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

Το περιεχόμενο βίντεο (video content) αποτελεί μεγάλο μέρος της κίνησης του διαδικτύου, πάνω από 80% το 2020. Το δίκτυο, και ειδικότερα το δίκτυο πέμπτης γενιάς (5G network), για τον καταμερισμό περιεχομένου βίντεο θα πρέπει να μπορεί να προσαρμοστεί αυτόματα στην ζήτηση περιεχομένου βίντεο, π.χ. να παρέχει περισσότερες πηγές (εικονικοί εξυπηρετητές) σε αυξανόμενη ζήτηση και να απελευθερώνει πηγές σε μειωμένη ζήτηση. Για αυτό είναι σημαντικό να γίνεται ανάλυση της ζήτησης του περιεχομένου βίντεο (video content popularity) αυτόματα και να εντοπίζονται αλλαγές στη ζήτηση του. Η ανάλυση χρονοσειρών περιεχομένου βίντεο θεωρώντας τη ζήτηση περιεχόμενου βίντεο ως δυναμικό σύστημα αποτελεί επίκαιρο αντικείμενο έρευνας.

Η υπολογιστική εργασία είναι στο θέμα της ανάλυσης χρονοσειρών ζήτησης περιεχόμενου βίντεο. Τα δεδομένα δίνονται στο αρχείο VideoViews.xlsx που έχει ένα πλήθος από χρονοσειρές (μια σε κάθε στήλη) με μήκος 1199 (γραμμές). Η κάθε χρονοσειρά έχει τον αριθμό προβολών περιεχομένου ενός βίντεο στο YouTube για 1199 συνεχόμενες ημέρες. Για μια πρόσφατη ανάλυση τέτοιων δεδομένων, όπου δίνεται και η πηγή των δεδομένων, δες S. Skaperas, L. Mamatas and A. Chorti, "Real-Time Video Content Popularity Detection Based on Mean Change Point Analysis," in IEEE Access, vol. 7, pp. 142246-142260, 2019.

Για κάθε ομάδα αντιστοιχούν δύο χρονοσειρές, η χρονοσειρά Α με αύξοντα αριθμό ίδιο με τον αύξοντα αριθμό της ομάδας και η χρονοσειρά Β με αύξοντα αριθμό αυτόν της Α συν 10.

Ανάλυση χρονοσειρών και σημεία αλλαγής ζήτησης

Βασικός σκοπός της ανάλυσης της χρονοσειράς περιεχομένου βίντεο είναι να εντοπίσουμε σημαντικές αλλαγές στη ζήτηση όπως αυτή αποτυπώνεται στις ημερήσιες προβολές του βίντεο.

Δεν ορίζονται συγκεκριμένα βήματα για την ανάλυση και θα πρέπει να επιλέξετε τη διαδικασία που απαντά καλύτερα στα παρακάτω στάδια.

1. Πρώτα θα διερευνήσετε αν η χρονοσειρά Α έχει αυτοσυσχέτιση που δεν οφείλεται σε τυχόν (στοχαστική) τάση που έχει.
2. Θα διερευνήστε το πιο κατάλληλο γραμμικό μοντέλο τύπου ARMA ή ARIMA, αντίστοιχα αν δεν έχει ή έχει τάση, που προσαρμόζεται στη χρονοσειρά Α και θα υπολογίστε κάποιο στατιστικό του σφάλματος προσαρμογής, π.χ. το NRMSE, για τη στάσιμη χρονοσειρά. Περιγράψτε τη διερεύνηση που κάνατε και αιτιολογείστε την επιλογή σας. Αν συμπεράνετε πως η χρονοσειρά είναι λευκός θόρυβος, τότε το μοντέλο είναι μηδενικό και NRMSE=1.
3. Επαναλάβετε τα παραπάνω δύο βήματα για την χρονοσειρά Β. Συγκρίνετε τα μοντέλα που βρήκατε ως πιο κατάλληλα στις δύο χρονοσειρές.
4. Στη συνέχεια θα αναπτύξετε ένα πρόγραμμα που θα εντοπίζει σημαντικές αλλαγές στη χρονοσειρά προβολών του βίντεο. Ο εντοπισμός ενός σημείου αλλαγής θα γίνει από κάποιο στατιστικό των σφαλμάτων πρόβλεψης, έτσι ώστε όταν το στατιστικό των σφαλμάτων πρόβλεψης υπερβαίνει κάποιο όριο (κατώφλι) α να σηματοδοτεί σημείο αλλαγής. Συγκεκριμένα στο πλαίσιο της εργασίας αυτής προτείνεται να δοκιμαστεί ως στατιστικό ο μέσος όρος των απολύτων τιμών των σφαλμάτων για προβλέψεις πολλών βημάτων μπροστά ως και το βήμα T

$$S_n = \frac{1}{\mathrm{T}} \sum_{k=1}^{\mathrm{T}} \left| x_{n+k} - x_n(k) \right|$$

