Μεταπτυχιακό Υπολογιστικής Φυσικής, Δεκέμβριος 2021
Εργασία στο μάθημα 'Ανάλυση Δεδομένων'

Δημήτρης Κουγιουμτζής

E-mail: dkugiu@auth.gr

30 Δεκεμβρίου 2021

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 Koygioymtzhw και το ζήτημα 5. Για τα προγράμματα τα ονόματα των αρχείων θα είναι KoygioymtzhwExe5Prog1.m, KoygioymtzhwExe5Prog2.m κτλ. Για τις συναρτήσεις τα ονόματα των αρχείων θα είναι KoygioymtzhwExe5Fun1.m, KoygioymtzhwExe5Fun2.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. Μη χρησιμοποιείτε διπλή απόστροφο (") αλλά απλή (') για να δηλώσετε οποιαδήποτε σειρά αλφαριθμητικών χαρακτήρων (όνομα αρχείου, κείμενο για εμφάνιση ή εκτύπωση κτλ) γιατί δεν είναι αποδεκτό σε παλιότερες εκδόσεις του Matlab.

- Θα υποβληθούν μόνο τα αρχεία Matlab και τυχόν αρχεία δεδομένων άλλα από αυτά που σας δίνονται και που έχετε χρησιμοποιήσει (μέσω του elearning).

- 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

Η εργασία περιλαμβάνει μια σειρά ζητημάτων που αφορούν την πανδημία του κορονοϊού. Τα δεδομένα που θα χρησιμοποιήσετε δίνονται στα αρχεία: 1) `ECDC-7Days-Testing.xlsx` και 2) `FullEodyData.xlsx`. Και τα δύο αρχεία είναι από την ιστοσελίδα https://www.stelios67pi.eu, όπου υπάρχουν και πολλά άλλα στοιχεία που ενδεχομένως να θέλετε να αντλήσετε. Το πρώτο αρχείο έχει εβδομαδιαία δεδομένα για αριθμό κρουσμάτων και τεστ και κυρίως για το δείκτη θετικότητας (στήλη positivity_rate) για χώρες της Ευρώπης όπως τα καταγράφει το European Center for Disease Control (ECDC). Το δεύτερο αρχείο έχει διάφορα στοιχεία και δείκτες σχετικά με τον Covid-19 για την Ελλάδα σε ημερήσια βάση,

όπως τεστ και κρούσματα, νοσηλείες σε απλές κλίνες και μονάδες εντατικής θεραπείας (ΜΕΘ), θάνατοι κτλ.

Στα ζητήματα της εργασίας που γίνεται αναφορά σε μια χώρα της Ευρώπης από τις 25 Ευρωπαϊκές χώρες που δίνονται στο αρχείο `EuropeanCountries.xlsx`, θα ορίσετε αυτήν τη χώρα ως εξής. Η χώρα που σας αντιστοιχεί είναι αυτή με αύξοντα αριθμό το υπόλοιπο της διαίρεσης αυξημένο κατά ένα του ΑΕΜ σας με το 25. Υπάρχει η δυνατότητα να χρησιμοποιηθεί μια άλλη γειτονική Ευρωπαϊκή χώρα αν κριθεί πως δεν υπάρχουν ικανοποιητικά δεδομένα για τα ζητήματα της εργασίας από την Ευρωπαϊκή χώρα που αρχικά ορίστηκε. Για παράδειγμα για ΑΕΜ=4321 ο αύξων αριθμός χώρας είναι 22 και αντιστοιχεί στη Σλοβακία. Ας ονομάσουμε τη χώρα που επιλέξατε ως χώρα Α.

Μπορείτε επίσης να αντλήσετε στοιχεία για τα ζητήματα της εργασίας από άλλες πηγές αν κρίνετε πως είναι πιο εύκολα επεξεργάσιμα ή πιο ακριβή και καλύτερα επικαιροποιημένα.

Ζητήματα εργασίας

Για όλα τα ζητήματα στην αρχή του κάθε προγράμματος θα φορτώνεται το σχετικό αρχείο δεδομένων. Αν είναι άλλο από τα τρία αρχεία δεδομένων που δίνονται για την εργασία θα πρέπει να υποβληθεί και αυτό. Υπάρχει ελεύθερη επιλογή στην παρουσίαση των αποτελεσμάτων (γραφήματα, συμπεράσματα και αποτελέσματα στη γραμμή εντολών). Αν για κάποια χώρα, βδομάδα και/ή μέρα δεν υπάρχει τιμή ή είναι αρνητική για το δείκτη που θα χρησιμοποιήσετε, μπορείτε είτε να την παραβλέψετε (και το δείγμα θα έχει μια παρατήρηση λιγότερη) ή να την αναπληρώσετε από τιμή που θα βρείτε σε άλλη πηγή ή να τη διορθώσετε με κάποιο αιτιολογημένο τρόπο.

