# Εργασία στο μάθημα 'Ανάλυση Δεδομένων', Ιανουάριος 2020

**Δημήτρης Κουγιουμτζής**

E-mail: dkugiu@auth.gr

31 Δεκεμβρίου 2020

**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 η ομάδα φοιτητών No 10 και το ζήτημα 5. Για τα προγράμματα τα ονόματα των αρχείων θα είναι Group10Exe5Prog1.m, Group10Exe5Prog2.m κτλ (αν χρειάζεται να δημιουργήσετε περισσότερα από ένα προγράμματα). Για τις συναρτήσεις τα ονόματα των αρχείων θα είναι Group10Exe5Fun1.m, Group10Exe5Fun2.m κτλ. Στην αρχή κάθε προγράμματος και συνάρτησης θα υπάρχουν (σε σχολιασμό) τα ονοματεπώνυμα των μελών της ομάδας.
- 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).
- Η κάθε εργασία (σύνολο προγραμμάτων και συναρτήσεων Matlab) 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.).

## Περιγραφή εργασίας

Η εργασία συμπεριλαμβάνει μια σειρά ζητημάτων που αφορούν την πανδημία του κορονοϊού. Τα δεδομένα που θα αναλύσετε είναι τα ημερήσια κρούσματα κορονοϊού και οι ημερήσιοι θάνατοι σχετικοί με κορονοϊό. Τα δεδομένα έχουν εξαχθεί από European Center for Disease Control (ECDC). Τα αρχεία `Covid19Confirmed.xlsx` και `Covid19Deaths.xlsx` που δίνονται στην ενότητα της εργασίας στο elearning έχουν τις τιμές των ημερήσιων νέων κρουσμάτων και θανάτων από την αρχή της πανδημίας και συγκεκριμένα την περίοδο 1/1/2020 ως 13/12/2020. Στην πρώτη στήλη είναι το όνομα της χώρας, στη δεύτερη στήλη η ήπειρος που ανήκει, στην τρίτη στήλη ο πληθυσμός της χώρας και στις υπόλοιπες στήλες οι τιμές για κάθε μέρα (η αντίστοιχη ημερομηνία δίνεται στην πρώτη γραμμή σε μορφή μέρα/μήνας/έτος). Στα δύο αρχεία υπάρχουν τα δεδομένα μόνο για χώρες με πληθυσμό πάνω από 1 εκ. και είναι 156 αυτές οι χώρες. Οι τρίτες πρώτες στήλες είναι ίδιες στα δύο αρχεία.

Θα αναλύσετε τους δύο τύπους δεδομένων από μια χώρα καθώς και από άλλες χώρες, όπως περιγράφεται στα ζητήματα της εργασίας. Η χώρα που αντιστοιχεί στην ομάδα δίνεται από το υπόλοιπο της διαίρεσης του ΑΕΜ (αν η ομάδα αποτελείται από δύο μέλη, του ενός από τα δύο μέλη) με το 156 αυξημένο κατά ένα, αν είναι Ευρωπαϊκή, και αν όχι η κοντινότερη Ευρωπαϊκή χώρα σε αυτό τον αύξοντα αριθμό. Υπάρχει η δυνατότητα να χρησιμοποιηθεί μια άλλη γειτονική Ευρωπαϊκή χώρα (ως +/-4 από τον αύξοντα αριθμό) αν κριθεί πως δεν υπάρχουν ικανοποιητικά δεδομένα για τα ζητήματα της εργασίας από την Ευρωπαϊκή χώρα που αρχικά επιλέχθηκε. Για παράδειγμα για ΑΕΜ=10000 ο αύξων αριθμός χώρας είναι 17 και αντιστοιχεί στην Μποτσουάνα και η πλησιέστερη Ευρωπαϊκή χώρα είναι η Βόσνια-Ερζεγοβίνη με αύξοντα αριθμό 16. Αν θεωρήσετε πως δεν υπάρχουν αρκετά στοιχεία για τη Βόσνια-Ερζεγοβίνη, μπορείτε να επιλέξετε τη Βουλγαρία ή το Βέλγιο που είναι στο εύρος των χωρών +/-4, δηλαδή με αύξοντα αριθμό από 12 ως 20. Ας ονομάσουμε τη χώρα που επιλέξατε ως χώρα Α.

