Εργασία στο μάθημα 'Ανάλυση Δεδομένων', Ιανουάριος 2020

Δημήτρης Κουγιουμτζής

E-mail: dkugiu@auth.gr

11 Ιανουαρίου 2020

Οδηγίες: Σχετικά με την παράδοση της εργασίας θα πρέπει:

- Για κάθε ζήτημα θα δημιουργήσετε ένα ή περισσότερα προγράμματα και συναρτήσεις Matlab. Τα ονόματα τους θα είναι ως εξής, όπου ως παράδειγμα δίνεται η ομάδα φοιτητών Νο 10 και το ζήτημα 5. Για τα προγράμματα τα ονόματα των αρχείων θα είναι Group10Exe5Prog1.m, Group10Exe5Prog2.m κτλ. Για τις συναρτήσεις τα ονόματα των αρχείων θα είναι Group10Exe5Fun1.m, Group10Exe5Fun2.m κτλ. Στην αρχή κάθε προγράμματος θα υπάρχουν (σε σχολιασμό) τα ονοματεπώνυμα των μελών της ομάδας.
- Τα προγράμματα θα πρέπει να είναι εκτελέσιμα και η εκτέλεση τους να δίνει τις απαντήσεις που ζητούνται σε κάθε ζήτημα. Επεξηγήσεις, σχολιασμοί αποτελεσμάτων και συμπεράσματα, όπου ζητούνται, θα δίνονται με μορφή σχολίων στο πρόγραμμα (τα συμπεράσματα στο τέλος του προγράμματος).
- Θα υποβληθούν μόνο τα αρχεία Matlab (μέσω του elearning).
- Η κάθε εργασία (σύνολο προγραμμάτων και συναρτήσεων Matlab) θα πρέπει να συντάσσεται αυτόνομα από την ομάδα. Ομοιότητες εργασιών θα οδηγούν σε μοίρασμα της βαθμολογίας (δύο 'όμοιες' άριστες εργασίες θα μοιράζονται το βαθμό δια δύο, τρεις δια τρία κτλ.).

Ζητήματα εργασίας

Η εργασία αφορά δασικές εκτάσεις, όπου κάποιες είναι καμμένες, και δίνονται για κάθε έκταση γεωγραφικοί, εποχικοί και περιβαλλοντικοί δείκτες.

Η μελέτη έγινε σε μια περιοχή της Πορτογαλίας (Montesinho). Τα δεδομένα έχουν δοθεί από την πηγή: P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimaraes, Portugal, pp. 512-523, 2007. (το άρθρο είναι διαθέσιμο στη διεύθυνση http://www3.dsi.uminho.pt/pcortez/fires.pdf)

Η οργάνωση των δεδομένων δίνεται παρακάτω:

Α/Α	*Όνομα*	*Περιγραφή*
1.	X	x-axis spatial coordinate within the Montesinho park map: 1 to 9
2.	Y	y-axis spatial coordinate within the Montesinho park map: 2 to 9
3.	month	month of the year: number from 1 to 12 corresponding from "jan" to "dec"
4.	day	day of the week: from 1 for Monday to 7 for Sunday
5.	FFMC	FFMC index from the FWI system: 18.7 to 96.20
6.	DMC	DMC index from the FWI system: 1.1 to 291.3
7.	DC	DC index from the FWI system: 7.9 to 860.6
8.	ISI	ISI index from the FWI system: 0.0 to 56.10
9.	temp	temperature in Celsius degrees: 2.2 to 33.30
10.	RH	relative humidity in %: 15.0 to 100
11.	wind	wind speed in km/h: 0.40 to 9.40
12.	rain	outside rain in mm/m2 : 0.0 to 6.4
13.	area	the burned area of the forest (in ha): 0.00 to 1090.84 (this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

Για όλα τα ζητήματα στην αρχή του κάθε προγράμματος θα φορτώνεται το αρχείο των δεδομένων.

