Εργασία στο μάθημα 'Ανάλυση Δεδομένων', Δεκέμβριος 2025

Δημήτρης Κουγιουμτζής

E-mail: dkugiu@auth.gr

23 Δεκεμβρίου 2025

Οδηγίες: Σχετικά με την παράδοση της εργασίας, θα πρέπει να ακολουθήσετε πιστά τα παρακάτω. Μη-τήρηση των οδηγιών μπορεί να επιφέρει ποινή στον βαθμό της εργασίας.

- Δεν θα πρέπει να υπάρχουν συναρτήσεις μέσα στο πρόγραμμα. Η κάθε συνάρτηση θα πρέπει να είναι σε ξεχωριστό αρχείο (για το όνομα της δες παρακάτω).
- 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 No 10 και το ζήτημα 5. Για τα προγράμματα τα ονόματα των αρχείων θα είναι Group10Exe5Prog1.m, Group10Exe5Prog2.m κτλ (αν χρειάζεται να δημιουργήσετε περισσότερα από ένα προγράμματα για το ζήτημα). Αντίστοιχα για τις συναρτήσεις τα ονόματα των αρχείων θα είναι Group10Exe5Fun1.m, Group10Exe5Fun2.m κτλ (αν χρειάζεται να δημιουργήσετε περισσότερες από μια συναρτήσεις). Στην αρχή κάθε προγράμματος και συνάρτησης 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 (τα προγράμματα θα φορτώνουν το αρχείο από τον ίδιο φάκελο). Επεξηγήσεις, σχολιασμοί αποτελεσμάτων και συμπεράσματα, όπου ζητούνται, θα δίνονται με μορφή σχολίων στο πρόγραμμα (τα συμπεράσματα στο τέλος του προγράμματος). Τα σχόλια θα πρέπει να είναι γραμμένα στην Αγγλική γλώσσα ή στην Ελληνική με λατινικούς χαρακτήρες (Greeklish) για να αποφευχθεί τυχόν πρόβλημα στην ανάγνωση τους.
- Θα υποβληθούν μόνο 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.).
- **Μπορεί ο διδάσκων να ζητήσει μια ομάδα να παρουσιάσει και συζητήσει για προγράμματα που έχει υποβάλει. Αυτό θα γίνει την επόμενη της τελευταίας ημέρας υποβολής, που είναι και η εξέταση στο μάθημα, μετά τη λήξη της εξέτασης. Το πρωί της ίδιας μέρας θα σταλεί email στα μέλη της ομάδας με τον σύνδεσμο zoom και την ακριβή ώρα σύνδεσης. Αν κάποιο μέλος της ομάδας δεν είναι διαθέσιμο (παρόν) θα μετρήσει αρνητικά στη βαθμολογία της εργασίας (ως και μηδενισμό).**

Περιγραφή και ζητήματα εργασίας

Η εργασία βασίζεται σε μια παλιά μελέτη που έγινε για την πρόβλεψη της απόδοσης του επεξεργαστή ηλεκτρονικού υπολογιστή (H/Y) [Kibler, D. & Aha,D. (1988). Instance-Based Prediction of Real-Valued Attributes. In Proceedings of the CSCSI (Canadian AI) Conference]. Οι συγγραφείς έκαναν προβλέψεις της απόδοσης του επεξεργαστή H/Y (CPU performance) διαφόρων κατασκευαστών από κάποιους σχετικούς δείκτες χαρακτηριστικών του επεξεργαστή με γραμμικό μοντέλο παλινδρόμησης. Τα δεδομένα δίνονται στο αρχείο `CPUperformance.xls` στην ιστοσελίδα του μαθήματος. Το αρχείο έχει 209 γραμμές με εγγραφές που αποτελούν το συνολικό δείγμα. Οι στήλες δίνονται στον παρακάτω πίνακα, όπου η στήλη 1 δηλώνει με έναν κωδικό από 1 ως 12 τον κατασκευαστή (για τους 12 κατασκευαστές με την πιο συχνή παρουσία στο δείγμα), η στήλη 2 έχει τα ονόματα όλων των κατασκευαστών, οι στήλες 3-8 τους δείκτες χαρακτηριστικών H/Y και η στήλη 9 έχει την απόδοση.

