Εργασία στο μάθημα 'Ανάλυση Δεδομένων', Δεκέμβριος 2024

Δημήτρης Κουγιουμτζής

E-mail: dkugiu@auth.gr

11 Δεκεμβρίου 2024

Οδηγίες: Σχετικά με την παράδοση της εργασίας:

- 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 το ζήτημα). Αντίστοιχα για τις συναρτήσεις τα ονόματα των αρχείων θα είναι Group10Exe5Fun1.m, Group10Exe5Fun2.m κτλ (αν χρειάζεται να δημιουργήσετε περισσότερες από μια συναρτήσεις). 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 (τα προγράμματα θα φορτώνουν το αρχείο από τον ίδιο φάκελο). 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 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.).

- **Μπορεί ο διδάσκων να ζητήσει μια ομάδα να παρουσιάσει και συζητήσει για προγράμματα που έχει υποβάλει. Αυτό θα γίνει την επόμενη της τελευταίας ημέρας υποβολής στις 12:00 - 15:00 απομακρυσμένα. Το πρωί της ίδιας μέρας θα σταλεί email στα μέλη της ομάδας με τον σύνδεσμο zoom και την ακριβή ώρα σύνδεσης και συζήτησης. Αν κάποιο μέλος της ομάδας δεν είναι διαθέσιμο (παρόν) θα μετρήσει αρνητικά στη βαθμολογία της εργασίας (ως και μηδενισμό).**

Περιγραφή εργασίας

Στη μελέτη επιληψίας έχει χρησιμοποιηθεί τα τελευταία έτη ο διακρανιακός μαγνητικός ερεθισμός (transcranial magnetic stimulation, TMS). Το TMS είναι μη-επεμβατική μέθοδος.

Με κάποιο μηχανισμό υψηλής τεχνολογίας δίνεται εξωτερικά μαγνητικό ερέθισμα που διαπερνά το κρανίο του ατόμου και επιδρά στο δυναμικό στην περιοχή του εγκεφάλου στη γειτονιά του ερεθίσματος. Πιστεύεται πως με κατάλληλο πειραματικό σχεδιασμό, που περιλαμβάνει μεταξύ άλλων σχήμα πηνίου, ένταση, συχνότητα και πλήθος ερεθισμάτων, η χορήγηση TMS στην αρχή επιληπτικών εκφορτίσεων (epileptic discharges, ED) μπορεί να μειώσει τη διάρκεια τους ή και να τις τερματίσει. Επίσης ίσως είναι σημαντικό το TMS να χορηγηθεί στην αρχή των ED, δηλαδή ο χρόνος preTMS από την έναρξη του ED ως τη χορήγηση του TMS να είναι μικρός, και τότε πιθανόν να μειώνεται ο χρόνος postTMS, δηλαδή ο χρόνος από τη χορήγηση του TMS ως το τέλος του ED.

Για να διερευνηθούν οι παραπάνω υποθέσεις έγιναν πειράματα στο Εργαστήριο Διακρανιακού Μαγνητικού Ερεθισμού, Γ Νευρολογική Κλινική, Ιατρική Σχολή ΑΠΘ. Με χρήση εξοπλισμού υψηλής τεχνολογίας που επιτρέπει το συνδυασμό του TMS με χρήση ηλεκτροεγκεφαλογραφήματος (EEG), μετρήθηκε η διάρκεια ED (EDduration) σε επιληπτικούς ασθενείς χωρίς τη χορήγηση TMS (135 μετρήσεις) καθώς και με τη χορήγηση TMS (119 μετρήσεις). Στην περίπτωση που χορηγήθηκε TMS κατά τη διάρκεια του ED, μετρήθηκε και ο χρόνος από την αρχή του ED ως τη στιγμή χορήγησης TMS (preTMS) καθώς και ο χρόνος από τη στιγμή χορήγησης του TMS ως το πέρας του ED (postTMS). Οι μετρήσεις έγιναν σε 6 διαφορετικές καταστάσεις μέτρησης (Setup) (διαφορετική περίοδος ή/και διαφορετικός ασθενής) και για τους δύο τύπους μετρήσεων (διάρκεια ED με ή χωρίς TMS). Καταγράφηκαν επίσης κάποιες παράμετροι του πειραματικού σχεδιασμού. Όλα τα δεδομένα δίνονται στο αρχείο `TMS.xlsx` (μορφή δεδομένων excel) που βρίσκεται στην ιστοσελίδα του μαθήματος, όπου η πρώτη γραμμή δηλώνει το όνομα του μεγέθους (μεταβλητή) και η κάθε μια από τις υπόλοιπες 254 γραμμές αφορά μια περίπτωση εμφάνισης ED με ή χωρίς TMS. Η οργάνωση των δεδομένων δίνεται παρακάτω:

