

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ, Θέμα εργασίας στο μάθημα «Χρονοσειρές»

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 και να υποβληθεί το σχετικό αρχείο, π.χ. τύπου Word, LaTeX, pdf.
- Τα γραφήματα και οι πίνακες θα πρέπει να παρουσιάζονται στο σημείο του κειμένου που αναφέρονται.
- Τα προγράμματα που χρησιμοποιήθηκαν θα πρέπει να είναι οργανωμένα σε αρχεία και να υποβληθούν μαζί με το αρχείο της αναφοράς.
- 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) ηλεκτρικής ενεργείας της Ιταλίας. Η επικράτεια της Ιταλίας χωρίζεται σε περιοχές / ζώνες ηλεκτρικής ενέργειας και σε κάθε περιοχή καταγράφεται η ζήτηση (demand) και η τιμή (price) της κιλοβατώρας σε ωριαία βάση. Για κάθε μέρα η τιμή καθορίζεται για όλες τις 24 ώρες από την προηγούμενη μέρα. Μας ενδιαφέρει να προβλέψουμε για κάθε ώρα 1,2,...,24, τη ζήτηση της επόμενης μέρας από τα δεδομένα ζήτησης ως και την προηγούμενη μέρα, και το ίδιο για την τιμή. Επίσης μας ενδιαφέρει να γνωρίζουμε τα χαρακτηριστικά του δυναμικού συστήματος ή στοχαστικής διαδικασίας που ορίζει την εξέλιξη της ζήτησης και τιμής στην επικράτεια της Ιταλίας και σε κάθε περιοχή.

Δεδομένα:

Τα ιστορικά δεδομένα δίνονται στο αρχείο `ElectricPowerItaly.xls` σε μορφή Excel στην Ενότητα «Εργασίες» στην ιστοσελίδα του μαθήματος στο elearning. Το αρχείο έχει δύο φύλλα εργασίας που ονομάζονται 'prices' και 'demand'. Τα δεδομένα στα δύο φύλλα εργασίας είναι οργανωμένα με τον ίδιο τρόπο. Στις στήλες 1-4 έχει το χρόνο (έτος, μήνας, μέρα, ώρα), όπου στην πρώτη στήλη η τιμή είναι πάντα 2011 και οι γραμμές είναι $365 \times 24 = 8760$, αφού τα στοιχεία είναι μόνο για το έτος 2011. Στη στήλη 5 είναι η τιμή (ζήτηση) για όλη την επικράτεια της Ιταλίας και στις στήλες 6-11 για κάθε μια από τις 6 περιοχές.

Η κάθε ομάδα θα μελετήσει δύο χρονοσειρές ως εξής. Σε κάθε ομάδα αντιστοιχεί ένας αριθμός. Ο αριθμός αυτός, έστω i , καθορίζει την ώρα της ημέρας που θα αναλύσει η ομάδα, που δίνεται ως $(i \bmod 24) + 1$ (π.χ. αν $i = 53$, $53 \bmod 24 = 5$ και η ώρα είναι $5 + 1 = 6$). Η πρώτη χρονοσειρά που θα μελετηθεί δίνεται από τις μετρήσεις ζήτησης (demand) και η δεύτερη από τις μετρήσεις τιμής (prices) για την i ώρα της ημέρας (χρονοσειρά μήκους 365) σε κάποια περιοχή της Ιταλίας ή σε όλη την επικράτεια, που καθορίζεται ως εξής. Αντιστοιχίζοντας την επικράτεια και τις 6 περιοχές με τη σειρά που εμφανίζονται στις στήλες 5-11 του αρχείου στους αριθμούς 1,2,...,7, καταχωρείται στην ομάδα με αριθμό i η περιοχή ή η επικράτεια που δίνεται ως $(i \bmod 7) + 1$ (π.χ. αν $i = 53$, $53 \bmod 7 = 4$ και η περιοχή είναι Sardegna που έχει αύξοντα αριθμό $4 + 1 = 5$). Στο ίδιο παράδειγμα, η ομάδα με αριθμό $i = 53$, θα κάνει την ανάλυση για την ώρα 6 (το πρωί) στις χρονοσειρές: 1) ζήτηση στη περιοχή της Σαρδηνίας (στήλη DEMAND Sardegna) και 2) τιμή στη περιοχή της Σαρδηνίας (στήλη PRICES Sardegna).

Πρώτο στάδιο - Γραμμική ανάλυση

Η γραμμική ανάλυση θα γίνει σε κάθε μια από τις δύο χρονοσειρές που αντιστοιχούν στην ομάδα σας. Δεν ορίζονται συγκεκριμένα βήματα για την ανάλυση και θα πρέπει να επιλέξετε τη διαδικασία που απαντά καλύτερα στα παρακάτω στάδια.

1. Απαλείψετε με βέλτιστο τρόπο την τάση και περιοδικότητα της χρονοσειράς.
2. Υπολογίστε και σχεδιάστε τη συνάρτηση αυτοσυσχέτισης στην στάσιμη χρονοσειρά. Ελέγξτε αν μπορεί η χρονοσειρά να θεωρηθεί λευκός θόρυβος.
3. Διερευνήστε το πιο κατάλληλο γραμμικό μοντέλο, προσαρμόστε το στη χρονοσειρά και υπολογίστε κάποιο στατιστικό του σφάλματος προσαρμογής, π.χ. το NRMSE. Περιγράψτε τη διερεύνηση που κάνατε και αιτιολογήστε την επιλογή σας. Αν συμπεράνετε πως η χρονοσειρά είναι λευκός θόρυβος, τότε το μοντέλο είναι μηδενικό και $\text{NRMSE}=1$.