Α/Α	*Όνομα*	*Περιγραφή*
1	code	Κωδικός κατασκευαστή με τιμή από 1 ως 12
2	VentorName	όνομα του κατασκευαστή
3	MYCT	machine cycle time in nanoseconds
4	MMIN	minimum main memory in kilobytes
5	MMAX	maximum main memory in kilobytes
6	CACH	cache memory in kilobytes
7	CHMIN	minimum channels in units
8	CHMAX	maximum channels in units
9	PRP	PRP: published relative performance

1. Φτιάξε μια συνάρτηση που να κάνει τα παρακάτω:

   (α') να επιλέγει τυχαία n παρατηρήσεις από το σύνολο των N παρατηρήσεων (όπου $N = 209$ και οι παρατηρήσεις είναι από κάποιον από τους δείκτες) και

   (β') να υπολογίζει την εμπειρική συνάρτηση πυκνότητας πιθανότητας (σππ) με τη μέθοδο του ιστογράμματος για κατάλληλη ισομερή διαμέριση.

   Φτιάξε ένα πρόγραμμα που για κάποιον από τους δείκτες να καλεί αυτή τη συνάρτηση $M = 50$ φορές για $n = 100$ και στη συνέχεια να σχηματίζει M καμπύλες για τα M ιστογράμματα σε ένα σχήμα. Επιπλέον στο ίδιο σχήμα να σχηματίζει επίσης και την εμπειρική σππ από όλα τα $N = 209$ δεδομένα για το δείκτη (με άλλο χρώμα ή μορφή καμπύλης για να ξεχωρίζει).

   Εφάρμοσε αυτή τη διαδικασία για τους δείκτες MYCT, CACH και CHMIN. Σχολίασε κατά πόσο οι σππ από τα M δείγματα των 100 παρατηρήσεων συμφωνούν με την σππ από το σύνολο των παρατηρήσεων. Φαίνεται η σππ για τον κάθε ένα από τους τρεις δείκτες να προσεγγίζει κάποια γνωστή κατανομή;

2. Φτιάξε μια συνάρτηση που να κάνει τα παρακάτω:

   (α') να επιλέγει τυχαία n παρατηρήσεις από το σύνολο των N παρατηρήσεων (όπου $N = 209$ και οι παρατηρήσεις είναι από κάποιον από τους δείκτες) και

(β') να κάνει έλεγχο X^2 καλής προσαρμογής σε κανονική κατανομή. Η συνάρτηση θα πρέπει να δίνει στην έξοδο την p-τιμή του ελέγχου.

Φτιάξε ένα πρόγραμμα που για κάποιον από τους δείκτες να καλεί αυτή τη συνάρτηση $M = 100$ φορές για $n = 40$ και να μετρά το ποσοστό που η υπόθεση της κανονικής κατανομής μπορεί να γίνει αποδεκτή σε επίπεδο σημαντικότητας $\alpha = 0.05$. Επανέλαβε το ίδιο αλλά αφού πρώτα λογαριθμήσεις τις τιμές (νεπέριος λογάριθμος), δηλαδή εφαρμόζοντας το μετασχηματισμό $y = \log(x)$.

Εφάρμοσε αυτή τη διαδικασία για τους δείκτες MYCT, MMAX και CHMIN. Σχολίασε αν η απόφαση για κανονική κατανομή του κάθε δείκτη αλλάζει με τη λογαρίθμηση των τιμών του δείγματος.

3. Επανέλαβε το ίδιο πρόγραμμα (και συνάρτηση) αλλά για να ελέγξεις αν η κατανομή του δείγματος (με ή χωρίς τη λογαρίθμηση) είναι αυτή του μεγάλου δείγματος των $N = 209$ παρατηρήσεων.

   Εφάρμοσε αυτή τη διαδικασία για τους δείκτες MYCT, MMAX και CHMIN. Σχολίασε αν η κατανομή του δείκτη με βάση το δείγμα μπορεί να διαφέρει από αυτήν του μεγάλου δείγματος των $N = 209$ παρατηρήσεων.

