Εργασία στο μάθημα 'Ανάλυση Δεδομένων', Δεκέμβριος 2023

Δημήτρης Κουγιουμτζής

E-mail: dkugiu@auth.gr

22 Δεκεμβρίου 2023

Οδηγίες: Σχετικά με την παράδοση της εργασίας:

- 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 (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 υποβληθεί μόνο ένα συμπιεσμένο αρχείο που θα πρέπει να περιέχει 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.).
- **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 στις 12:00 - 15:00 απομακρυσμένα. Το πρωί της ίδιας μέρας θα σταλεί email στα μέλη της ομάδας με τον σύνδεσμο zoom και την ακριβή ώρα σύνδεσης και συζήτησης. Αν κάποιο μέλος της ομάδας δεν είναι διαθέσιμο (παρόν) θα μετρήσει αρνητικά στη βαθμολογία της εργασίας (ως και μηδενισμό).**

Περιγραφή εργασίας

Στο αρχείο SeoulBike.xlsx που υπάρχει στην ιστοσελίδα του μαθήματος, δίνονται ωριαία δεδομένα για πλήθος νοικιασμένων δημόσιων ποδηλάτων στο Seoul Bike Sharing System

καθώς και αντίστοιχα μετεωρολογικά δεδομένα. Περισσότερες πληροφορίες υπάρχουν στην ιστοσελίδα https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand, από όπου λήφθηκαν τα δεδομένα (ώρες που το σύστημα δεν ήταν σε λειτουργία έχουν αφαιρεθεί). Συγκεκριμένα οι δείκτες που αντιστοιχούν στις στήλες του πίνακα δεδομένων στο αρχείο είναι

Α/Α	*Συντόμευση*	*Τίτλος*
1	Date	Date
2	Bikes	Rented Bike Count
3	Hour	Hour
4	Temperature	Temperature(^{o}C)
5	Humidity	Humidity(%)
6	WindSpeed	Wind speed (m/s)
7	Visibility	Visibility (10m)
8	DewPoint	Dew point temperature(^{o}C)
9	Solar	Solar Radiation (MJ/m2)
10	Rainfall	Rainfall(mm)
11	Snowfall	Snowfall (cm)
12	Season	Seasons
13	Holiday	Holiday

Για ευκολία στη χρήση των δύο τελευταίων κατηγορηματικών δεικτών, έχουν αντιστοιχηθεί στις κατηγορίες ακέραιες τιμές ως εξής:

Season	Winter->1, Spring->2, Summer->3, Autumn->4
Holiday	No Holiday->0, Holiday->1

Ζητήματα εργασίας

1. Θεωρείστε τα νοικιασμένα ποδήλατα ανά ώρα (Bikes) σε μια εποχή του έτους (για μια τιμή του Season). Βρείτε την κατάλληλη γνωστή (παραμετρική) κατανομή πιθανότητας που προσαρμόζεται καλύτερα σε αυτά τα δεδομένα. Επανέλαβε το ίδιο για όλες τις 4 εποχές. Διαφέρει η κατάλληλη κατανομή στις 4 εποχές; [Βοήθεια: Για λίστα κατανομών δες συνάρτηση fitdist. Για τη σύγκριση κατανομών, η καλή προσαρμογή μπορεί να αξιολογηθεί από την τιμή του αντίστοιχου στατιστικού ελέγχου X^2].

2. Θεωρείστε νοικιασμένα ποδήλατα ανά ώρα (Bikes) σε δύο εποχές του έτους (για δύο διαφορετικές τιμές του Season). Πάρετε ένα τυχαίο δείγμα 100 παρατηρήσεων από την κάθε εποχή. Με βάση αυτά τα δύο δείγματα, ελέγξετε αν οι κατανομές πιθανότητας των ποδηλάτων στις δύο εποχές μπορούν να θεωρηθούν πως είναι ίδιες. Για αυτό θα κάνετε παρόμοιο έλεγχο με τον έλεγχο καλής προσαρμογής X^2, αλλά οι παρατηρούμενες τιμές θα προέρχονται από το ιστόγραμμα της πρώτης εποχής και οι αναμενόμενες τιμές από το ιστόγραμμα της δεύτερης εποχής. Επαναλάβετε αυτό M φορές, π.χ. $M = 100$, και υπολογίστε το ποσοστό που η κατανομή των ποδηλάτων δε διαφέρει στις δύο εποχές. Κάντε την παραπάνω διαδικασία για όλα τα ζεύγη εποχών (6 ζεύγη). Σχολιάστε με βάση το ποσοστό για τη διαφορά κατανομών για κάθε ζεύγος εποχών αν φαίνεται να υπάρχουν διαφορετικές κατανομές για κάποιες εποχές.

