

Μεταπτυχιακό Υπολογιστικής Φυσικής, Ιανουάριος 2024

Εργασία στο μάθημα ‘Ανάλυση Δεδομένων’

Δημήτρης Κουγιουμτζής

E-mail: dkugiu@auth.gr

18 Ιανουαρίου 2024

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 το όνομα φοιτητή Koygioumtzhw και το ζήτημα 5. Για τα προγράμματα τα ονόματα των αρχείων θα είναι KoygioumtzhwExe5Prog1.m, KoygioumtzhwExe5Prog2.m κτλ. Για τις συναρτήσεις τα ονόματα των αρχείων θα είναι KoygioumtzhwExe5Fun1.m, KoygioumtzhwExe5Fun2.m κτλ.
- 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). 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 τα αρχεία Matlab (μέσω του elearning και όχι σε συμπιεσμένο αρχείο).
- Η κάθε εργασία (σύνολο προγραμμάτων και συναρτήσεων Matlab) θα πρέπει να συντάσσεται αυτόνομα από τον/την φοιτητή/τρια. Ομοιότητες εργασιών θα οδηγούν σε μοίρασμα της βαθμολογίας (δύο ‘όμοιες’ άριστες εργασίες θα μοιράζονται το βαθμό δια δύο, τρεις δια τρία κτλ.).

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

A/A	Συντόμευση	Τίτλος
1	Date	Date
2	Bikes	Rented Bike Count
3	Hour	Hour
4	Temperature	Temperature(°C)
5	Humidity	Humidity(%)
6	WindSpeed	Wind speed (m/s)
7	Visibility	Visibility (10m)
8	DewPoint	Dew point temperature(°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. Επαναλάβετε το ίδιο για όλες τις 4 εποχές. Διαφέρει η κατάλληλη κατανομή στις 4 εποχές; [Βοήθεια: Για λίστα κατανομών δεξ 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). Πάρτε ένα τυχαίο δείγμα 100 παρατηρήσεων από την κάθε εποχή. Με βάση αυτά τα δύο δείγματα, ελέγξτε 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). Ελέγξτε αν διαφέρει το 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. 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 μεγέθους 24×24 για μέση διαφορά και αντίστοιχα για τον έλεγχο.

4. Υπολογίστε τη διάμεσο του πλήθους νοικιασμένων ποδηλάτων ανά ώρα (Bikes) από όλη την καταγραφή (για όλες τις εποχές, ημέρες και ώρες) στις δύο περιπτώσεις: όταν είναι διακοπές (Holiday->1) και όταν δεν είναι (No Holiday->0). Ας ονομάσουμε αυτές τις δύο διαμέσους ως μ_1 και μ_0 , αντίστοιχα. Θεωρείστε νοικιασμένα ποδήλατα ανά ώρα (Bikes) σε μια εποχή του έτους (για μια τιμή του Season) και για μια ώρα της ημέρας (για μια τιμή του Hour). Μας ενδιαφέρει να εξετάσουμε αν για την συγκεκριμένη ώρα της ημέρας και εποχή και μόνο για τις ημέρες διακοπών η διάμεσος μπορεί να είναι αυτή του συνόλου, δηλαδή το μ_1 . Το ίδιο για τις ημέρες που δεν είναι διακοπές (για το μ_0). Για αυτό θα υπολογίσετε 95% διάστημα εμπιστοσύνης bootstrap για διάμεσο (δες άσκηση 5.5, όπου το δείγμα δεν είναι ζευγαρωτές παρατηρήσεις αλλά παρατηρήσεις από μια τ.μ. και το στατιστικό είναι η διάμεσος). Αυτό θα το κάνετε για τις δύο περιπτώσεις (για μ_1 και μ_0 αντίστοιχα). Θα επαναλάβετε την ίδια διαδικασία για κάθε μια από τις 24 ώρες της ημέρας. Θα παρουσιάσετε δύο σχήματα, ένα για μ_1 και ένα για μ_0 . Στο κάθε σχήμα θα παρουσιάζονται τα διαστήματα εμπιστοσύνης ως προς την ώρα της ημέρας όπου θα σημειώνεται και η τιμή της διαμέσου, μ_1 και μ_0 αντίστοιχα. Διαφέρει η διάμεσος για κάποια ώρα της ημέρας από τη διάμεσο της καταγραφής και για ποιές ώρες και περίπτωση; Υπάρχουν διαφορές στα αποτελέσματα στην περίπτωση διακοπών και μη-διακοπών;
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 οι 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. 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 δε-

δομένα 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 ($20 \cdot 24 = 480$ 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).

9. 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 $\text{Holiday}=1$) 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, $t-1, t-2, \dots, t-p$, 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 το γραμμικό μοντέλο βηματικής παλινδρόμησης. Για κάθε μια από τις 24 ώρες θα υπολογίσετε τον συντελεστή προσδιορισμού του μοντέλου. 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; Υπάρχει κάποια μέγιστη υστέρηση p που δίνει καλύτερη πρόβλεψη;
10. Θέλουμε να εξετάσουμε αν το πλήθος νοικιασμένων ποδηλάτων (Bikes) σε κάποια ώρα συσχετίζεται με το πλήθος ενοικίασης ποδηλάτων σε κάποια προηγούμενη ώρα, δηλαδή αν υπάρχει αυτοσυσχέτιση στη χρονοσειρά του πλήθους ενοικίασης ποδηλάτων. Θα περιορίσετε τη μελέτη στη χρονοσειρά για μια εποχή (για μια τιμή του Season) που αντιστοιχεί σε ένα χρονικό παράθυρο πολλών συνεχόμενων ωρών και ημερών που θα επιλέξετε έτσι ώστε να μην περιέχει διακοπές ($\text{Holiday}=1$), π.χ χρονικό παράθυρο 40 συνεχόμενων η-

μερών που δεν περιέχει ημέρες και ώρες διακοπών και η αντίστοιχη χρονοσειρά πλήθους ποδηλάτων ανά ώρα έχει $40 \cdot 24 = 960$ παρατηρήσεις. Από αυτήν την χρονοσειρά θα απαλείψετε την εποχικότητα και θα υπολογίσετε την συνάρτηση αυτοσυσχέτισης (δες άσκηση 6.1). Θα κάνετε τρία σχήματα: α) το διάγραμμα ιστορίας της αρχικής χρονοσειράς, β) το διάγραμμα ιστορίας της χρονοσειράς αφού έχετε αφαιρέσει την εποχικότητα, γ) το διάγραμμα της συνάρτησης αυτοσυσχέτισης (με με τα όρια της αυτοσυσχέτισης λευκού θορύβου). Θα επαναλάβετε αυτήν τη διαδικασία για κάθε μια από τις 4 εποχές. Φαίνεται να υπάρχουν στατιστικά σημαντικές αυτοσυσχετίσεις στη χρονοσειρά του πλήθους νοικιασμένων ποδηλάτων ανά ώρα (πέρα από αυτήν που δημιουργεί η εποχικότητα) και για ποιες εποχές;