   Με βάση κάποια χρονική στιγμή n οι προβλέψεις για τα T επόμενα χρονικά βήματα είναι $x_n(k)$, k=1,...,5, και οι αντίστοιχες πραγματικές τιμές είναι x_{n+k}, k=1,...,5. Όταν $S_n > \alpha$, θα θεωρήσετε πως η ζήτηση έχει αλλάξει και θα ορίσετε το χρονικό σημείο $n+T$ ως σημείο αλλαγής. Θα πρέπει να ορίσετε το πλήθος χρονικών βημάτων T και το όριο α, π.χ. T=3 και α=2s, όπου s η τυπική απόκλιση των παρατηρήσεων στο σύνολο εκμάθησης του μοντέλου.

   Το πρόγραμμα θα ξεκινά με την προσαρμογή του μοντέλου που βρήκατε στο βήμα 2 στις πρώτες 400 παρατηρήσεις. Στη συνέχεια για n=400 θα υπολογίζει το S_n. Αν το κριτήριο αλλαγής ικανοποιείται, $S_n > \alpha$, η χρονική στιγμή $n+T$ θα

προστίθεται στη λίστα των σημείων αλλαγής και ο χρόνος θα αυξάνει κατά *T*, δηλαδή $n \leftarrow n+T$, αλλιώς ο χρόνος θα αυξάνει κατά ένα, δηλαδή $n \leftarrow n+1$. Στη συνέχεια θα εξεταστεί το κριτήριο για το νέο *n* και αυτό θα συνεχίζεται ως το τέλος της χρονοσειράς. Για τη χρήση του μοντέλου έχετε τρεις επιλογές: a) το μοντέλο μπορεί να παραμείνει το ίδιο, δηλαδή αυτό που εκτιμήθηκε στις πρώτες 400 παρατηρήσεις, b) το μοντέλο εκτιμάται σε κάθε νέα χρονική στιγμή *n* στις τελευταίες 400 παρατηρήσεις, δηλαδή για χρόνους n-399, n-398,…,n-1,n, c) το μοντέλο παραμένει το ίδιο όσο δεν εντοπίζεται σημείο αλλαγής και εκτιμάται στις τελευταίες 400 παρατηρήσεις κάθε φορά που εντοπίζεται σημείο αλλαγής (για $n \leftarrow n+T$). Από τις τρεις προσεγγίσεις, προγραμματιστικά η προσέγγιση b) είναι η πιο απλή αφού μπορείτε σε κάθε νέα χρονική στιγμή *n* να καλείται τη συνάρτηση predictARMAmultistep, ενώ για τις a) και c) θα πρέπει εσείς να υπολογίζετε τις προβλέψεις χρησιμοποιώντας τις παραμέτρους του πρόσφατου μοντέλου (όταν δε χρειάζεται να το εκτιμήσετε).

Στο τέλος θα παρουσιάζετε στο γράφημα της χρονοσειράς Α τα χρονικά σημεία αλλαγής που βρήκατε.

5. Θα επαναλάβετε το βήμα 4 για τη χρονοσειρά Β. Θα σχολιάσετε για το κατώφλι α και τα χρονικά βήματα Τ που επιλέξατε για τις δύο χρονοσειρές (αν είναι τα ίδια και γιατί, ή αν είναι διαφορετικά και γιατί). Φαίνεται η προσέγγιση αυτή να δίνει με αυτόματα τρόπο σημεία αλλαγής που να φαίνονται χρήσιμα?
6. Θα επαναλάβετε τα βήματα 4 και 5 χρησιμοποιώντας μη-γραμμικό μοντέλο για την στάσιμη χρονοσειρά (που θα μετατρέπετε σε στάσιμη αν δεν είναι). Το μη-γραμμικό μοντέλο θα είναι ένα τοπικό μοντέλο κοντινότερων γειτόνων που θα επιλέξετε αιτιολογώντας την επιλογή σας. Υπάρχει σχετική συνέπεια στα αποτελέσματα με το γραμμικό και το μη-γραμμικό μοντέλο για τη χρονοσειρά Α και Β?
7. Σχολιάστε αν τα αποτελέσματα σας δείχνουν να είναι χρήσιμη η προσέγγιση του εντοπισμού σημείων αλλαγής με το στατιστικό σφαλμάτων πρόβλεψης, ή/και αν προτείνετε κάποιον άλλον τρόπο.

Στην αναφορά που θα παρουσιάζετε 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 συζητιούνται (όχι στο τέλος του κειμένου).