Η μελέτη θα επικεντρωθεί στο λεγόμενο πρώτο κύμα, δηλαδή στην πρώτη έξαρση και πτώση της διασποράς του κορονοϊού. Ο χρόνος έναρξης και λήξης για το πρώτο κύμα ορίζεται από την ομάδα με κριτήρια που θέτει και μπορεί να διαφέρει για τα ημερήσια κρούσματα και τους ημερήσιους θανάτους. Μια καλή πρακτική είναι για την έναρξη η αρχή της ημερήσια αύξησης των κρουσμάτων / θανάτων και για τη λήξη η ισοπέδωση ή μηδενισμός μετά από πτώση του πλήθους κρουσμάτων / θανάτων. Η ομάδα μπορεί επίσης να επικαιροποιήσει και διορθώσει στοιχεία (για κρούσματα και θανάτους) από άλλες πηγές αν κρίνει πως κάποια στοιχεία (για χώρες και μέρες) δεν είναι ακριβή. Για άλλες πηγές δες π.χ. `https://www.worldometers.info/coronavirus` και COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University στη διεύθυνση `https://github.com/CSSEGISandData/COVID-19`.

## Ζητήματα εργασίας

Για όλα τα ζητήματα στην αρχή του κάθε προγράμματος θα φορτώνεται το σχετικό αρχείο δεδομένων. Υπάρχει ελεύθερη επιλογή στην παρουσίαση των αποτελεσμάτων (γραφήματα, συμπεράσματα και αποτελέσματα στη γραμμή εντολών). Για κάποιες χώρες και μέρες μπορεί να υπάρχουν αρνητικές τιμές ή να μην υπάρχουν στοιχεία ημερήσιων κρουσμάτων ή θανάτων (κενά κελιά στο αρχείο δεδομένων και NaN στους αντίστοιχους πίνακες στο Matlab). Μπορείτε να επιλέξετε να τα αναπληρώσετε από στοιχεία που θα βρείτε σε άλλες πηγές ή να τα αφαιρέσετε (αν αφαιρέσετε αρνητική τιμή ή κενό για μια μεταβλητή π.χ. ημερήσια κρούσματα, σε κάποια μέρα θα πρέπει να αφαιρέσετε την τιμή στην αντίστοιχη ημέρα και για την άλλη μεταβλητή, π.χ. για τους ημερήσιους θανάτους). Μπορείτε επίσης να διορθώσετε τις τιμές με κάποιο αιτιολογημένο τρόπο. Για παράδειγμα για την Ισπανία, από μια ημερομηνία και μετά δίνονται μηδενικές τιμές ημερήσιων κρουσμάτων για το Σαββατοκύριακο και υψηλότερη τιμή τη Δευτέρα. Μπορείτε να επιλέξετε να μοιράσετε με κάποιο (εύλογο ;) τρόπο την τιμή της Δευτέρας στις προηγούμενες δύο μέρες.

1. Θεώρησε πως οι τιμές ημερήσιων κρουσμάτων στις μέρες του πρώτου κύματος αντιστοιχούν σε συχνότητες εμφάνισης νέων κρουσμάτων για κάθε μέρα, δηλαδή το γράφημα των τιμών είναι το ραβδόγραμμα (καταχρηστικά ιστόγραμμα), όπου η τ.μ. είναι η μέρα του πρώτου κύματος. Με αυτήν την θεώρηση, βρες την παραμετρική κατανομή πιθανότητας