1. Θέλουμε να συγκρίνουμε αν η κατανομή δεικτών διαφέρει στις περιπτώσεις εκτάσεων καμμένων δασών και εκτάσεων χωρίς καμμένα δάση και αν η κατανομή σε κάθε περίπτωση είναι κανονική. Χώρισε τα δεδομένα σε δύο υποσύνολα ως εξής. Το πρώτο υποσύνολο, δείγμα Α, θα έχει τις μετρήσεις που αντιστοιχούν σε μηδενική έκταση καμμένων δασών, δηλαδή είναι οι περιπτώσεις που η τελευταία στήλη (area) έχει τιμή μηδέν. Το δεύτερο υποσύνολο, δείγμα Β, θα έχει τις υπόλοιπες μετρήσεις. Θα διερευνήσεις την τυχόν διαφορά ποιοτικά από την κατανομή του δείκτη σε κάθε μια από τις δύο περιπτώσεις. Για αυτό θα σχηματίσεις την καμπύλη της εμπειρικής συνάρτησης πυκνότητας πιθανότητας (σππ) του δείκτη με τη μέθοδο του ιστογράμματος για κατάλληλη ισομερή διαμέριση στις δύο περιπτώσεις (δύο καμπύλες σε ένα σχήμα). Επίσης

για κάθε ένα από τα δύο δείγματα Α και Β θα ελέγξεις (με έλεγχο υπόθεσης καλής προσαρμογής X^2) αν η αντίστοιχη σππ προσεγγίζεται από κανονική κατανομή. Αν δεν προσεγγίζεται από κανονική κατανομή, θα κάνεις τον ίδιο έλεγχο για κατανομή Poisson. Το πρόγραμμα θα εφαρμόζει τα παραπάνω για κάθε έναν από τους δείκτες 9,10,11 (θερμοκρασία (temp), υγρασία (RH) και άνεμο (wind)) και θα δίνει για κάθε δείκτη και σε κάθε μια από τις δύο περιπτώσεις την κατανομή που ακολουθεί. Συμφωνούν οι κατανομές στις δύο περιπτώσεις και για ποιους δείκτες;

2. Σε συνέχεια του Ζητήματος 1, θέλουμε να εξετάσουμε αν οι μέσες τιμές του δείκτη διαφέρουν στις δύο περιπτώσεις (καμμένες και μη-καμμένες δασικές εκτάσεις) και κατά πόσο όταν θεωρούμε το δείγμα όλων των παρατηρήσεων και όταν χρησιμοποιούμε ένα μικρό δείγμα. Με βάση τα δείγματα Α και Β όλων των παρατηρήσεων (δες Ζήτημα 1), υπολόγισε το 95% διάστημα εμπιστοσύνης για τη διαφορά των μέσων τιμών του δείκτη στην περίπτωση μηδενικής καμμένης έκτασης και στην αντίθετη περίπτωση. Κάνε το ίδιο για ένα μεγάλο πλήθος M, π.χ. $M = 50$, μικρότερων δειγμάτων $n = 20$ παρατηρήσεων με τυχαία επιλογή των παρατηρήσεων από το κάθε ένα από τα δείγματα Α και Β αντίστοιχα. Σχημάτισε τα διαστήματα εμπιστοσύνης για τα M μικρά δείγματα και το διάστημα εμπιστοσύνης για τα μεγάλα δείγματα Α και Β (σε ένα ή δύο σχήματα). Συμφωνούν οι απαντήσεις από τα μεγάλα δείγματα Α και Β και τα M μικρά δείγματα στο ερώτημα για την ισότητα των μέσων τιμών του δείκτη στις δύο περιπτώσεις (καμμένες και μη-καμμένες δασικές εκτάσεις); Το πρόγραμμα θα εφαρμόζει τα παραπάνω (και θα απαντάει στο παραπάνω ερώτημα) για κάθε έναν από τους δείκτες 9,10,11 (θερμοκρασία (temp), υγρασία (RH) και άνεμο (wind)).

