submit.php?course=ECON1332&exerciseId=7403

Θέμα 2

Μαζί με την εκφώνηση της εργασίας, θα βρείτε δύο δημοσιεύσεις (αρχεία paper1.pdf και paper2.pdf). Αφού διαβάσετε τις δημοσιεύσεις αυτές, γράψτε μία σύντομη περίληψη (όχι παραπάνω από 750-800 λέξεις) για κάθε μία. Η περίληψη θα πρέπει απαραιτήτως να απαντά στα ακόλουθα ερωτήματα:

- Ποιος είναι ο σκοπός της έρευνας;
- Ποιες στατιστικές μέθοδοι χρησιμοποιήθηκαν;
- Τί δεδομένα χρησιμοποιήθηκαν και από ποια πηγή βρήκαν τα δεδομένα ;
- Ποιο ήταν το πλήθος των παρατηρήσεων και η διάσταση των δεδομένων που χρησιμοποιήθηκαν ;
- Για ποιον λόγο χρησιμοποιήθηκε η μέθοδος της Ανάλυσης Κυρίων Συνιστωσών (PCA);
- Πως χρησιμοποιήθηκαν ή ερμηνεύτηκαν τα αποτελέσματα της Ανάλυσης Κυρίων Συνιστωσών;
- Τι νέο/καινούργιο συνεισφέρει η δημοσίευση; (ΣΗΜ: όπως αναφέρει η ίδια)

Θέμα 3

Γράψτε πρόγραμμα σε R και Python που να εκτελεί τον αλγόριθμο Ανάλυσης Κύριων Συνιστωσών (PCA) σε σύνολα δεδομένων που περιγράφονται παρακάτω. Ειδικότερα:

1) Το πρόγραμμα που θα γράψετε σε R θα πρέπει να εκτελεί τον αλγόριθμο PCA στο σύνολο δεδομένων Cars93, που δίνεται έτοιμο από τη βιβλιοθήκη MASS της R. Εκτελέσετε PCA χρησιμοποιώντας πίνακα συνδιακύμανσης (covariance matrix) στο σύνολο δεδομένων Cars93.

2) Το πρόγραμμα που θα γράψετε σε Python θα πρέπει να εκτελεί τον αλγόριθμο PCA στο σύνολο δεδομένων “Wine Quality Data Set”. Για να κατεβάσετε το σύνολο δεδομένων θα πρέπει να επισκεφτείτε την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/Wine+Quality και τα κατεβάσετε το αρχείο δεδομένων με όνομα winequality-white.csv, που περιέχει παρατηρήσεις για την ποιότητα 4898 λευκών κρασιών. Ο κώδικας Python που θα συγγράψετε θα πρέπει απαραιτήτως να κάνει χρήση της βιβλιοθήκης pandas για την ανάγνωση του αρχείου

δεδομένων. Για τη δημιουργία του προγράμματός σας κάνετε χρήση της υλοποίησης του αλγορίθμου PCA που υπάρχει στη βιβλιοθήκη scikit-learn, αφού βεβαιωθείτε ότι έχει εγκατασταθεί στο σύστημά σας.

Και τα δύο προγράμματα (R και Python) θα πρέπει, αφού εκτελέσουν τον αλγόριθμο PCA στα σύνολα δεδομένων που αναφέρθηκαν παραπάνω, να εμφανίζουν στην οθόνη τα παρακάτω στοιχεία:

a) Όλες τις ιδιοτιμές των ιδιοδιανυσμάτων που προκύπτουν από την εκτέλεση της μεθόδου PCA
b) Όλα τα ιδιοδιανύσματα που προκύπτουν από την εκτέλεση της μεθόδου PCA
c) Το ποσοστό της διακύμανσης που εξηγεί η κάθε κύρια συνιστώσα που προέκυψε από τον αλγόριθμο PCA.