με την καλύτερη προσαρμογή στα δεδομένα ημερήσιων κρουσμάτων του πρώτου κύματος για τη χώρα Α. Η επιλογή θα γίνει από 5 τουλάχιστον διαφορετικές υποψήφιες παραμετρικές κατανομές με ελεύθερη επιλογή. Για να δοκιμάσεις διαφορετικές παραμετρικές κατανομές θα πρέπει να έχεις την εργαλειοθήκη Statistics and Machine Learning Toolbox και για να δεις τη λίστα των παραμετρικών κατανομών, δες π.χ. τη βοήθεια για την εντολή `fitdist`. Η καλή προσαρμογή κάποιας κατανομής πιθανότητας μπορεί να γίνει με έλεγχο καλής προσαρμογής $X^2$ (χρησιμοποιώντας ως κριτήριο την $p$-τιμή του ελέγχου). Λόγω των μεγάλων τιμών συχνοτήτων, ο έλεγχος καλής προσαρμογής $X^2$ ενδέχεται να δίνει πολύ χαμηλές $p$-τιμές. Για αυτό εναλλακτικά μπορείς να χρησιμοποιήσεις κάποιο στατιστικό σφάλματος προσαρμογής, π.χ. το μέσο τετραγωνικό σφάλμα, και με βάση αυτό να επιλέξεις την παραμετρική κατανομή πιθανότητας με την καλύτερη προσαρμογή. Κάνε το ίδιο για τους ημερήσιους θανάτους. Είναι ίδια η πιο κατάλληλη κατανομή για ημερήσια κρούσματα και θανάτους; Αν όχι, θα μπορούσε η κατανομή που επιλέχτηκε για τα ημερήσια κρούσματα να είναι κατάλληλη και για τους ημερήσιους θανάτους, και το αντίθετο (όπου εφαρμόζεται);

2. Επίλεξε έναν ικανό αριθμό άλλων Ευρωπαϊκών χωρών ($\geq 10$) και προσάρμοσε την παραμετρική κατανομή πιθανότητας που βρέθηκε καλύτερη στο Ζήτημα 1, ξεχωριστά για ημερήσια κρούσματα και ημερήσιους θανάτους. Η χρονική περίοδος του πρώτου κύματος θα πρέπει να προσδιοριστεί ξεχωριστά για κάθε χώρα (και για κρούσματα και θανάτους). Κατάταξε τις χώρες ως προς την καλή προσαρμογή της παραμετρικής κατανομής στα δεδομένα της κάθε χώρας (ξεχωριστά για κρούσματα και θανάτους). Φαίνεται η παραμετρική κατανομή πιθανότητας που επιλέχτηκε στο Ζήτημα 1 να προσαρμόζεται γενικά καλά στα δεδομένα των Ευρωπαϊκών χωρών; Απάντησε ξεχωριστά για ημερήσια κρούσματα και ημερήσιους θανάτους.

3. Στο ζήτημα αυτό θέλουμε να εκτιμήσουμε το χρονικό διάστημα από την κορύφωση των ημερήσιων κρουσμάτων ως την κορύφωση των ημερήσιων θανάτων (θεωρώντας γενικά πως η κορύφωση θανάτων έπεται της κορύφωσης κρουσμάτων). Σε συνέχεια των Ζητημάτων 1 και 2, πρώτα κάνε μια συνάρτηση που θα εκτιμά την κορύφωση του πρώτου κύματος ως προς τα ημερήσια κρούσματα (την ίδια συνάρτηση θα καλέσεις και για τους θανάτους) για δεδομένη χώρα και με βάση τη μέγιστη τιμή της προσαρμοσμένης πιο κατάλληλης κατανομής πιθανότητας. Δηλαδή η συνάρτηση θα κάνει ότι έκανες στο πρόγραμμα στο Ζήτημα 1 και επιπλέον θα βρίσκει και την ημερομηνία της κορύφωσης. Στη συνέχεια θεώρησε το σύνολο των χωρών στο Ζήτημα 2, δηλαδή τη χώρα Α και τις άλλες ($\geq 10$) Ευρωπαϊκές χώρες. Κάνε ένα πρόγραμμα που θα καλεί τη συνάρτηση για κάθε μια από τις χώρες, τη μια φορά για τα ημερήσια κρούσματα και την άλλη φορά για τους ημερήσιους θανάτους, και θα υπολογίζει τις μέρες μεταξύ κορύφωσης κρουσμάτων και θανάτων. Με βάση το δείγμα των χρονικών υστερήσεων για το πλήθος των χωρών που θεώρησες θα υπολογίσεις 95% παραμετρικό και bootstrap διάστημα εμπιστοσύνης για το μέσο χρόνο υστέρησης και θα ελέγξεις σε αυτό το επίπεδο εμπιστοσύνης αν ο χρόνος μεταξύ κορύφωσης ημερήσιων κρουσμάτων και θανάτων μπορεί να είναι 14 μέρες.