3. Σε συνέχεια του Ζητήματος 1, θέλουμε να εξετάσουμε αν οι διάμεσοι (median) του δείκτη διαφέρουν στις δύο περιπτώσεις (καμμένες και μη-καμμένες δασικές εκτάσεις) και κατά πόσο όταν θεωρούμε το δείγμα όλων των παρατηρήσεων και όταν χρησιμοποιούμε ένα μικρό δείγμα. Με βάση τα δείγματα Α και Β όλων των παρατηρήσεων (δες Ζήτημα 1), κάνε έλεγχο υπόθεσης (σε επίπεδο σημαντικότητας $\alpha = 0.05$) για την ισότητα των διαμέσων του δείκτη στην περίπτωση μηδενικής καμμένης έκτασης και στην αντίθετη περίπτωση. Για τον έλεγχο θα χρησιμοποιήσεις έλεγχο τυχαιοποίησης (τυχαίας αντιμετάθεσης) **ή** έλεγχο bootstrap δημιουργώντας $B = 1000$ τυχαία δείγματα χωρίς επανάθεση **ή** με επανάθεση, αντίστοιχα. Κάνε το ίδιο για ένα μεγάλο πλήθος M, π.χ. $M = 50$, μικρότερων δειγμάτων $n = 20$ παρατηρήσεων με τυχαία ε-

πιλογή των παρατηρήσεων από το κάθε ένα από τα δείγματα Α και Β αντίστοιχα. Συμφωνούν οι απαντήσεις (απόρριψη ή μη-απόρριψη της μηδενικής υπόθεσης για ίσες διαμέσους του δείκτη στις δύο περιπτώσεις καμμένων και μη-καμμένων δασικών εκτάσεων) από τα μεγάλα δείγματα Α και Β και τα M μικρά δείγματα; Το πρόγραμμα θα εφαρμόζει τα παραπάνω (και θα απαντάει στο παραπάνω ερώτημα) για κάθε έναν από τους δείκτες 9,10,11 (θερμοκρασία (temp), υγρασία (RH) και άνεμο (wind)).

4. Θέλουμε να διερευνήσουμε αν υπάρχει συσχέτιση μεταξύ των δεικτών ανά ζεύγη. Από το σύνολο των παρατηρήσεων, επέλεξε τυχαία ένα μικρό δείγμα 40 παρατηρήσεων για όλους τους δείκτες (δηλαδή 40 εκτάσεις από το σύνολο των εκτάσεων με τις αντίστοιχες τιμές των δεικτών). Με βάση αυτό το δείγμα, κάνε κατάλληλο παραμετρικό έλεγχο για μηδενική συσχέτιση για κάθε ζεύγος, χρησιμοποιώντας το στατιστικό της κατανομής Student. Κάνε επίσης τον αντίστοιχο έλεγχο τυχαιοποίησης χρησιμοποιώντας $B = 1000$ τυχαιοποιημένα δείγματα. Φαίνεται να υπάρχει κάποια συσχέτιση μεταξύ των δεικτών και ποιών με βάση τον παραμετρικό και έλεγχο τυχαιοποίησης για κάθε ζεύγος δεικτών από τους δείκτες 5,...,11 (FFMC,...,wind); Συμφωνούν οι δύο τύποι ελέγχου υπόθεσης;

5. Θέλουμε να διερευνήσουμε αν η υγρασία (RH) και ο άνεμος (wind) εξαρτώνται από τη θερμοκρασία (temp). Με βάση το δείγμα που επέλεξες στο Ζήτημα 4, εκτίμησε το μοντέλο γραμμικής παλινδρόμησης του RH ως προς το temp, υπολόγισε το συντελεστή προσδιορισμού (ή εναλλακτικά τον προσαρμοσμένο συντελεστή προσδιορισμού) και διερεύνησε την καταλληλότητα του μοντέλου (αν τα υπόλοιπα είναι ασυσχέτιστα). Επανέλαβε την παραπάνω ανάλυση για την εξάρτηση του δείκτη wind από το temp. Διαφέρουν τα αποτελέσματα; Σε ποια περίπτωση η προσαρμογή του μοντέλου παλινδρόμησης είναι καλύτερη; Θα ήταν χρήσιμο να επεκτείνουμε το μοντέλο παλινδρόμησης σε πολυωνυμικό κάποιου βαθμού σε κάθε μια από τις δύο περιπτώσεις;