Α/Α	*Όνομα*	*Περιγραφή*
1	TMS	1: με TMS, 0: χωρίς TMS
2	EDduration	διάρκεια ED (σε δευτερόλεπτα)
3	preTMS	χρόνος από την αρχή του ED ως τη στιγμή χορήγησης TMS (σε δευτερόλεπτα)
4	postTMS	χρόνος από τη στιγμή χορήγησης του TMS ως το πέρας του ED (σε δευτερόλεπτα)
5	Setup	κωδικός κατάστασης μέτρησης 1 ως 6 (διαφορετική περίοδος ή/και διαφορετικός ασθενής)
6	Stimuli	πλήθος ερεθισμάτων που συνιστούν μια χορήγηση TMS
7	Intensity	ένταση ερεθισμού (ως ποσοστό του μέγιστου ερεθισμού)
8	Spike	-1: το πρώτο ερέθισμα δόθηκε κατά την περίοδο φόρτισης του ED, 0: το πρώτο ερέθισμα δόθηκε στην κορυφή φόρτισης του ED, 1: το πρώτο ερέθισμα δόθηκε κατά την περίοδο εκφόρτισης του ED
9	Frequency	συχνότητα ερεθισμάτων που συνιστούν μια χορήγηση TMS
10	CoilCode	κωδικός πηνίου, 1: σχήμα οκτάρι, 0: στρογγυλό

Ζητήματα εργασίας

1. Θεωρώντας το σύνολο των δεδομένων, βρείτε την κατάλληλη γνωστή (παραμετρική) κατανομή πιθανότητας που προσαρμόζεται καλύτερα στα δεδομένα για τη διάρκεια ED (EDduration) με TMS. Επανέλαβε το ίδιο για τη διάρκεια ED χωρίς TMS. Στη συνέχεια φτιάξε ένα σχήμα που να περιέχει:

   - την καμπύλη της εμπειρικής συνάρτησης πυκνότητας πιθανότητας (σππ) με τη μέθοδο του ιστογράμματος για κατάλληλη ισομερή διαμέριση για τη διάρκεια ED χωρίς TMS.
   - την καμπύλη της εμπειρικής σππ με τη μέθοδο του ιστογράμματος για την ίδια διαμέριση όπως παραπάνω για τη διάρκεια ED με TMS.
   - την καμπύλη της κατανομής που βρέθηκε ως πιο κατάλληλη για τη διάρκεια ED χωρίς TMS.
   - την καμπύλη της κατανομής που βρέθηκε ως πιο κατάλληλη για τη διάρκεια ED με TMS.

   Φαίνεται η σππ για τη διάρκεια ED να είναι ίδια με ή χωρίς TMS;

[Βοήθεια: Για λίστα κατανομών δες συνάρτηση fitdist. Για τη σύγκριση κατανομών, η καλή προσαρμογή μπορεί να αξιολογηθεί από την τιμή του αντίστοιχου στατιστικού ελέγχου X^2].

2. Θεωρείστε ως δύο διαφορετικά δείγματα τα δεδομένα διάρκειας του ED (EDduration) με TMS και πηνίο σε σχήμα οκτάρι και σε σχήμα στρογγυλό (οι δύο τιμές της μεταβλητής CoilCode). Για κάθε ένα από τα δύο δείγματα θα κάνετε έλεγχο καταλληλότητας κατανομής για την εκθετική κατανομή με επαναδειγματοληψία (resampling). Συγκεκριμένα θα δημιουργήσετε 1000 τυχαία δείγματα από την εκθετική κατανομή που ελέγχετε με παράμετρο κατανομής αυτήν που θα εκτιμήσετε από το αρχικό δείγμα. Στην συνέχεια θα θεωρήσετε έλεγχο $\mathcal{X}^2$ για εκθετική κατανομή για κάθε ένα από τα 1000 τυχαία δείγματα και θα υπολογίσετε 1000 τιμές του χ^2 στατιστικού. Θα υπολογίσετε επίσης το χ^2 στατιστικό για το αρχικό δείγμα, ${\chi_0}^2$ (για τον ίδιο έλεγχο υπόθεσης εκθετικής κατανομής). Θα εξετάσετε αν το ${\chi_0}^2$ είναι στη δεξιά ουρά της εμπειρικής κατανομής που σχηματίζεται από τις 1000 τιμές χ^2 για τα 1000 τυχαία δείγματα (μονόπλευρος έλεγχος) για να καταλήξετε στην απόφαση του ελέγχου. Συγκρίνετε τον έλεγχο επαναδειγματοληψίας με τον παραμετρικό έλεγχο $\mathcal{X}^2$ για εκθετική κατανομή. Διαφέρουν τα αποτελέσματα των δύο ελέγχων στα δύο δείγματα (πηνίο οκτάρι και πηνίο στρογγυλό);