Σημειώνεται πως για κάθε Ζήτημα θα πρέπει πρώτα να παρουσιάσετε τα δεδομένα που θα χρησιμοποιήσετε στην ανάλυση με τρόπο που σας επιτρέπει να αποκτήσετε μια πρώτη υποκειμενική εντύπωση για το ερώτημα του Ζητήματος (δεν είναι απαραίτητο να αναφέρεται στο Ζήτημα).

1. *Ποια είναι η κατανομή του δείκτη θετικότητας στις Ευρωπαϊκές χώρες για δύο συγκεκριμένες ημερομηνίες (σε επίπεδο εβδομάδας); Υπάρχει κάποια γνωστή παραμετρική κατανομή να την προσεγγίζει;* Για να απαντήσετε αυτά τα ερωτήματα θα χρησιμοποιήσετε όλες τις 25 Ευρωπαϊκές χώρες που δίνονται στο αρχείο `EuropeanCountries.xlsx` και θα θεωρήσετε δύο διαφορετικές ημερομηνίες (βδομάδες) στο αρχείο `ECDC-7Days-Testing.xlsx`: 1) μια από τις τελευταίες 6 βδομάδες του 2021 (W45-W50) στην οποία ο δείκτης θετικότητας της χώρας Α είναι μέγιστος και 2) το ίδιο για τις αντίστοιχες 6 βδομάδες του 2020 (W45-W50). Θα σχηματίσετε τα ιστογράμματα του δείκτη θετικότητας από τις 25 χώρες για τις δύο ημερομηνίες. Θα προσαρμόσετε μια κατάλληλη γνωστή παραμετρική κατανομή (κανονική, ομοιόμορφη, εκθετική, δες επίσης συνάρτηση `fitdist`). Επίσης θα διερευνήσετε κατά πόσο μπορούν (ή δε μπορούν) να περιγραφούν ικανοποιητικά οι δύο κατανομές για τις δύο ημερομηνίες με την ίδια παραμετρική κατανομή. Η απάντηση σας εδώ μπορεί να δοθεί απλά με βάση τα κατάλληλα γραφήματα των παραμετρικών κατανομών στα αντίστοιχα ιστογράμματα.

2. *Διαφέρουν οι δύο κατανομές του δείκτη θετικότητας στο Ζήτημα 1);* Για να απαντήσετε σε αυτό το ερώτημα θα κάνετε έλεγχο ισότητας δύο κατανομών. Για τον έλεγχο θα

χρησιμοποιήσετε το στατιστικό Kolmogorov-Smirnov που είναι η μέγιστη διαφορά μεταξύ των αθροιστικών σχετικών συχνοτήτων για το δείκτη θετικότητας στις δύο χρονικές περιόδους. Η αθροιστική σχετική συχνότητα εκτιμά την αθροιστική συνάρτηση κατανομής $\hat{F}_X(x) = i/n$, όπου i είναι η τάξη της παρατήρησης x της τ.μ. X στη λίστα αύξουσας σειράς των n παρατηρήσεων, και δηλώνει την αναλογία των παρατηρήσεων στο δείγμα που είναι μικρότερες ή ίσες της x. Το στατιστικό Kolmogorov-Smirnov είναι $\max_x |\hat{F}_X(x) - \hat{F}_Y(x)|$, όπου X είναι ο δείκτης θετικότητας στην Ευρώπη για την πρώτη περίοδο και Y για τη δεύτερη περίοδο. Θεωρώντας πως δε γνωρίζουμε κάποια γνωστή παραμετρική κατανομή αυτού του στατιστικού θα κάνετε έλεγχο τυχαιοποίησης. Συγκεκριμένα, θα θεωρήσετε το κοινό δείγμα μεγέθους $n+m$ και των δύο περιόδων μαζί, όπου n και m είναι τα μεγέθη της πρώτης και δεύτερης περιόδου (όπου στην περίπτωση μας τα n και m είναι ίσα). Θα τυχαιοποιήσετε τη σειρά των $n + m$ τιμών του δείκτη θετικότητας και τις n πρώτες τιμές θα τις αντιστοιχίσετε στο X και τις υπόλοιπες m στο Y. Θα επαναλάβετε αυτό M φορές για να πάρετε M τυχαιοποιημένα δείγματα για X και Y και να υπολογίσετε σε αυτά το στατιστικό Kolmogorov-Smirnov.