3. Θεωρείστε νοικιασμένα ποδήλατα ανά ώρα (Bikes) σε μια εποχή του έτους (για μια τιμή του Season) και για δύο διαφορετικές ώρες της ημέρας (για δύο διαφορετικές τιμές του Hour). Ελέγξετε αν διαφέρει το μέσο πλήθος νοικιασμένων ποδηλάτων στις δύο ώρες της ημέρας και κατά πόσο. Για αυτό θα θεωρήσετε νέο δείγμα από τις διαφορές των αντίστοιχων τιμών ποδηλάτων στις δύο ώρες για την ίδια μέρα (για όλες τις ημέρες της εποχής). Με βάση αυτό το δείγμα θα κάνετε έλεγχο για μέση τιμή μηδέν. Θα κάνετε τον έλεγχο για όλα τα ζευγάρια ωρών (276 ζεύγη). Παρουσιάστε τα αποτελέσματα (μέση διαφορά ποδηλάτων και απόρριψη ή μη του ελέγχου για μέση διαφορά μηδέν) για όλα τα ζεύγη ωρών, π.χ. χρωματικός πίνακας (colormap) μεγέθους 24×24 για μέση διαφορά και αντίστοιχα για τον έλεγχο. Κάντε την παραπάνω διαδικασία για κάθε εποχή και σχολιάστε αν οι πίνακες αποτελεσμάτων μοιάζουν στις 4 εποχές ή υπάρχουν σημαντικές διαφορές για κάποια εποχή.

4. Θεωρείστε νοικιασμένα ποδήλατα ανά ώρα (Bikes) σε δύο εποχές του έτους (για δύο διαφορετικές τιμές του Season) και για συγκεκριμένη ώρα της ημέρας (για μια τιμή του Hour). Θέλουμε να διερευνήσουμε αν διαφέρει το πλήθος νοικιασμένων ποδηλάτων στις δύο εποχές του έτους και για την ίδια ώρα της ημέρας. Για αυτό θα χρησιμοποιήσετε διαστήματα εμπιστοσύνης bootstrap για διαφορά διαμέσων (αντίστοιχα με αυτά για τη διαφορά μέσων τιμών). Θα επαναλάβετε την ίδια διαδικασία για κάθε μια από τις 24 ώρες της ημέρας. Θα παρουσιάσετε τα διαστήματα εμπιστοσύνης ως προς την ώρα της ημέρας σε ένα σχήμα όπου θα σημειώνεται και αν υπάρχει στατιστικά σημαντική διαφορά. Θα επαναλάβετε την ίδια διαδικασία για όλα τα ζεύγη των εποχών (6 ζεύγη). Υπάρχουν διαφορές στο μέσο πλήθος νοικιασμένων ποδηλάτων μεταξύ εποχών και για ποια ώρα και ποιες εποχές;

5. Θεωρείστε νοικιασμένα ποδήλατα ανά ώρα (Bikes) και θερμοκρασία (Temperature) για συγκεκριμένη εποχή του έτους (τιμή του Season) και ώρα της ημέρας (τιμή του Hour). Θέλουμε να διερευνήσουμε αν υπάρχει γραμμική συσχέτιση της ενοικίασης ποδηλάτων με τη θερμοκρασία και για ποιες ώρες της ημέρας και εποχές εμφανίζονται οι υψηλότερες συσχετίσεις. Για αυτό για κάθε συνδυασμό ώρας της ημέρας και εποχής θα υπολογίσετε τον συντελεστή συσχέτισης Pearson και θα κάνετε έλεγχο σημαντικότητας του. Θα παρουσιάσετε τους συντελεστές συσχέτισης ως προς την ώρα της ημέρας σε ένα σχήμα και για κάθε μια από τις 4 εποχές (στο ίδιο σχήμα ή σε 4 σχήματα), όπου θα πρέπει να ξεχωρίζουν οι στατιστικά σημαντικές συσχετίσεις από τις στατιστικά μη-σημαντικές. Εντοπίστε τις ισχυρότερες συσχετίσεις και σχολιάστε αν φαίνεται να είναι την/τις ίδια/ες ώρα/ες σε κάθε εποχή.