6. Θέλουμε να διερευνήσουμε την ακρίβεια της εκτίμησης του συντελεστή της γραμμικής παλινδρόμησης από μικρό δείγμα για την περίπτωση της γραμμικής εξάρτησης της σχετική υγρασίας (RH) από τη θερμοκρασία (temp). Θεωρούμε ως πραγματικό συντελεστή β της γραμμικής παλινδρόμησης της υγρασίας από τη θερμοκρασία αυτόν που εκτιμούμε με τη μέθοδο ελαχίστων τετραγώνων στο σύνολο των ζευγαρωτών παρατηρήσεων θερμοκρασίας και υγρασίας. Θα δημιουργήσεις $M = 100$ μικρά ζευγαρωτά δείγματα $n = 40$ παρατηρήσεων από το αρχικό σύνολο των

παρατηρήσεων θερμοκρασίας και υγρασίας. Για κάθε δείγμα θα υπολογίσεις το συντελεστή b της ευθείας ελαχίστων τετραγώνων, το αντίστοιχο παραμετρικό 95% διάστημα εμπιστοσύνης, καθώς και το bootstrap 95% διάστημα εμπιστοσύνης. Το bootstrap διάστημα εμπιστοσύνης για το συντελεστή γραμμικής παλινδρόμησης σχηματίζεται με βάση B ζευγαρωτά δείγματα bootstrap, όπου το κάθε δείγμα σχηματίζεται από το σύνολο των ζευγαρωτών παρατηρήσεων με επαναδειγματοληψία με επανάθεση. Θα σχηματίσεις την κατανομή του b (ιστόγραμμα) από τα M μικρά δείγματα και θα μετρήσεις το ποσοστό των αντίστοιχων παραμετρικών και bootstrap διαστημάτων εμπιστοσύνης (σε σύνολο M διαστημάτων) που περιέχουν την τιμή β, η οποία υπολογίστηκε στο σύνολο των παρατηρήσεων. Συμφωνούν τα ποσοστά για το παραμετρικό και bootstrap διάστημα εμπιστοσύνης;

7. Θέλουμε να διερευνήσουμε αν η έκταση δασών (area, καμμένων ή μη) μπορεί να εξαρτάται από κάποιον από τους άλλους παρατηρούμενες δείκτες (στήλες 1 - 12). Η εξάρτηση μπορεί να είναι γραμμική ή μηγραμμική. Κάνε διερεύνηση του μοντέλου παλινδρόμησης της έκτασης ως προς κάποιον δείκτη (με ελεύθερη επιλογή του μοντέλου, γραμμικού, πολυωνυμικού, άλλου) που δίνει την καλύτερη προσαρμογή, δηλαδή δίνει τη μεγαλύτερη τιμή του προσαρμοσμένου συντελεστή προσδιορισμού. Η διερεύνηση θα γίνει ως προς κάθε έναν από τους 12 δείκτες.

8. Θέλουμε να ελέγξουμε το κατάλληλο μοντέλο πολλαπλής γραμμικής παλινδρόμησης για τον άνεμο (wind). Η εφαρμογή του μοντέλου βηματικής παλινδρόμησης στο σύνολο των παρατηρήσεων μπορεί να προσθέσει μεταβλητές που δεν έχουν σημαντική προβλεπτική πληροφορία για τον άνεμο, όπου οι μεταβλητές είναι οι δείκτες 1,...,10,12,13. Πρώτα θα εφαρμόσεις το μοντέλο βηματικής παλινδρόμησης στο σύνολο των παρατηρήσεων και θα κρατήσεις το σύνολο X_0 των επεξηγηματικών μεταβλητών που επιλέχτηκαν. Στη συνέχεια θα δημιουργήσεις $M = 100$ μικρά ζευγαρωτά δείγματα $n = 40$ παρατηρήσεων από το αρχικό σύνολο των παρατηρήσεων. Για κάθε δείγμα $i = 1, \ldots, M$, θα εφαρμόσεις το μοντέλο βηματικής παλινδρόμησης και θα κρατήσεις το σύνολο X_i των επεξηγηματικών μεταβλητών που επιλέχτηκαν. Θα υπολογίσεις πόσο συχνά εμφανίζεται κάθε μεταβλητή από το σύνολο X_0 στα σύνολα X_i, $i = 1, \ldots, M$. Φαίνεται κάποιες μεταβλητές του μοντέλου βηματικής παλινδρόμησης στο σύνολο των παρατηρήσεων (X_0) να εμφανίζονται συχνά στο σύνολο των μεταβλητών του μοντέλου βηματικής παλινδρόμησης στα μικρά δείγματα;