3. *Πότε ο εβδομαδιαίος δείκτης θετικότητας της Ελλάδας είναι στατιστικά σημαντικά διαφορετικός από αυτόν της Ευρωπαϊκής Ένωσης (ΕΕ);* Για να απαντήσετε αυτό το ερώτημα πρώτα θα δημιουργήσετε μια συνάρτηση που θα δίνει την απάντηση σε αυτό το ερώτημα για μια οποιαδήποτε εβδομάδα. Οι ημερήσιες τιμές του δείκτη θετικότητας για την Ελλάδα μπορούν να υπολογιστούν από τα ημερήσια τεστ (rapid και PCR στις στήλες `AS` και `AT` αντίστοιχα στο αρχείο `FullEodyData.xlsx`) και τα ημερήσια νέα κρούσματα (στη στήλη `B` στο ίδιο αρχείο). Ο εβδομαδιαίος δείκτης θετικότητας είναι ο μέσος όρος του δείκτη θετικότητας των 7 ημερών της εβδομάδας. Τον εβδομαδιαίο δείκτη θετικότητας της ΕΕ μπορείτε να τον υπολογίσετε από τις τιμές των εβδομαδιαίων δεικτών θετικότητας των 25 χωρών που δίνονται στο αρχείο `ECDC-7Days-Testing.xlsx` ή πιο εύκολα διαβάζοντας τις τιμές από το αντίστοιχο γράφημα στην ιστοσελίδα https://www.stelios67pi.eu/testing.html. Η συνάρτηση θα δέχεται τις 7 συνεχόμενες ημερήσιες τιμές του δείκτη θετικότητας στην Ελλάδα για μια συγκεκριμένη βδομάδα και τον αντίστοιχο εβδομαδιαίο δείκτη θετικότητας της ΕΕ, θα υπολογίζει το 95% διάστημα εμπιστοσύνης bootstrap (δες άσκηση 5.5) για το μέσο δείκτη θετικότητας της Ελλάδας σε μια βδομάδα (7 μέρες) και η απάντηση για το αν διαφέρει με στατιστική σημαντικότητα θα δίνεται συγκρίνοντας την τιμή του εβδομαδιαίου δείκτη θετικότητας της ΕΕ με το διάστημα εμπιστοσύνης. Στην έξοδο η συνάρτηση θα πρέπει να δίνει και το πρόσημο της διαφοράς αν υπάρχει διαφορά. Σε ένα πρόγραμμα θα κάνετε τους υπολογισμούς για μια περίοδο συνεχόμενων 12 βδομάδων (καλώντας τη συνάρτηση 12 φορές) από την εβδομάδα τελευταίας κορύφωσης του δείκτη θετικότητας της χώρας Α και προς τα πίσω (τα στοιχεία για τη χώρα Α υπάρχουν στο αρχείο `ECDC-7Days-Testing.xlsx`). Το πρόγραμμα θα πρέπει να σχηματίζει το διάγραμμα για τον εβδομαδιαίο δείκτη θετικότητας για Ελλάδα και ΕΕ για την περίοδο ενδιαφέροντος και να σημειώνονται / φαίνονται με κάποιον τρόπο οι στατιστικά σημαντικές διαφορές.

4. *Υπάρχουν σημαντικές διαφορές στο δείκτη θετικότητας στην Ευρώπη στο τελευταίο δίμηνο με το αντίστοιχο του 2020;* Για να απαντήσετε σε αυτό το ερώτημα για τη χώρα Α θα συγκρίνετε στις δύο περιόδους 2 μηνών, δηλαδή W42-W50 για το 2021 και 2020, το μέσο εβδομαδιαίο δείκτη θετικότητας (σε περίοδο δύο μηνών) στην πρώτη και δεύτερη περίο-

δο. Η σύγκριση θα γίνει με κατάλληλο παραμετρικό έλεγχο και έλεγχο τυχαιοποίησης (όπως για το Ζήτημα 2). Θα επαναλάβετε τους ίδιους ελέγχους για άλλες 4 χώρες, που είναι αλφαβητικά γειτονικές της χώρας Α στη λίστα των 25 Ευρωπαϊκών χωρών. Υπάρχει συμφωνία στα αποτελέσματα των συγκρίσεων στις 5 χώρες;