6. Η αμοιβαία πληροφορία $I(X, Y)$ δύο τ.μ. X και Y χρησιμοποιείται και ως μέτρο γραμμικής και μη-γραμμικής συσχέτισης και ορίζεται ως $I(X, Y) = H(X) + H(Y) - H(X, Y)$, όπου $H(X) = -\sum_x f_X(x) \log f_X(x)$ είναι η εντροπία (του Shannon) της X και $f_X(x)$ η συνάρτηση μάζας πιθανότητας της διακριτοποιημένης X (αν η X είναι συνεχής). Όμοια ορίζεται η κοινή εντροπία των X και Y, $H(X, Y) = -\sum_{x,y} f_{X,Y}(x, y) \log f_{X,Y}(x, y)$, όπου $f_{X,Y}(x, y)$ είναι η από κοινού συνάρτηση μάζας πιθανότητας των διακριτοποιημένων X και Y (αν η X και Y είναι συνεχείς). Για την υλοποίηση της εκτίμησης αμοιβαίας πληροφορίας $I(X, Y)$ για συνεχείς μεταβλητές X και Y θα θεωρήσετε τη διακριτοποίηση των X και Y σε k τιμές, που αντιστοιχούν στα ισομήκη διαστήματα της διαμέρισης του πεδίου

τιμών τους (όπως και για τα ιστογράμματα). Το k είναι ίδιο και για τις δύο μεταβλητές. Στην ιστοσελίδα του μαθήματος δίνεται η συνάρτηση `MutualInformationXY.m` του Matlab που υπολογίζει αυτήν την εκτίμηση της αμοιβαίας πληροφορίας. Επίσης είναι γνωστό πως όταν οι δύο μεταβλητές X και Y ακολουθούν διμεταβλητή κανονική (Γκαουσιανή) κατανομή ισχύει $I(X,Y) = -0.5\log(1-\rho^2)$, όπου ρ είναι ο συντελεστής συσχέτισης Pearson. Ας την ονομάσουμε Γκαουσιανή αμοιβαία πληροφορία. Αυτή είναι η αμοιβαία πληροφορία που αντιστοιχεί μόνο σε γραμμική συσχέτιση (θεωρώντας πως οι X και Y είναι Γκαουσιανές). Θεωρείστε επίσης την κανονικοποιημένη αμοιβαία πληροφορία που λέγεται και συντελεστής μέγιστης πληροφορίας (Maximal information coefficient, MIC) διαιρώντας την αμοιβαία πληροφορία με το $\log(k)$. Αντίστοιχα θεωρείστε και την κανονικοποιημένη Γκαουσιανή αμοιβαία πληροφορία διαιρώντας επίσης με το $\log(k)$ (Gaussianized Maximal information coefficient, GMIC).

Θεωρείστε νοικιασμένα ποδήλατα ανά ώρα (Bikes) και θερμοκρασία (Temperature) για συγκεκριμένη εποχή του έτους (τιμή του Season) και ώρα της ημέρας (τιμή του Hour). Θέλουμε να διερευνήσουμε αν υπάρχει γραμμική και μη-γραμμική συσχέτιση της ενοικίασης ποδηλάτων με τη θερμοκρασία για μια εποχή του έτους. Για κάθε ώρα της ημέρας θα υπολογίσετε αντίστοιχα το GMIC και το MIC και θα κάνετε έλεγχο σημαντικότητας και για τα δύο με τυχαιοποίηση (δες άσκηση 5.2). Θα παρουσιάσετε ένα διάγραμμα διασποράς ενοικίασης ποδηλάτων με θερμοκρασία για κάθε ώρα (24 σχήματα). Στο ίδιο σχήμα θα δίνονται οι τιμές των δύο συντελεστών συσχέτισης και θα σημειώνεται αν είναι ή όχι στατιστικά σημαντικοί. Σχολιάστε αν φαίνεται να διαφέρουν οι δύο συντελεστές συσχέτισης για κάποια/ες ώρα/ες της ημέρας;

7. Θεωρείστε νοικιασμένα ποδήλατα ανά ώρα (Bikes) και θερμοκρασία Temperature για μια εποχή του έτους (κάποια τιμή του Season) και μια ώρα της ημέρας (κάποια τιμή του Hour). Βρείτε το καλύτερο μοντέλο παλινδρόμησης των ποδηλάτων ως προς τη θερμοκρασία για τη συγκεκριμένη ώρα της ημέρας και εποχή του έτους. Θα δοκιμαστούν το γραμμικό μοντέλο και κάποια μη-γραμμικά μοντέλα, όπως εγγενής συνάρτησης ή πολυωνύμου (μπορείτε να επιλέξετε και οποιαδήποτε άλλη μη-γραμμική συνάρτηση για την παλινδρόμηση). Η αξιολόγηση των μοντέλων θα πρέπει να γίνει με τον προσαρμοσμένο συντελεστή προσδιορισμού. Θα επαναλάβετε αυτήν τη διαδικασία για κάθε μια από τις 24 ώρες της εποχής. Θα παρουσιάσετε, π.χ. σε μια λίστα ή πίνακα, το πιο κατάλληλο μοντέλο (το όνομα του ή την μαθηματική έκφραση του) και τον προσαρμοσμένο συντελεστή προσδιορισμού για κάθε ώρα της ημέρας. Σχολιάστε αν φαίνεται να υπάρχει μη-γραμμική εξάρτηση την ενοικίασης ποδηλάτων από τη θερμοκρασία και για ποιες ώρες της ημέρας.

