Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών,

Θέμα εργασίας στο μάθημα «Χρονοσειρές» για το ακαδημαϊκό έτος 2022/23

Οδηγίες:

Σχετικά με την παράδοση της εργασίας θα πρέπει:

- *Το κείμενο της αναφοράς της ανάλυσης που ζητείται να είναι γραμμένο σε κάποιο πρόγραμμα επεξεργασίας κειμένου, π.χ. τύπου Word, LaTeX, pdf. Τα γραφήματα και οι πίνακες θα πρέπει να παρουσιάζονται στο σημείο του κειμένου που αναφέρονται.*
- *Μπορείτε να χρησιμοποιήσετε γνωστά προγράμματα (που θα τροποποιήσετε) και συναρτήσεις σε Matlab και Python που δίνονται στην ιστοσελίδα του μαθήματος στο elearning. Τα προγράμματα που θα χρησιμοποιήσετε θα πρέπει να είναι οργανωμένα σε αρχεία και να υποβληθούν μαζί με το αρχείο της αναφοράς, καθώς και το/α αρχείο/α δεδομένων που χρησιμοποιήθηκαν. Η υποβολή θα γίνει μέσω της ιστοσελίδας του μαθήματος στο elearning.*
- *Η κάθε εργασία 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.).*

Γενικά

Η υπολογιστική εργασία είναι στην ανάλυση ιστορικών δεδομένων (χρονοσειρών) από το θερμοπίδακα (geyser) του Old Faithful στις ΗΠΑ (βλέπε http://www.geyserstudy.org/geyser.aspx?pGeyserNo=OLDFAITHFUL). Τα δεδομένα αφορούν το χρόνο μεταξύ διαδοχικών εξάρσεων (eruption) του θερμοπίδακα, δηλαδή το χρόνο σε λεπτά της ώρας από την έναρξη της μιας έξαρσης ως την έναρξη της επόμενης έξαρσης. Η διάρκεια των εξάρσεων είναι πολύ μικρή σε σχέση με το χρόνο μεταξύ των εξάρσεων. Μας ενδιαφέρει να διερευνήσουμε αν υπάρχει κάποια δομή στους διαδοχικούς χρόνους αναμονής (χρόνους μεταξύ διαδοχικών εξάρσεων), που ενδεχομένως να μας επιτρέψει να προβλέψουμε το χρόνο αναμονής ως την επόμενη έξαρξη. Μας ενδιαφέρει επίσης αν η υποτιθέμενη δομή αλλάζει στις διαφορετικές περιόδους που μελετάμε το πρόβλημα, δηλαδή από το 1989 ως το 2011.

Δεδομένα:

Οι μετρήσεις έχουν χωριστεί και καταχωρηθεί σε αρχεία για διαφορετικά έτη, ως εξής:

`eruption1989.dat` : Διαδοχικές μετρήσεις σε χρονικό διάστημα μικρότερου του ενός μήνα (298 μετρήσεις).

`eruption2000.dat`: Διαδοχικές μετρήσεις για όλες τις μέρες την περίοδο 11/10 – 25/11 (739 μετρήσεις).

`eruption2001.dat, …, eruption2011.dat`: Διαδοχικές μετρήσεις για όλο το έτος για κάθε ένα από τα έτη 2001, 2002, …, 2011.

Τα αρχεία έχουν διαφορετικό αριθμό μετρήσεων.

Η κάθε ομάδα φοιτητών θα μελετήσει 4 χρονοσειρές, τις χρονοσειρές για τα έτη 1989, 2000 και 2011, καθώς και μια από τις χρονοσειρές για τα έτη 2001 ως 2010, αντίστοιχα με τον αριθμό από 1 ως 10 που του / τους αντιστοιχεί. Ο αριθμός αυτός ορίζεται από το άθροισμα των δύο ΑΕΜ mod 10, δηλαδή αν τα δύο ΑΕΜ είναι 6 και 8 τότε ο αριθμός που αντιστοιχεί είναι 14 mod 10 = 4.

Θεωρούμε πως όλες οι χρονοσειρές είναι στάσιμες.

Πρώτο στάδιο - Γραμμική ανάλυση για το έτος 1989, 2000 και 2011

Θέλουμε να ελέγξουμε αν το σύστημα που δημιουργεί τις εξάρσεις, και ειδικότερα την αναμονή των εξάρσεων, παραμένει το ίδιο στις διαφορετικές περιόδους, και συγκεκριμένα, για το 1989, το 2000 και το 2011 (διαφορά κατά 11 έτη). Θα χρησιμοποιήσετε τη χρονοσειρά των 298 παρατηρήσεων για το έτος 1989, και θα επιλέξετε ένα τμήμα ίδιου μήκους (298 παρατηρήσεων) της χρονοσειράς για το έτος 2000 και 2011. Η επιλογή του τμήματος για το 2000 και 2011 θα πρέπει να είναι τυχαία. Η ανάλυση θα γίνει σε κάθε μια από τις τρεις χρονοσειρές ίδιου μήκους για τα έτη 1989, 2000 και 2011, χρησιμοποιώντας τις ίδιες μεθόδους της γραμμικής ανάλυσης σε κάθε μια από τις 3 χρονοσειρές. Θα πρέπει να αντιμετωπίσετε τα παρακάτω ερωτήματα:

1. Είναι η χρονοσειρά λευκός θόρυβος (white noise) ή υπάρχουν σημαντικές αυτοσυσχετίσεις; Για αυτό θα πρέπει να συμπεριλάβετε και κατάλληλο έλεγχο υπόθεσης με βάση την αυτοσυσχέτιση. Είναι τα συμπεράσματα σας ίδια για τις 3 χρονοσειρές;

2. Σε συνέχεια του ερωτήματος 1 και αν το επιτρέπει το συμπέρασμα σας στο ερώτημα 1, ποιο είναι το πιο κατάλληλο γραμμικό μοντέλο προσαρμογής και πρόβλεψης; Πόσο καλή είναι η πρόβλεψη για ένα χρονικό βήμα μπροστά (ή και δύο ή τρία βήματα); Διαφέρουν τα μοντέλα και οι προβλέψεις για τις 3 χρονοσειρές;

Δεύτερο στάδιο – Γραμμική και Μη-γραμμική ανάλυση για το έτος που σας αντιστοιχεί

Στο δεύτερο στάδιο ανάλυσης, θα θεωρήσετε δύο χρονοσειρές που θα σχηματιστούν από το αρχείο που σας αντιστοιχεί, δηλαδή για κάποιο από τα έτη 2001 ως 2010. Η πρώτη χρονοσειρά θα περιλαμβάνει όλες τις παρατηρήσεις και η δεύτερη ένα τμήμα της χρονοσειράς μήκους 500 παρατηρήσεων.

Θα διερευνήσετε για το έτος που σας αντιστοιχεί αν η σειρά των αναμονών εξάρσεων έχει γραμμικές ή/και μη-γραμμικές (αυτο)συσχετίσεις και πως αυτό ανιχνεύεται αν θεωρήσουμε μια μικρή σειρά (500 παρατηρήσεων, τη δεύτερη χρονοσειρά) ή όλη τη διαθέσιμη σειρά (την πρώτη χρονοσειρά). Κάποια από τα βήματα της ανάλυσης που μπορείτε να κάνετε για κάθε μια από τις δύο χρονοσειρές είναι:

1. Σχεδιάγραμμα της χρονοσειράς.

2. Στατιστικός έλεγχος ανεξαρτησίας με βάση την αυτοσυσχέτιση (Portmanteau test).

3. Εκτίμηση της υστέρησης *τ* που δίνει το κριτήριο της αμοιβαίας πληροφορίας.

4. Εκτίμηση της διάστασης εμβύθισης *m* που σας δίνει το κριτήριο των ψευδών κοντινότερων γειτόνων (false nearest neighbors) χρησιμοποιώντας ως υστέρηση αυτή που βρήκατε παραπάνω.

5. Πρόβλεψη με τοπικό μοντέλο μέσου όρου και τοπικό γραμμικό μοντέλο για υστέρηση *τ* και διάσταση εμβύθισης *m* που εκτιμήσατε παραπάνω (και τον ίδιο αριθμό γειτονικών σημείων).

6. Σχεδιάγραμμα των προβλέψεων ενός βήματος μπροστά. Σύγκριση με τις αντίστοιχες προβλέψεις του γραμμικού μοντέλου (από το πρώτο μέρος της εργασίας). Σύγκριση του γραμμικού και μη-γραμμικού μοντέλου ως προς την πρόβλεψη για ένα βήμα μπροστά με βάση το nrmse.

7. Εκτίμηση της διάστασης συσχέτισης για διαστάσεις m=1,…,10, χρησιμοποιώντας την υστέρηση από το κριτήριο της αμοιβαίας πληροφορίας με κατάλληλα σχήματα.

Με βάση τα αποτελέσματα από τα μέτρα θα πρέπει να σχολιάσετε για τη μορφή του συστήματος της χρονοσειράς, δηλαδή αν είναι πλήρως στοχαστικό ή όχι και αν είναι γραμμικό ή μη-γραμμικό. Φαίνεται να διαφέρει το σύστημα της χρονοσειράς σας από το σύστημα του έτους 1989, 2000 και 2011;

Στην αναφορά που θα παρουσιάζονται τα αποτελέσματα της ανάλυσης θα πρέπει να συμπεριλάβετε πίνακες αποτελεσμάτων και σχήματα με αρίθμηση (π.χ. Πίνακας 1, Σχήμα 1) μέσα στο κείμενο στο σημείο που συζητούνται (όχι στο τέλος του κειμένου).