5. *Με ποια από 5 χώρες της Ευρώπης η πορεία του εβδομαδιαίου δείκτη θετικότητας της Ελλάδας συσχετίζεται το τελευταίο τρίμηνο;* Για να απαντήσετε αυτό το ερώτημα θα θεωρήσετε τις ίδιες 5 χώρες που χρησιμοποιήσατε στο Ζήτημα 4 και την περίοδο 3 μηνών, δηλαδή W38-W50 για το 2021, και θα σχηματίσετε την πορεία των 5 εβδομαδιαίων δεικτών θετικότητας στην περίοδο ενδιαφέροντος. Θα υπολογίσετε τον συντελεστή συσχέτισης Pearson για το ζευγάρι του εβδομαδιαίου δείκτη θετικότητας της Ελλάδας με κάθε μια από τις 5 χώρες. Θα κάνετε επίσης έλεγχο σημαντικότητας του συντελεστή συσχέτισης, παραμετρικό και τυχαιοποίησης, για κάθε ένα από τα 5 ζευγάρια χωρών σε επίπεδο σημαντικότητας α=0.05 και α=0.01. Υπάρχει στατιστικά σημαντική συσχέτιση και με ποια χώρα είναι μεγαλύτερη;

6. *Συσχετίζεται σημαντικά πιο ισχυρά με κάποια από 5 χώρες της Ευρώπης ο εβδομαδιαίος δείκτης θετικότητας της Ελλάδας το τελευταίο τρίμηνο;* Σε συνέχεια του Ζητήματος 5 θέλουμε να ελέγξουμε αν, για τις δύο χώρες που ο εβδομαδιαίος δείκτης θετικότητας συσχετίζεται περισσότερο με το δείκτη της Ελλάδας, η διαφορά των αντίστοιχων συντελεστών συσχέτισης είναι στατιστικά σημαντική. Έστω η μέγιστη τιμή του συντελεστή συσχέτισης είναι για το ζευγάρι χωρών (Α,Β) και η δεύτερη μεγαλύτερη τιμή για το ζευγάρι (Α,Γ). Για να ελέγξετε αν ο συντελεστής συσχέτισης των (Α,Β) διαφέρει σημαντικά με αυτόν του (Α,Γ) θα κάνετε έλεγχο ισότητας συντελεστών συσχέτισης τυχαιοποίησης, επιλέγοντας τυχαία χωρίς επανάθεση από το κοινό δείγμα ζευγαρωτών παρατηρήσεων των (Α,Β) και (Α,Γ) (όπως για το Ζήτημα 2).

7. *Μπορώ να προβλέψω τους θανάτους από κορονοϊό μιας χώρας από τον εβδομαδιαίο δείκτη θετικότητας κάποιας προηγούμενης βδομάδας;* Θα απαντήσετε αυτό το ερώτημα για τη χώρα Α. Θα πρέπει να βρείτε για ποια υστέρηση εβδομάδας του εβδομαδιαίου δείκτη θετικότητας με αναφορά στην εβδομάδα που αναφέρονται οι θάνατοι είναι καλύτερο το γραμμικό μοντέλο απλής παλινδρόμησης του εβδομαδιαίου αριθμού θανάτων (ανά εκατομμύριο κατοίκων) ως προς τον εβδομαδιαίο δείκτη θετικότητας κάποιας εβδομάδας πριν. Τον εβδομαδιαίο αριθμό θανάτων (ανά εκατομμύριο κατοίκων) για τη χώρα Α μπορείτε να το διαβάσετε στο σχετικό γράφημα στην ιστοσελίδα https://www.stelios67pi.eu/testing.html. Θα δοκιμάσετε τα μοντέλα για υστέρηση ως και 5 βδομάδες πριν. Θα πρέπει να επιλέξετε δύο διαφορετικές περιόδους 4 μηνών (16 εβδομάδων, ελεύθερη επιλογή) και θα εφαρμόσετε τη διαδικασία με προσαρμογή των μοντέλων στα δεδομένα για την κάθε μια από τις δύο περιόδους ξεχωριστά. Φαίνεται να συμφωνούν τα συμπεράσματα για την υστέρηση εβδομάδας του δείκτη θετικότητας που δίνει την καλύτερη πρόβλεψη θανάτων στις δύο περιόδους;

8. *Για την Ελλάδα, μπορώ να προβλέψω ημερήσιους θανάτους λόγω κορονοϊού γνωρίζοντας τους δείκτες θετικότητας για πολλές μέρες πριν;* Θα πρέπει να βρείτε το καλύτερο μοντέλο πολλαπλής γραμμικής παλινδρόμησης που προβλέπει τον ημερήσιο αριθμό θανάτων για την Ελλάδα από το δείκτη θετικότητας σε προηγούμενες μέρες, πηγαίνοντας ως και 30