8. Θεωρείστε νοικιασμένα ποδήλατα ανά ώρα (Bikes) και θερμοκρασία Temperature για μια ώρα της ημέρας (κάποια τιμή του Hour) και για δύο εποχές του έτους (δύο τιμές του Season). Θέλουμε να διερευνήσουμε αν το γραμμικό μοντέλο της ενοικίασης ποδηλάτων από τη θερμοκρασία για μια συγκεκριμένη ώρα της ημέρας προσαρμόζεται το ίδιο καλά στις δύο εποχές. Για αυτό προσαρμόστε γραμμικό μοντέλο παλινδρόμησης των ποδηλάτων ως προς τη θερμοκρασία για τη συγκεκριμένη ώρα της ημέρας και ξεχωριστά σε κάθε μια από τις δύο εποχές, και υπολογίστε κάποιο στατιστικό καλής προσαρμογής, όπως τον συντελεστή προσδιορισμού (ελεύθερη επιλογή). Για να ελέγξετε αν ο συντελεστής προσ-

διορισμού (ή όποιο άλλο στατιστικό επιλέξατε) διαφέρει στις δύο εποχές θα κάνετε έλεγχο ισότητας των δύο συντελεστών προσδιορισμού. Ο έλεγχος θα γίνει είτε με bootstrap ή με τυχαιοποίηση, με παρόμοιο τρόπο με τον έλεγχο ισότητας δύο μέσω τιμών (δεν στην αντίστοιχη ενότητα των σημειώσεων). Το κάθε bootstrap δείγμα σχηματίζεται επιλέγοντας τυχαία και με επανάθεση ζευγαρωτές παρατηρήσεις από το κοινό δείγμα ζευγαρωτών παρατηρήσεων (θερμοκρασία και ποδήλατα) των δύο εποχών. Αντίστοιχη είναι η διαδικασία σχηματισμού τυχαιοποιημένων δειγμάτων αλλά χωρίς επανάθεση. Εφαρμόστε αυτήν την διαδικασία για μια συγκεκριμένη ώρα της ημέρας και για όλους τους συνδυασμούς δύο εποχών (6 ζεύγη) και παρουσιάστε τα αποτελέσματα των αντίστοιχων ελέγχων. Σχολιάστε αν φαίνεται η προσαρμογή του γραμμικού μοντέλου να είναι το ίδιο καλή για όλες τις εποχές ή κάποιες εποχές.