Θέμα 4

Γράψτε σε R και Python **μία συνάρτηση** με όνομα euclideanDistance, η οποία δέχεται ως ως είσοδο δύο διανύσματα και εμφανίζει στην οθόνη την Ευκλείδεια απόσταση των δύο διανυσμάτων διαστάσεων n, που δίνονται ως όρισμα. Θεωρήστε ότι τα διανύσματα έχουν μόνο πραγματικές τιμές. Επιπλέον, ζητούνται τα εξής:

I) Για κάθε ζεύγος διανυσμάτων x και y που εμφανίζονται παρακάτω, υπολογίστε την Ευκλείδεια απόστασή τους χρησιμοποιώντας τη συνάρτηση euclideanDistance που έχετε δημιουργήσει σε R και Python, καλώντας την με τα κατάλληλα ορίσματα.

a) x=(1,2,3,4,5,6) και y= (1,2,3,4,5,6)
b) x=(-0.5, 1, 7.3, 7, 9.4, -8.2, 9, -6, -6.3) και y=(0.5, -1, -7.3, -7, -9.4, 8.2, -9, 6, 6.3)
c) x=(-0.5, 1, 7.3, 7, 9.4, -8.2) και y=(1.25, 9.02, -7.3, -7, 5, 1.3)
d) x=(0, 0, 0.2) και y=(0.2, 0.2, 0)

Ο κώδικάς σας τόσο στο περιβάλλον R όσο και στην Python θα πρέπει να φαίνεται ο ορισμός της συνάρτησης euclideanDistance που έχετε δημιουργήσει καθώς επίσης και η κλήση της συνάρτησης euclideanDistance για τον υπολογισμό των αποστάσεων των διανυσμάτων που δίνονται στην εκφώνηση. Το πρόγραμμά σας θα πρέπει να εμφανίζει στην οθόνη την Ευκλείδεια απόσταση των τεσσάρων ζευγών διανυσμάτων που δίνονται παραπάνω.

II) Δίνεται ο παρακάτω πίνακας, ο οποίος εμφανίζει το προφίλ χρηστών (γραμμές) μιας εταιρείας κινητής τηλεφωνίας. Η στήλη «Χρήστης» περιέχει έναν μοναδικό κωδικό για κάθε χρήστη:

ΑΑ χρήστη	Διάρκεια ομιλίας (σε λεπτά)	Πλήθος SMS που έχουν σταλεί	Χρήση Internet (σε MB)
1	25000	14	7
2	42000	17	9
3	55000	22	5
4	27000	13	11
5	58000	21	13

A. Χρησιμοποιώντας τη συνάρτηση υπολογισμού της Ευκλείδειας απόστασης euclideanDistance που έχετε δημιουργήσει στο περιβάλλοντης R και σε Python, κάντε τους απαραίτητους υπολογισμούς (στο περιβάλλον της R και Python) για να βρείτε το προφίλ εκείνου του χρήστη, το οποίο μοιάζει περισσότερο με το προφίλ του χρήστη με κωδικό 5.

B. Από την απάντηση και τις τιμές που βρήκατε στο ερώτημα Α), τί θα απαντούσατε αν σας ρωτούσε κάποιος το εξής: «Ποιες τιμές επηρεάζουν περισσότερο την τιμή της Ευκλείδειας απόστασης που προκύπτει σε κάθε περίπτωση;»

Θέμα 5

Γράψτε στο περιβάλλον R και Python **μία συνάρτηση** με όνομα cosineSimilarity η οποία δέχεται ως είσοδο δύο διανύσματα και υπολογίζει ή επιστρέφει την ομοιότητα συνημιτόνου των δύο διανυσμάτων. Θεωρήστε ότι τα διανύσματα έχουν μόνο πραγματικές τιμές. Επιπλέον, ζητούνται τα εξής:

Για κάθε ζεύγος διανυσμάτων x και y που εμφανίζονται παρακάτω, υπολογίστε την ομοιότητα συνημίτονου των διανυσμάτων που αναφέρονται παρακάτω σε κάθε περίπτωση, χρησιμοποιώντας τη συνάρτηση cosineSimilarity που έχετε ορίσει, καλώντας την με τα κατάλληλα ορίσματα στο περιβάλλον R και Python. Το πρόγραμμά σας θα πρέπει επίσης να εμφανίζει στην οθόνη την ομοιότητα συνημιτονου των διανυσμάτων.