μέρες πίσω (30 ανεξάρτητες μεταβλητές). Θα πρέπει να συγκρίνετε ως προς την προσαρμογή του στα δεδομένα το πλήρες μοντέλο με τις 30 ανεξάρτητες μεταβλητές με το μοντέλο βηματικής παλινδρόμησης. Για τη σύγκριση των δύο μοντέλων θα πρέπει να χρησιμοποιήσετε τον προσαρμοσμένο συντελεστή πολλαπλού προσδιορισμού. Θα πρέπει να επιλέξετε δύο διαφορετικές περιόδους 3 μηνών (12 εβδομάδων, ελεύθερη επιλογή) και να εφαρμόσετε τη διαδικασία με προσαρμογή των δύο μοντέλων στα δεδομένα για την κάθε μια από τις δύο περιόδους ξεχωριστά. Συγκρίνετε τα βέλτιστα μοντέλα που επιλέξατε στις δύο περιόδους ως προς τη δομή τους και την καταλληλότητα προσαρμογής τους.

9. *Πως μπορώ να ελέγξω αν το μοντέλο μου είναι κατάλληλο για προβλέψεις;* Θα απαντήσετε αυτό το ερώτημα στο πλαίσιο του Ζητήματος 8 και θα χρησιμοποιήσετε διασταυρωμένη επικύρωση (cross-validation) για να υπολογίσετε κάποιο στατιστικό σφάλματος πρόβλεψης ή τον (προσαρμοσμένο) συντελεστή πολλαπλού προσδιορισμού. Συγκεκριμένα θα χωρίσετε το σύνολο των δεδομένων σε 5 ισοπληθή μέρη, π.χ. αν το σύνολο των παρατηρήσεων είναι 56 θα χωρίσετε σε 5 μέρη των 11 παρατηρήσεων (το τελευταίο μπορεί να έχει 12 ώστε να περιέχονται όλες οι παρατηρήσεις). Θα αφαιρείτε το ένα από τα 5 μέρη και θα προσαρμόζετε το μοντέλο στα υπόλοιπα 4 μέρη. Στη συνέχεια θα προβλέπετε στο μέρος που θα αφαιρέσετε. Θα επαναλάβετε τη διαδικασία αυτή 5 φορές (μια για κάθε μέρος) και στο τέλος θα συλλέξετε τις προβλέψεις για να υπολογίσετε το στατιστικό σφάλματος πρόβλεψης ή τον (προσαρμοσμένο) συντελεστή πολλαπλού προσδιορισμού. Θα επαναλάβετε τη διαδικασία σύγκρισης μοντέλων όπως στο Ζήτημα 8 με την προσέγγιση της διασταυρωμένης επικύρωσης. Σχολιάστε τα αποτελέσματα με την απλή προσαρμογή στο Ζήτημα 8 και την διασταυρωμένη επικύρωση.

10. *Για την Ελλάδα, μπορώ να προβλέψω ημερήσιους θανάτους λόγω κορονοϊού γνωρίζοντας άλλους σχετικούς δείκτες και για πολλές μέρες πριν;* Ενώ στο Ζήτημα 8 διερευνήσαμε την εξάρτηση του αριθμού ημερήσιων θανάτων στην Ελλάδα από τον δείκτη θετικότητας τις προηγούμενες μέρες, εδώ θέλουμε να συμπεριλάβουμε στο μοντέλο άλλους (ή και άλλους) δείκτες που μπορεί να έχουν προβλεπτική ικανότητα. Τέτοιους δείκτες (όπως δείκτες νοσηλείας σε απλές κλίνες, διασωληνομένων εμβολιασμένων και ανεμβολίαστων), μπορείτε να βρείτε στο αρχείο `FullEodyData.xlsx` (ελεύθερη επιλογή). Μπορείτε να θεωρήσετε και υστέρηση ημερών για τον κάθε δείκτη όπως στο Ζήτημα 8 για το δείκτη θετικότητας (ελεύθερη επιλογή υστερήσεων). Μπορείτε να επιλέξετε ελεύθερα την περίοδο στην οποία θα προσαρμόσετε και αξιολογήσετε τα μοντέλα αλλά θα πρέπει αυτή να επεκτείνεται ως τις 27/12/2021. Έχοντας επιλέξει το κατάλληλο μοντέλο (πλήρες μοντέλο ή αυτό από τη βηματική παλινδρόμηση) θα πρέπει να κάνετε προβλέψεις ημερήσιων θανάτων τουλάχιστον για την 28/12/2021. Μπορείτε επίσης να επεκτείνετε τις προβλέψεις και σε μεταγενέστερες ημέρες αντλώντας πιο πρόσφατα στοιχεία για τους δείκτες σας από το επικαιροποιημένο αρχείο `FullEodyData.xlsx` που δίνεται στην ιστοσελίδα https://www.stelios67pi.eu/eody.html.