9. Μας ενδιαφέρει να δούμε αν μπορούμε να προσδιορίσουμε το πλήθος νοικιασμένων ποδηλάτων ανά ώρα (Bikes) με ένα κατάλληλο γραμμικό μοντέλο που να περιλαμβάνει τους πιο σχετικούς από τους 10 δείκτες, όπου δεν περιλαμβάνονται οι δείκτες εποχής (Season) και διακοπών (Holiday). Θα περιορίσετε τη μελέτη στα δεδομένα που δεν αντιστοιχούν σε διακοπές (θα αφαιρεθούν τα δεδομένα για Holiday=1) και για μια εποχή (για μια τιμή του Season). Ειδικότερα θέλουμε να συγκρίνουμε την πρόβλεψη του πλήθους νοικιασμένων ποδηλάτων από τους σχετικούς δείκτες σε μια ώρα όταν εκπαιδεύουμε το μοντέλο μόνο στα δεδομένα αυτής της ώρας της ημέρας (μοντέλο 1) και όταν το εκπαιδεύουμε στα δεδομένα όλων των ωρών της ημέρας (24 φορές περισσότερα δεδομένα, μοντέλο 2), για την ίδια πάντα εποχή. Θα υπολογίσετε κάποιο κατάλληλο στατιστικό πρόβλεψης τιμών ποδηλάτων (π.χ. συντελεστής προσδιορισμού, μέσο τετραγωνικό σφάλμα) για προβλέψεις στις ώρες των τελευταίων 20 ημερών της εποχής (20*24=480 τιμές στο σύνολο αξιολόγησης) που δεν θα έχετε χρησιμοποιήσει για να εκτιμήσετε το μοντέλο (οι προηγούμενες μέρες από αυτές τις 20 ημέρες της εποχής αποτελούν το σύνολο εκμάθησης). Ως μοντέλο 1 θα έχετε ουσιαστικά 24 μοντέλα, ένα για κάθε ώρα της ημέρας που θα το εκτιμήσετε στην αντίστοιχη ώρα για κάθε μέρα στο σύνολο εκμάθησης, π.χ. για το χειμώνα υπάρχουν δεδομένα για 90 ημέρες και το μοντέλο για κάθε ώρα της ημέρας θα εκτιμηθεί από 70 παρατηρήσεις κάθε δείκτη για αυτήν την ώρα (οι άλλες 20 παρατηρήσεις ανήκουν στο σύνολο αξιολόγησης). Αφού κάνετε τις προβλέψεις και με τα 24 μοντέλα (480 τιμές για όλες τις ώρες) θα τις μαζέψετε για να υπολογίσετε το στατιστικό πρόβλεψης. Για το μοντέλο 1 και για το μοντέλο 2 θα θεωρήσετε α) το πλήρες μοντέλο πολλαπλής παλινδρόμησης, και β) το μοντέλο που προκύπτει από τη μέθοδο της βηματικής παλινδρόμησης. Θα παρουσιάσετε ένα σχήμα που θα δείχνει τις πραγματικές τιμές στο σύνολο αξιολόγησης και τις προβλέψεις με τα 4 μοντέλα (1α, 2α, 1β, 2β) και 4 διαγνωστικά διαγράμματα (τυποποιημένα σφάλματα προς πραγματικές τιμές), ένα για κάθε μοντέλο. Συγκρίνετε τα μοντέλα και σχολιάστε αν η πρόβλεψη βελτιώνεται με κατάλληλη επιλογή δεικτών (βηματική παλινδρόμηση) και με την ξεχωριστή πρόβλεψη ανά ώρα (μοντέλο 1).

10. Μας ενδιαφέρει να δούμε αν μπορούμε να προβλέψουμε με ένα κατάλληλο γραμμικό μοντέλο το πλήθος νοικιασμένων ποδηλάτων ανά ώρα (Bikes) από κάποιους σχετικούς δείκτες από τους 10 δείκτες σε προηγούμενες ώρες, όπου δεν περιλαμβάνονται οι δείκτες εποχής (Season) και διακοπών (Holiday). Θα περιορίσετε τη μελέτη στα δεδομένα που δεν αντιστοιχούν σε διακοπές (θα αφαιρεθούν τα δεδομένα για Holiday=1) και για μια

εποχή (για μια τιμή του Season). Ειδικότερα θέλουμε να συγκρίνουμε την πρόβλεψη του πλήθους νοικιασμένων ποδηλάτων για κάποια ώρα της ημέρας t από τους σχετικούς δείκτες τις προηγούμενες ώρες, $t-1, t-2, \ldots, t-p$, για κάποια μέγιστη υστέρηση p. Αυτό θα γίνει προσαρμόζοντας στο σύνολο των δεδομένων για την εποχή (π.χ. 90 μέρες και παρατηρήσεις για τον χειμώνα) ένα μοντέλο ξεχωριστά για κάθε ώρα της ημέρας, δηλαδή συνολικά θα εκτιμήσετε 24 μοντέλα. Θα χρησιμοποιήσετε δύο τύπους γραμμικών μοντέλων που έχουν μείωση διάστασης (ελεύθερη επιλογή). Για κάθε μια από τις 24 ώρες θα υπολογίσετε τον προσαρμοσμένο συντελεστή προσδιορισμού για κάθε ένα από τα δύο γραμμικά μοντέλα μείωσης διάστασης. Θα παρουσιάσετε τα αποτελέσματα αυτά για κάθε ώρα της ημέρας σε έναν πίνακα ή ένα σχήμα. Θα επαναλάβετε αυτήν τη διαδικασία για διαφορετικά p. Μπορεί να προβλεφθεί το πλήθος νοικιασμένων ποδηλάτων για κάποια ώρα της ημέρας από τους μετεωρολογικούς δείκτες τις προηγούμενες ώρες; Υπάρχουν κάποιες ώρες της ημέρας για τις οποίες η πρόβλεψη νοικιασμένων ποδηλάτων είναι καλύτερη και με ποιο μοντέλο μείωσης διάστασης; Υπάρχει κάποια μέγιστη υστέρηση p που δίνει καλύτερη πρόβλεψη;