a) x=(9.32, -8.3, 0.2) και y= (-5.3, 8.2, 7)
b) x=(6.5, 1.3, 0.3, 16, 2.4, -5.2, 2, -6, -6.3) και y=(0.5, -1, -7.3, -7, -9.4, 8.2, -9, 6, 6.3)
c) x=(-0.5, 1, 7.3, 7, 9.4, -8.2) και y=(1.25, 9.02, -7.3, -7, 15, 12.3)
d) x=(2, 8, 5.2) και y=(2, 8, 5.2)

Στον κώδικα R και Python που θα παραδώσετε θα πρέπει να φαίνεται ο ορισμός της συνάρτησης cosineSimilarity καθώς επίσης και η κλήση της συνάρτησης cosineSimilarity με τα κατάλληλα ορίσματα για τον υπολογισμό των αποστάσεων των διανυσμάτων που δίνονται στην εκφώνηση.

Θέμα 6

Γράψτε στο περιβάλλον R και σε Python **μία συνάρτηση** με όνομα nominalDistance, η οποία δέχεται ως είσοδο δύο διανύσματα και υπολογίζει την απόστασή τους. Τα διανύσματα που δίνονται ως είσοδο στις συναρτήσεις θα έχουν μόνο ονομαστικές (nominal) τιμές. Ορίστε την συνάρτηση nominalDistance σε R και Python όπως εσείς κρίνετε εσείς σκόπιμο.

Ακολούθως, για κάθε ζεύγος διανυσμάτων x και y που εμφανίζονται παρακάτω, υπολογίστε την απόστασή τους χρησιμοποιώντας τη συνάρτηση nominalDistance που έχετε ορίσει στην R και στην Python, καλώντας την με τα κατάλληλα ορίσματα στο περιβάλλον R.

a) x=(“Green”, “Potato”, “Ford”) και y=(“Tyrian purple”, “Pasta”, “Opel”)
b) x=(“Eagle”, “Ronaldo”, “Real madrid”, “Prussian blue”, “Michael Bay”) και y=(“Eagle”, “Ronaldo”, “Real madrid”, “Prussian blue”, “Michael Bay”)
c) x=(“Werner Herzog”, “Aquirre, the wrath of God”, “Audi”, “Spanish red”) και y=(“Martin Scorsese”, “Taxi driver”, “Toyota”, “Spanish red”)

Τα προγράμματά σας σε R και Python που θα δημιουργήσετε θα πρέπει για κάθε ζεύγος διανυσμάτων να εμφανίζει στην οθόνη την απόστασή τους χρησιμοποιώντας τη συνάρτηση nominalDistance που έχετε ορίσει.

Ομάδες εργασίας

Η εργασία θα εκπονηθεί ομαδικά. Γι’αυτό τον λόγο θα πρέπει να σχηματίσετε ομάδες 3 ή 4 ατόμων της αρεσκείας σας (**όχι λιγότερο από 3 και όχι περισσότερα από 4 άτομα σε κάθε ομάδα**). Σε περίπτωση που υπάρχει οποιοδήποτε πρόβλημα με τη δημιουργία ομάδας, θα πρέπει **να επικοινωνήσετε άμεσα με τον διδάσκοντα.**
Κάθε ομάδα εργασίας πρέπει να δηλωθεί με email **που θα στείλει ένας από κάθε ομάδα** στη διεύθυνση tzagara@upatras.gr **.** Με τον όρο δήλωση ομάδας εννοείται να ενημερώσετε

τον διδάσκοντα του μαθήματος για το ποια άτομα αποτελούν την ομάδα εργασίας. Το email δήλωσης ομάδας εργασίας, θα πρέπει για κάθε άτομο της ομάδας να αναφέρει τα εξής στοιχεία: το **όνομά του, τον ΑΜ του και το email του (στο domain .upnet.gr). Ένα άτομο από κάθε ομάδα θα στείλει ένα τέτοιο μήνυμα δήλωσης ομάδας για λογαριασμό της ομάδας.**

Το email που θα αποσταλεί για τη δήλωση ομάδας εργασίας πρέπει απαραιτήτως να έχει ως θέμα (subject) *«ΠΜΣ: Δήλωση ομάδας εργασίας Διαχείριση Μεγάλων Δεδομένων ακ. έτους 2023-2024»* (προσοχή! Δίχως τα « »).