4. Για κάποιο δείκτη, επέλεξε τυχαία $M = 100$ δείγματα των $n = 20$ παρατηρήσεων το καθένα, και υπολόγισε με βάση το κάθε δείγμα το 95% διάστημα εμπιστοσύνης για τη μέση τιμή, χρησιμοποιώντας το κατάλληλο παραμετρικό διάστημα εμπιστοσύνης και το διάστημα εμπιστοσύνης με τη μέθοδο bootstrap (ελεύθερη επιλογή του τύπου bootstrap). Υπολόγισε επίσης τη μέση τιμή του δείκτη στο σύνολο των δεδομένων. Για κάθε μια από τις δύο προσεγγίσεις εκτίμησης διαστήματος εμπιστοσύνης, θα πρέπει να παρουσιάσεις το ποσοστό των M διαστημάτων εμπιστοσύνης της μέσης τιμής του δείκτη που περιλαμβάνουν τη μέση τιμή του δείκτη υπολογισμένη στο σύνολο των δεδομένων. Είναι το ποσοστό αυτό αναμενόμενο και για τις δύο προσεγγίσεις; Επανέλαβε την ίδια διαδικασία αλλά λογαριθμίζοντας πρώτα τις τιμές ($y = \log(x)$) και απάντησε στο ίδιο ερώτημα. Εφάρμοσε αυτή τη διαδικασία στους δείκτες MYCT, MMAX και CHMIN και σχολίασε αντίστοιχα.

5. Επέλεξε τυχαία δύο κατασκευαστές, δηλαδή δύο κωδικούς στη στήλη 1 με τιμές από 1 ως 12. Οι τιμές στα αντίστοιχα κελιά της στήλης 9 αποτελούν τα δύο δείγματα απόδοσης του H/Y PRP. Με βάση τα δύο αυτά δείγματα θέλουμε να εκτιμήσουμε αν η μέση απόδοση PRP διαφέρει στους δύο κατασκευαστές. Για αυτό παρουσίασε πρώτα τα θηκογράμματα για τα δύο δείγματα (σε ένα σχήμα) και για κάθε δείγμα κάνε έλεγχο X^2 καλής προσαρμογής σε κανονική κατανομή. Αν με βάση τους δύο ελέγχους η απόδοση PRP ακολουθεί κανονική κατανομή και για τους δύο κατασκευαστές, τότε κάνε παραμετρικό (Student) 95% διάστημα εμπιστοσύνης για τη διαφορά μέσων τιμών. Αν έστω σε έναν από τους δύο ελέγχους υπάρχει απόρριψη της μηδενικής υπόθεσης κανονικής κατανομής υπολόγισε το 95% διάστημα εμπιστοσύνης χρησιμοποιώντας την μέθοδο bootstrap (ελεύθερη επιλογή τύπου bootstrap). Επανέλαβε την ίδια διαδικασία 10 φορές για τυχαία και διαφορετικά ζεύγη κατασκευαστών. Στο τέλος το πρόγραμμα θα πρέπει να δίνει πόσες φορές απορρίφθηκε η μηδενική υπόθεση κανονικής κατανομής και πόσες

φορές βρέθηκε να υπάρχει στατιστικά σημαντική διαφορά στην μέση απόδοση στους δύο κατασκευαστές. Φαίνεται να διαφέρει γενικά με τον κατασκευαστή η απόδοση του H/Y PRP;

6. Με βάση τα M δείγματα των $n = 20$ στο Ερώτημα 4 για τους δείκτες MMAX και CHMIN υπολόγισε M παραμετρικά 95% διαστήματα εμπιστοσύνης για το συντελεστή συσχέτισης μεταξύ αυτών των δεικτών κάνοντας χρήση του μετασχηματισμού Fisher. Υπολόγισε επίσης τον ίδιο συντελεστή συσχέτισης στο σύνολο των δεδομένων. Σε τι ποσοστό τα M διαστήματα εμπιστοσύνης του συντελεστή συσχέτισης περιλαμβάνουν το συντελεστή συσχέτισης υπολογισμένο στο σύνολο των δεδομένων; Είναι το ποσοστό αυτό αναμενόμενο; Επανέλαβε την ίδια διαδικασία αλλά λογαριθμίζοντας πρώτα τις τιμές ($y = \log(x)$) και απάντησε στο ίδιο ερώτημα.