4. Στο ζήτημα αυτό θα υποθέσουμε πως η πορεία του κύματος είναι παρόμοια για τα ημερήσια κρούσματα και τους ημερήσιους θανάτους, αλλά αναμένουμε η πορεία των ημερήσιων

θανάτων να έπεται αυτής των ημερήσιων κρουσμάτων με κάποια χρονική υστέρηση κάποιων ημερών (όπως για την κορύφωση στο Ζήτημα 3. Για αυτό θέλουμε να βρούμε την υστέρηση που οι δύο πορείες των ημερήσιων κρουσμάτων και θανάτων κατά τη διάρκεια του πρώτου κύματος συσχετίζονται περισσότερο. Θα επιλέξεις τη χώρα Α και 5 άλλες χώρες (από αυτές που χρησιμοποίησες στα Ζητήματα 2 και 3) και θα κάνεις ένα πρόγραμμα που για κάθε χώρα θα βρίσκει εκείνη την υστέρηση $\tau$ για την οποία ο συντελεστής συσχέτισης Pearson μεταξύ ημερήσιων κρουσμάτων και θανάτων έχει τη μέγιστη τιμή, για τιμές υστέρησης σε ένα χρονικό παράθυρο, π.χ. [-20,20]. Πιο συγκεκριμένα, αν $x(t)$ και $y(t)$ είναι η τιμή νέων ημερήσιων κρουσμάτων και θανάτων την ημερομηνία $t$ αντίστοιχα, θέλουμε να βρούμε την υστέρηση $\tau$ που δίνει τη μέγιστη τιμή στον συντελεστή συσχέτισης Pearson $r(x(t), y(t+\tau))$, όπου το δείγμα αποτελείται από τις ζευγαρωτές τιμές των $x(t)$ και $y(t+\tau))$ στις μέρες του πρώτου κύματος. Το πρόγραμμα θα βρίσκει την υστέρηση μέγιστης συσχέτισης για κάθε χώρα και θα παρουσιάζονται τα αποτελέσματα (ελεύθερη χρήση μορφής αποτελεσμάτων στο παράθυρο εντολών και σχημάτων). Φαίνεται αυτή η προσέγγιση να εκτιμά σωστά την υστέρηση της πορείας των ημερήσιων θανάτων ως προς την πορεία των ημερήσιων κρουσμάτων; Συμφωνεί η εκτίμηση αυτής της υστέρησης με αυτήν στο Ζήτημα 3; Σχολίασε για τις τυχόν δυσκολίες και προβληματισμούς έχεις.