9. Θέλουμε να εξετάσουμε αν οι 13 δείκτες έχουν πληροφορία που μπορεί να παρασταθεί με λιγότερες μεταβλητές, δηλαδή να διερευνήσουμε τη μείωση διάστασης του χώρου των παρατηρήσιμων μεταβλητών. Για αυτό θα εφαρμόσεις την ανάλυση PCA στο σύνολο των δεδομένων, αφού έχουν κανονικοποιηθεί πρώτα, δηλαδή έγινε αφαίρεση με τη μέση τιμή και διαίρεση με την τυπική απόκλιση για κάθε δείκτη ξεχωριστά. Αυτό γίνεται γιατί οι δείκτες έχουν διαφορετικό εύρος τιμών που μπορεί να επηρεάσει την PCA. Θα εκτιμήσεις τη διάσταση $d \leq 13$ για τη μείωση διάστασης με PCA με το λεγόμενο scree plot, όπου χρησιμοποιείται ως κατώφλι η μέση τιμή των ιδιοτιμών $\bar{\lambda}$ (οι ιδιοτιμές που είναι μεγαλύτερες του $\bar{\lambda}$ θεωρούνται σημαντικές και το πλήθος τους συνιστά την εκτίμηση του d). Θα δοκιμάσεις τον καθορισμό του d με χρήση τυχαιοποιημένων δειγμάτων. Κάθε τυχαιοποιημένο δείγμα των 13 δεικτών παράγεται με τυχαία αντιμετάθεση των παρατηρήσεων της κάθε μεταβλητής. Θα δημιουργήσεις $B = 1000$ τυχαιοποιημένα δείγματα, θα εφαρμόσεις το PCA σε κάθε ένα από αυτά και θα κρατήσεις τις 13 ιδιοτιμές για το καθένα. Με αυτόν τον τρόπο σχηματίζουμε την εμπειρική κατανομή της κάθε ιδιοτιμής $\lambda_1, \ldots, \lambda_{13}$, αν οι 13 μεταβλητές ήταν ανεξάρτητες μεταξύ τους. Ιδιοτιμές στο αρχικό δείγμα που είναι στην δεξιά ουρά αυτής της εμπειρικής κατανομής (για ανεξάρτητα δείγματα) μπορούν να θεωρηθούν στατιστικά σημαντικές (θέτοντας κάποιο όριο σημαντικότητας, π.χ. $\alpha = 0.05$). Με αυτόν τον τρόπο θα ελέγξετε αν κάποιες (από τις πρώτες) ιδιοτιμές είναι σημαντικές και το πλήθος τους θα δώσει την εκτίμηση του d με τυχαιοποίηση. Συμφωνεί αυτή η εκτίμηση με την εκτίμηση του κατωφλίου $\bar{\lambda}$;

10. Θέλουμε να βρούμε κατάλληλο μοντέλο πολλαπλής γραμμικής παλινδρόμησης για τη σχετική υγρασία (RH) από τις υπόλοιπες μεταβλητές. Η διερεύνηση του κατάλληλου μοντέλου θα γίνει χωρίζοντας το σύνολο των παρατηρήσεων σε δύο υποσύνολα (ελεύθερη επιλογή), στο σύνολο εκμάθησης (training set) στο οποίο θα εκτιμήσεις τους συντελεστές του μοντέλου και στο σύνολο αξιολόγησης (validation set) στο οποίο θα υπολογίσεις τα σφάλματα και το στατιστικό του συντελεστή προσδιορισμού. Θα δοκιμάσεις το πλήρες μοντέλο παλινδρόμησης με όλες τις 12 μεταβλητές καθώς και δύο μοντέλα παλινδρόμησης που εφαρμόζουν μείωση διάστασης. Για τους δύο τύπους μοντέλων με μείωση διάστασης θα δοκιμάσεις διαφορετικές διαστάσεις $d < 12$. Θα συγκρίνεις το συντελεστή προσδιορισμού στο σύνολο αξιολόγησης με το πλήρες μοντέλο και τα μοντέλα μείωσης διάστασης για διαφορετικό d.