Δείγμα ενός τέτοιου email δήλωσης ομάδας εργασίας φαίνεται παρακάτω:

> *Από: up-424242@upnet.gr*
> *Προς (To): tzagara@upatras.gr,*
> *Θέμα (Subject): ΠΜΣ: Δήλωση ομάδας εργασίας Διαχείριση Μεγάλων Δεδομένων ακ. έτους 2023-2024*
>
> *Η ομάδα εργασίας αποτελείται από τους εξής φοιτητές:*
>
> *Σάκης Ρουβάς, -424242, up424242@upnet.gr*
> *Δέσποινα Μαλέα, -872238976, up102340@upnet.gr*
> *Αντώνης Πασχαλίδης, 000000000, up105522@upnet.gr*
> *Michael Bay, 123456789, kabooooom@upnet.gr*
>
> *Thnx!*

Αφού στείλετε το μήνυμα δήλωσης ομάδας, θα λάβετε ως απάντηση ένα email με πληροφορίες για έναν ειδικό ιδιωτικό χώρο που θα δημιουργηθεί για κάθε ομάδα, στον οποίο θα έχουν πρόσβαση μόνο τα μέλη της ομάδα και όπου η ομάδα μπορεί να υποβάλλει/παραδώσει την εργασία της.
Οι ομάδες δεν αλλάζουν κατά τη διάρκεια του μαθήματος: η δήλωση της κάθε ομάδας γίνεται μία φορά το εξάμηνο και όχι για κάθε εργασία. Έτσι η ίδια ομάδα ατόμων, θα παραδώσει όλες τις εργασίες του μαθήματος.

ΠΡΟΣΟΧΗ! Email δήλωσης ομάδας εργασίας θα πρέπει να σταλεί το αργότερο έως και Πέμπτη, 26 Οκτωβρίου 2023 και ώρα 00:00:00.

Οι ομάδες δεν αλλάζουν μεταξύ των εργασιών. Η ίδια ομάδα θα εκτελέσει και τις υπόλοιπες εργασίες του μαθήματος.

Παράδοση της εργασίας: Τί και πως;

Κάθε ομάδα θα πρέπει να παραδώσει μία αναφορά σε αρχείο μορφής .pdf, γραμμένη σε LaTeχ, η οποία θα περιέχει τις απαντήσεις στα ζητούμενα των θεμάτων και μέσα στην ίδια αναφορά τον κώδικα σε R και Python που υπολογίζει τα ζητούμενα της εκφώνησης. Επιπλέον, ο κώδικας σε R και Python που έχετε δημιουργήσει για όλα τα θέματα, θα πρέπει να σταλεί και σε ξεχωριστά αρχεία (ένα ή περισσότερα) με τη μορφή κειμένου, ώστε να μπορεί να εκτελείται. Το αρχείο της αναφοράς καθώς και τα αρχεία με τον κώδικα R και Python θα πρέπει να συμπιεστούν σε ένα αρχείο (μορφή .zip ή .rar) και να σταλούν στον διδάσκοντα.
Αναλυτικότερες πληροφορίες για τον τρόπο παράδοσης θα δοθεί κατά τη διάρκεια των διαλέξεων.

ΠΡΟΣΟΧΗ! Καταληκτική ημερομηνία παράδοσης της 1ης εργασίας είναι: Πέμπτη 2 Νοεμβρίου 2023.

Ερωτήσεις/Απορίες

Για οποιαδήποτε ερώτηση ή απορία σχετικά με την εργασία μπορείτε να στείλετε email στη διεύθυνση tzagara@upatras.gr . Απορίες μπορούν επίσης (**και συστήνεται!**) να συζητηθούν κατά τη διάρκεια του μαθήματος.

Βαρύτητα της εργασίας

Η 1η εργασία είναι υποχρεωτική για τους φοιτητές και συνεισφέρει το 10% του τελικού τους βαθμού.

Καλή επιτυχία!