7. Κάνε παραμετρικό έλεγχο σε επίπεδο σημαντικότητας $\alpha = 0.05$ της μηδενικής υπόθεσης για μηδενική συσχέτιση των δεικτών MMAX και CHMIN, χρησιμοποιώντας το στατιστικό της κατανομής Student σε κάθε ένα από τα M δείγματα των $n = 20$ στο Ερώτημα 6. Κάνε επίσης έλεγχο τυχαιοποίησης για την ίδια μηδενική υπόθεση για τα ίδια M δείγματα. Σε τι ποσοστό απορρίπτεται η μηδενική υπόθεση στους M ελέγχους σημαντικότητας του συντελεστή συσχέτισης για τον παραμετρικό έλεγχο και για τον έλεγχο τυχαιοποίησης; Επανέλαβε την ίδια διαδικασία αλλά λογαριθμίζοντας πρώτα τις τιμές ($y = \log(x)$) και απάντησε στο ίδιο ερώτημα. Υπάρχει διαφορά στους δύο ελέγχους (παραμετρικός και τυχαιοποίησης) πριν και μετά τον μετασχηματισμό λογαρίθμου;

8. Επίλεξε τυχαία 50 από τις $N = 209$ καταγραφές. Σε αυτό το δείγμα, υπολόγισε το κατάλληλο μοντέλο παλινδρόμησης (γραμμικό, πολυωνυμικό, άλλο) που να αποδίδει καλύτερα την εξάρτηση της απόδοσης του H/Y PRP (στήλη 9) από το δείκτη MYCT. Για την επιλογή του κατάλληλου μοντέλου παλινδρόμησης κάνε διαγνωστικό έλεγχο με το διάγραμμα διασποράς των τυποποιημένων υπολοίπων για κάθε μοντέλο που δοκιμάζεις και υπολόγισε το συντελεστή προσδιορισμού καθώς και τον προσαρμοσμένο συντελεστή προσδιορισμού. Εφάρμοσε τη διαδικασία προσαρμογής μοντέλου και στο σύνολο των $N = 209$ δεδομένων. Διαφέρουν τα δύο μοντέλα που κατέληξες; Φαίνεται κάποιο από τα δύο μοντέλα να είναι πιο ακριβές;

9. Για κάθε ένα από τα δύο σύνολα δεδομένων στο Ερώτημα 8 (των 50 και των 209 παρατηρήσεων), διερεύνησε το κατάλληλο μοντέλο πολλαπλής γραμμικής παλινδρόμησης για την απόδοση του H/Y PRP με βάση τους 6 δείκτες στις στήλες 3-8. Δοκίμασε το μοντέλο με όλες τις 6 ανεξάρτητες μεταβλητές και σύγκρινε το με το μοντέλο που δίνει η μέθοδος βηματικής παλινδρόμησης. Υπολόγισε για το κάθε μοντέλο τη διασπορά των σφαλμάτων και τον συντελεστή προσδιορισμού καθώς και τον προσαρμοσμένο συντελεστή προσδιορισμού. Είναι ίδιες οι επιλεγμένες ανεξάρτητες μεταβλητές στα δύο μοντέλα από τη βηματική παλινδρόμηση με βάση τα δύο δείγματα των 50 και 209 παρατηρήσεων;

10. Το πρόγραμμα θα ξεκινά με το να βρίσκει τα δείγματα των κατασκευαστών που έχουν μέγεθος πάνω από 10, καθώς αυτά θα χρησιμοποιηθούν στη συνέχεια. Για κάθε τέτοιο πολυμεταβλητό δείγμα κατασκευαστή, θα συγκρίνεις τρία μοντέλα πολλαπλής γραμμικής παλινδρόμησης για την απόδοση του H/Y PRP με βάση τους 6 δείκτες στις στήλες 3-8: 1)

το πλήρες γραμμικό μοντέλο με όλους τους 6 δείκτες, 2) το μοντέλο μείωσης διάστασης PCR και 3) το μοντέλο μείωσης διάστασης LASSO. Για τα δύο τελευταία μοντέλα θα χρησιμοποιήσεις κάποο κριτήριο μείωσης διάστασης. Η σύγκριση θα γίνει υπολογίζοντας το μέσο τετραγωνικό σφάλμα στο 35% των παρατηρήσεων που δεν θα έχουν χρησιμοποιηθεί για την εκμάθηση του μοντέλου (ο υπολογισμός του κάθε μοντέλου θα γίνει στο υπόλοιπο 65% των παρατηρήσεων).