Με βάση τα αποτελέσματα της ανάλυσης και μοντελοποίησης στις δύο χρονοσειρές σχολιάστε αν η ζήτηση και η τιμή για την ώρα και περιοχή που αντιστοιχεί στην ομάδα σας φαίνεται να έχουν κοινά γραμμικά χαρακτηριστικά (π.χ. παρόμοια συνάρτηση αυτοσυσχέτισης ή μερικής αυτοσυσχέτισης, παρόμοιο τύπο βέλτιστου μοντέλου, ίδιο επίπεδο σφάλματος προσαρμογής).

4. Κάνετε προβλέψεις της ζήτησης για την επόμενη μέρα (ένα βήμα μπροστά) με μοντέλο AR(5). Χρησιμοποιείτε διαφορετικό χρονικό σημείο για να χωρίσετε τη χρονοσειρά των 365 παρατηρήσεων σε δύο μέρη (σύνολο εκμάθησης και σύνολο αξιολόγησης) και για κάθε χρονικό σημείο υπολογίσετε το NRMSE για τις προβλέψεις στο σύνολο αξιολόγησης (π.χ. χρησιμοποιείτε τέσσερα διακριτά χρονικά σημεία). Σχολιάστε αν το NRMSE παραμένει σταθερό

με τη μετακίνηση του χρονικού σημείου διαχωρισμού. Κάνοντας το ίδιο και για την άλλη χρονοσειρά (της τιμής) συγκρίνετε τα δύο αποτελέσματα. Παρουσιάζεται το χαμηλότερο NRMSE στο ίδιο περίπου χρονικό σημείο διαχωρισμού για τη ζήτηση και την τιμή?

Δεύτερο στάδιο – Μη-γραμμική ανάλυση

Στο δεύτερο στάδιο ανάλυσης θέλουμε να διερευνήσουμε αν η κάθε μια από τις δύο χρονοσειρές (αφού έχει απαλειφθεί η τυχόν τάση και εποχικότητα) έχει μη-γραμμικές (αυτο)συσχετίσεις. Θα πρέπει να ακολουθήσετε τα παρακάτω βήματα:

1. Θα θεωρήσετε τη χρονοσειρά των υπολοίπων από την προσαρμογή με το βέλτιστο γραμμικό μοντέλο που βρήκατε στην γραμμική ανάλυση που κάνατε (ερώτημα 3 στο πρώτο στάδιο).
2. Θα δημιουργήσετε ένα πλήθος 20 χρονοσειρών iid (σειρά ανεξάρτητων και ισόνομων τυχαίων μεταβλητών), όπου η κάθε μια χρονοσειρά iid θα προέρχεται από την αρχική με τυχαία αντιμετάθεση των στοιχείων της (δες εντολή `randperm` στο Matlab).
3. Σε κάθε μια από τις 21 χρονοσειρές (αρχική χρονοσειρά και 20 χρονοσειρές iid) θα εκτιμήσετε κάποια γραμμικά και μη-γραμμικά στατιστικά (χαρακτηριστικά). Γραμμικό χαρακτηριστικό μπορεί να είναι η αυτοσυσχέτιση ή η μερική αυτοσυσχέτιση για κάποια συγκεκριμένη υστέρηση. Αντίστοιχα, μη-γραμμικό χαρακτηριστικό μπορεί να είναι η αμοιβαία πληροφορία για κάποια συγκεκριμένη υστέρηση, αλλά και άλλα όπως το ποσοστό ψευδών γειτόνων, η εκτιμώμενη διάσταση συσχέτισης, ο μέγιστος εκθέτης Lyapunov και το στατιστικό σφάλμα πρόβλεψης με τοπικό μοντέλο για κάποια συγκεκριμένη τιμή της διάστασης εμβύθισης ή/και άλλων παραμέτρων (υστέρησης, πλήθος κοντινότερων γειτόνων).
4. Θα διερευνήσετε κατά πόσο η τιμή του κάθε στατιστικού (που υπολογίσατε στο βήμα 3) στην πρώτη πραγματική χρονοσειρά διαφέρει από τις τιμές των 20 άλλων χρονοσειρών iid. Αν q_0 είναι η τιμή του στατιστικού για την αρχική χρονοσειρά και q_1, q_2, \dots, q_{20} , οι τιμές του ίδιου στατιστικού για τις 20 χρονοσειρές iid, θα εξετάσετε αν το q_0 ανήκει στην κατανομή των $\{q_1, q_2, \dots, q_{20}\}$. Αυτό μπορείτε να το δείξετε με ένα ιστόγραμμα και να το αποφασίσετε απλά εξετάζοντας αν το q_0 είναι μεγαλύτερο (ή μικρότερο) από τα q_1, q_2, \dots, q_{20} .

Η παραπάνω ανάλυση θα γίνει και στις δύο χρονοσειρές (ζήτησης και τιμής).

Με βάση τα αποτελέσματα από την παραπάνω ανάλυση θα πρέπει να σχολιάσετε για τη μορφή του συστήματος της χρονοσειράς, π.χ. στοχαστικό/αιτιοκρατικό, γραμμικό/μη-γραμμικό, χαμηλής/υψηλής διάστασης και πολυπλοκότητας, μικρή/μεγάλη μνήμη.

Στην αναφορά που θα παρουσιάζετε τα αποτελέσματα της ανάλυσης για τα δύο στάδια 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).