3. Υπολόγισε τη μέση τιμή της διάρκειας του ED (EDduration) στο σύνολο των δεδομένων διάρκειας ED χωρίς TMS και έστω αυτή είναι μ_0. Θεώρησε τα 6 δείγματα διάρκειας του ED χωρίς TMS για κάθε μια από τις 6 καταστάσεις μέτρησης (μεταβλητή Setup). Χρησιμοποιώντας είτε διάστημα εμπιστοσύνης ή έλεγχο υπόθεσης, έλεγξε αν μπορούμε να δεχτούμε ότι η μέση διάρκεια ED χωρίς TMS για την κάθε μια από τις 6 καταστάσεις μέτρησης είναι μ_0. Θα πρέπει να χρησιμοποιήσεις παραμετρικό διάστημα εμπιστοσύνης (ή έλεγχο υπόθεσης αν επιλέξεις αυτό) αν η κατανομή μπορεί να θεωρηθεί κανονική (θα χρειαστεί έλεγχος καταλληλότητας $\mathcal{X}^2$ για κανονική κατανομή) ή bootstrap διάστημα εμπιστοσύνης (και αντίστοιχο έλεγχο υπόθεσης αν επιλέξεις αυτό, δες Κεφ.3, Άσκηση 10) αν η κατανομή δε μπορεί να θεωρηθεί κανονική. Επανέλαβε την παραπάνω ανάλυση για την περίπτωση που χορηγείται TMS. Συγκέντρωσε τα αποτελέσματα για τις 6 καταστάσεις μέτρησης και για τις δύο περιπτώσεις (με και χωρίς TMS) σε έναν πίνακα και σχολίασε αν συμφωνούν τα αποτελέσματα για τις δύο περιπτώσεις.

4. Για κάθε μια από τις 6 καταστάσεις μέτρησης (μεταβλητή Setup), θέλουμε να διερευνήσουμε αν υπάρχει συσχέτιση μεταξύ του χρόνου από την αρχή του ED ως τη στιγμή χορήγησης TMS (preTMS) και του χρόνου από τη στιγμή χορήγησης του TMS ως το πέρας του ED (postTMS). Κάνε κατάλληλο παραμετρικό έλεγχο για μηδενική συσχέτιση των δεικτών preTMS και postTMS, χρησιμοποιώντας το στατιστικό της κατανομής Student. Κάνε επίσης τον αντίστοιχο έλεγχο τυχαιοποίησης χρησιμοποιώντας 1000 τυχαιοποιημένα δείγματα (δες Κεφ.5, Άσκηση 2). Επανέλαβε τα παραπάνω για κάθε μια από τις 6 καταστάσεις μέτρησης και συγκέντρωσε τα αποτελέσματα ώστε να παρουσιάζονται σε μορφή πίνακα. Φαίνεται να υπάρχει κάποια συσχέτιση μεταξύ των preTMS και postTMS με βάση τον παραμετρικό και έλεγχο τυχαιοποίησης και για ποια κατάσταση μέτρησης; Ποιον έλεγχο εμπιστεύεσαι περισσότερο και γιατί για κάθε κατάσταση μέτρησης (ή για όλες συνολικά);

5. Θέλουμε να διερευνήσουμε αν η διάρκεια ED όταν δε χορηγείται TMS εξαρτιέται από την κατάσταση μέτρησης (Setup). Εκτίμησε το μοντέλο γραμμικής παλινδρόμησης του EDduration ως προς το Setup, υπολόγισε το συντελεστή προσδιορισμού (ή εναλλακτικά τον προσαρμοσμένο συντελεστή προσδιορισμού) και διερεύνησε την καταλληλότητα του μοντέλου (αν τα υπόλοιπα είναι ασυσχέτιστα). Επανέλαβε την παραπάνω ανάλυση όταν χορηγείται TMS. Διαφέρουν τα αποτελέσματα; Σε ποια περίπτωση η προσαρμογή του μοντέλου παλινδρόμησης είναι καλύτερη; Θα ήταν χρήσιμο να επεκτείνουμε το μοντέλο παλινδρόμησης σε πολυωνυμικό κάποιου βαθμού σε κάθε μια από τις δύο περιπτώσεις (με ή χωρίς TMS);