5. Στο ζήτημα αυτό θέλουμε να διερευνήσουμε τη δυνατότητα πρόβλεψης των ημερήσιων θανάτων από τα ημερήσια κρούσματα στο πρώτο κύμα μιας χώρας. Συγκεκριμένα θέλουμε να συγκρίνουμε μοντέλα απλής γραμμικής παλινδρόμησης των νέων θανάτων σε μια μέρα από τα νέα κρούσματα την ίδια μέρα, ή μιας μέρας πριν, δύο ημερών πριν κτλ. Θα κάνεις ένα πρόγραμμα που θα υλοποιεί αυτήν τη διερεύνηση για τις ίδιες χώρες που χρησιμοποίησες στο Ζήτημα 4. Για κάθε υστέρηση $\tau$ σε ένα χρονικό παράθυρο, π.χ. [0,20], θα προσαρμόσεις το μοντέλο απλής γραμμικής παλινδρόμησης της τ.μ. νέων θανάτων $y(t)$ την ημερομηνία $t$ από τη μεταβλητή των κρουσμάτων $x(t-\tau)$ την ημερομηνία $t-\tau$. Θα συγκρίνεις την καλή προσαρμογή των μοντέλων για διαφορετικά $\tau$ και θα βρεις την υστέρηση $\tau$ που δίνει την καλύτερη προσαρμογή. Θα κάνεις επίσης διαγνωστικό έλεγχο (γράφημα τυποποιημένων σφαλμάτων) για κάθε υστέρηση. Το πρόγραμμα θα κάνει τα παραπάνω για κάθε χώρα και θα παρουσιάζει τα αποτελέσματα (ελεύθερη χρήση μορφής αποτελεσμάτων στο παράθυρο εντολών και σχημάτων). Φαίνεται η προσαρμογή να είναι το ίδιο καλή στις χώρες που δοκίμασες; Φαίνεται ο διαγνωστικός έλεγχος να υποδεικνύει καταλληλότητα του μοντέλου για τις διαφορετικές υστερήσεις και χώρες; Φαίνεται να είναι δυνατή η πρόβλεψη ημερήσιων θανάτων από τα ημερήσια κρούσματα; Συμφωνούν οι βέλτιστες υστερήσεις από την προσαρμογή μοντέλου με τις βέλτιστες υστερήσεις από τη συσχέτιση στο Ζήτημα 4 και τη διαφορά κορυφώσεων στο Ζήτημα3 για τις χώρες που δοκίμασες; Σχολίασε για τις τυχόν δυσκολίες και προβληματισμούς έχεις.

6. Σε συνέχεια του Ζητήματος 5 θέλουμε να διερευνήσουμε αν η πρόβλεψη των ημερήσιων θανάτων στο πρώτο κύμα βελτιώνεται αν αντί του μοντέλου απλής γραμμικής παλινδρόμησης ως προς τα ημερήσια κρούσματα της ίδιας η κάποιας προηγούμενης ημέρας (όπως έκανες στο Ζήτημα 5), χρησιμοποιήσουμε μοντέλο πολλαπλής γραμμικής παλινδρόμησης ημερήσιων κρουσμάτων σε περισσότερες από μια υστερήσεις. Για τις ίδιες χώρες και τον ίδιο σχεδιασμό όπως στο Ζήτημα 5, θα βρεις το βέλτιστο μοντέλο με ανεξάρτητες μεταβλητές τα ημερήσια κρούσματα για υστερήσεις $\tau$ στο ίδιο χρονικό παράθυρο που

χρησιμοποίησες στο Ζήτημα 5, π.χ. [0,20], δηλαδή οι υποψήφιες ανεξάρτητες μεταβλητές για το μοντέλο στο παράδειγμα είναι $x(t), x(t-1), \ldots, x(t-20)$. Μπορείς να επιλέξεις μια από τις προσεγγίσεις που κάναμε στο μάθημα για μείωση διάστασης σε προβλήματα παλινδρόμησης. Θα συγκρίνεις αυτό το μοντέλο με το βέλτιστο μοντέλο απλής γραμμικής παλινδρόμησης που βρήκες στο Ζήτημα 5 (αν δεν είναι το ίδιο), καθώς και με το πλήρες μοντέλο όλων των μεταβλητών υστέρησης, δηλαδή το μοντέλο με όλες τις 21 μεταβλητές υστέρησης στο παραπάνω παράδειγμα. Θα κάνεις επίσης διαγνωστικό έλεγχο (γράφημα τυποποιημένων σφαλμάτων) για κάθε ένα από τα τρία μοντέλα (μοντέλο μιας μεταβλητής, όλων των μεταβλητών, και επιλεγμένων μεταβλητών ή μείωσης διάστασης). Το πρόγραμμα που θα αναπτύξεις θα το κάνει αυτό για κάθε μια από τις χώρες που χρησιμοποίησες στο Ζήτημα 5 και θα παρουσιάζει τα αποτελέσματα (ελεύθερη χρήση μορφής αποτελεσμάτων στο παράθυρο εντολών και σχημάτων). Φαίνεται η προσαρμογή του βέλτιστου μοντέλου να είναι το ίδιο καλή στις χώρες που δοκίμασες; Φαίνεται να βελτιώνεται η πρόβλεψη ημερήσιων θανάτων όταν χρησιμοποιείται μοντέλο πολλαπλής γραμμικής παλινδρόμησης σε σχέση με την πρόβλεψη με μοντέλο απλής γραμμικής παλινδρόμησης; Σχολίασε για τις τυχόν δυσκολίες και προβληματισμούς έχεις.

7. Σε συνέχεια του Ζητήματος 6, θέλουμε να χρησιμοποιήσουμε το καλύτερο μοντέλο που βρήκαμε στο Ζήτημα 6 για την πρόβλεψη των ημερήσιων θανάτων από τα ημερήσια κρούσματα (την ίδια ή/και άλλες προηγούμενες ημέρες στο πρώτο κύμα) για να κάνουμε προβλέψεις σε χρονικές περιόδους μετά το πρώτο κύμα. Η περίοδος του πρώτου κύματος αποτελεί το σύνολο δεδομένων εκμάθησης (εκπαίδευσης) και η δεύτερη χρονική περίοδος αποτελεί το σύνολο αξιολόγησης. Θα ορίσετε τη δεύτερη χρονική περίοδο για κάθε χώρα να καλύπτει όσο το δυνατόν καλύτερα το δεύτερο κύμα, όπου αυτό εφαρμόζεται (αν υπάρχει κορύφωση του δεύτερου κύματος να την καλύπτει). Το πρόγραμμα θα χρησιμοποιεί το μοντέλο που βρέθηκε ως καλύτερο στο Ζήτημα 6 και θα κάνει προβλέψεις στο σύνολο αξιολόγησης. Θα συγκρίνει με κάποιο στατιστικό (π.χ. προσαρμοσμένος συντελεστής προσδιορισμού) τα σφάλματα στο σύνολο εκμάθησης και στο σύνολο αξιολόγησης. Αυτό θα επαναλαμβάνεται για κάθε μια από τις χώρες που χρησιμοποίησες στα Ζητήματα 5 και 6. Για κάποιες χώρες το πεδίο τιμών των ανεξάρτητων μεταβλητών (ημερήσιων κρουσμάτων) και εξαρτημένης μεταβλητής (ημερήσιοι θάνατοι) διαφέρει στις δύο περιόδους (στο σύνολο εκμάθησης και αξιολόγησης). Για παράδειγμα για κάποια χώρα τα ημερήσια κρούσματα μπορεί να είναι σε πολύ υψηλότερο επίπεδο στη δεύτερη περίοδο. Μπορείς να το αντιμετωπίσεις με κάποιο τρόπο ώστε οι προβλέψεις στο σύνολο αξιολόγησης να είναι συγκρίσιμες με αυτές στο σύνολο εκμάθησης;

8. Επανέλαβε την ίδια διαδικασία όπως στο Ζήτημα 7 αλλά χρησιμοποίησε ένα άλλο μοντέλο μείωσης διάστασης. Σύγκρινε τα δύο μοντέλα (αυτό που χρησιμοποίησες στα Ζητήματα 6 και 7 και το νέο μοντέλο) με βάση τα σφάλματα προσαρμογής στο πρώτο κύμα (σύνολο εκμάθησης) και τα σφάλματα πρόβλεψης στο δεύτερο κύμα (σύνολο ελέγχου). Η σύγκριση θα γίνει για κάθε μια από τις χώρες που χρησιμοποίησες στα Ζητήματα 5, 6 και 7. Και πάλι θα πρέπει να αντιμετωπίσεις τη διαφορά των πεδίων τιμών ημερήσιων κρουσμάτων και θανάτων στα δύο κύματα, για όποια χώρα προκύπτει.