6. Για την περίπτωση που χορηγείται TMS, διερεύνησε το κατάλληλο μοντέλο πολλαπλής γραμμικής παλινδρόμησης για τη διάρκεια ED. Δοκίμασε το μοντέλο με τις ανεξάρτητες μεταβλητές Setup, Stimuli, Intensity, Spike, Frequency και CoilCode. Σύγκρινε το πλήρες μοντέλο (με τις 6 ανεξάρτητες μεταβλητές) με δύο μοντέλα επιλεγμένων μεταβλητών, α) αυτό που δίνει η μέθοδος βηματικής παλινδρόμησης, και β) αυτό που δίνει η μέθοδος LASSO. Για την LASSO θα επιλέξετε κατάλληλη τιμή της παραμέτρου λ που ορίζει και τις επιλεγμένες μεταβλητές. Υπολόγισε για το κάθε μοντέλο τη διασπορά των σφαλμάτων (μέσο τετραγωνικό σφάλμα) και τον προσαρμοσμένο συντελεστή προσδιορισμού. Ποιο μοντέλο προσαρμόζεται καλύτερα;

   Η μεταβλητή Spike έχει πολλά κενά (missing values). Αγνόησε αυτήν την μεταβλητή και επανέλαβε την παραπάνω διαδικασία. Αλλάζουν τα αποτελέσματα;

7. Σε συνέχεια του ζητήματος 6, θα θεωρήσεις τα ίδια τρία μοντέλα πολλαπλής παλινδρόμησης της διάρκειας ED όταν χορηγείται TMS (για τα μοντέλα α) και β) θα χρησιμοποιήσεις τις μεταβλητές που επιλέχτηκαν στο σύνολο των δεδομένων), αλλά τώρα θα χωρίσεις το σύνολο δεδομένων σε ένα τυχαίο επιλεγμένο σύνολο εκμάθησης, όπου θα προσαρμόσεις τα μοντέλα, και θα υπολογίσεις για κάθε ένα από τα τρία μοντέλα το μέσο τετραγωνικό σφάλμα στο υπόλοιπο σύνολο δεδομένων, το σύνολο ελέγχου (ελεύθερη επιλογή του μεγέθους του συνόλου εκμάθησης). Ποιο μοντέλο προβλέπει καλύτερα; Στη συνέχεια, για το μοντέλα της βηματικής παλινδρόμησης και LASSO θα επαναλάβεις την ίδια διαδικασία αλλά τώρα η επιλογή των μεταβλητών θα γίνει στο σύνολο εκμάθησης. Διαφέρουν τα αποτελέσματα; [Μπορείς να αποφασίσεις πως θα διαχειριστείς την μεταβλητή Spike (αν θα την συμπεριλάβεις ή όχι), σύμφωνα και με τα αποτελέσματα στο ζήτημα 6.]

8. Επανέλαβε την ανάλυση στο ζήτημα 6 για μοντέλο πολλαπλής γραμμικής παλινδρόμησης για τη διάρκεια ED, θεωρώντας εκτός από τις 6 ανεξάρτητες μεταβλητές και τη μεταβλητή preTMS. Σύγκρινε τα αποτελέσματα του πλήρους μοντέλου (με τις 7 ανεξάρτητες μεταβλητές) με τα δύο μοντέλα επιλογής μεταβλητών (βηματική παλινδρόμηση και LASSO στις 7 ανεξάρτητες μεταβλητές). Βελτιώθηκε η προσαρμογή των τριών μοντέλων σε σχέση με αυτά του ζητήματος 5; Στη σύγκριση των μοντέλων συμπεριλάβετε και το μοντέλο παλινδρόμησης κυρίων συνιστωσών (principal component regression, PCR) στις 7 ανεξάρτητες μεταβλητές. Θα πρέπει να προσδιορίσετε το πλήθος των κυρίων συνιστωσών με κάποιο κριτήριο. Σχολιάστε για την επίδοση του μοντέλου PCR σε σχέση με τα άλλα τρία μοντέλα. Αλλάζουν τα τρία μοντέλα μείωσης διάστασης (βηματική παλινδρόμηση, LASSO και PCR) όταν προσθέτουμε και τη μεταβλητή postTMS στις ήδη 7 υπάρχουσες ανεξάρτητες μεταβλητές (δηλαδή πριν την επιλογή μεταβλητών); [Μπορείς να αποφασίσεις πως

θα διαχειριστείς την μεταβλητή Spike (αν θα την συμπεριλάβεις ή όχι), σύμφωνα και με τα αποτελέσματα στο ζήτημα 6.]