Ε Θ Ν Ι Κ Ο Μ Ε Τ Σ Ο Β Ι Ο Π Ο Λ Υ Τ Ε Χ Ν Ε Ι Ο
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
Τομέας Μαθηματικών
Πολυτεχνειούπολη – Ζωγράφου ΑΘΗΝΑ - 157 80
ΤΗΛ. : 772 1774
FAX : 772 1775

(α) ΔΠΜΣ ΣΤΙΣ ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ
(β) ΔΠΜΣ ΣΤΗΝ ΜΑΘΗΜΑΤΙΚΗ ΠΡΟΤΥΠΟΠΟΙΗΣΗ ΣΕ ΣΥΓΧΡΟΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΚΑΙ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ
(γ) ΔΠΜΣ ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ
(δ) 8ο ΕΞΑΜΗΝΟ ΤΟΥ ΠΡΟΠΤΥΧΙΑΚΟΥ ΠΡΟΓΡΑΜΜΑΤΟΣ ΣΠΟΥΔΩΝ ΤΗΣ ΣΕΜΦΕ

ΤΙΤΛΟΣ ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΜΑΘΗΜΑΤΟΣ: *ΥΠΟΛΟΓΙΣΤΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΣΤΟΧΑΣΤΙΚΗ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ*
ΤΙΤΛΟΣ ΠΡΟΠΤΥΧΙΑΚΟΥ ΜΑΘΗΜΑΤΟΣ: *ΥΠΟΛΟΓΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ*
ΔΙΔΑΣΚΩΝ: *ΔΗΜΗΤΡΗΣ ΦΟΥΣΚΑΚΗΣ (τηλ: 210 7721702 – email: fouskakis@math.ntua.gr)*

ΕΡΓΑΣΙΑ

1.

(α) Με χρήση δικού σας κώδικα στην R, προσομοιώστε 1000 τιμές από την τυποποιημένη κανονική κατανομή N(0, 1) με τη μέθοδο της απόρριψης. Χρησιμοποιήστε ως κατανομή εισήγησης την

$$g(x) = \frac{1}{\pi(1+x^2)},$$

δηλαδή την κατανομή Cauchy και βρείτε θεωρητικά τη βέλτιστη σταθερά M. Η προσομοίωση τιμών από την Cauchy να γίνει με τη μέθοδο της αντιστροφής. Εκτιμήστε την ολική πιθανότητα αποδοχής του αλγορίθμου σας και συγκρίνετέ την με την θεωρητική. Υπολογίστε (θεωρητικά) το μέσο αριθμό προσπαθειών που θα χρειαστεί κανείς για να έχει μια αποδοχή. Συγκρίνετε το ιστόγραμμα των προσομοιωμένων τιμών με το γράφημα της σ.π.π. της τυποποιημένης κανονικής κατανομής. Επιπλέον συγκρίνετε τον μέσο και την τυπική απόκλιση των προσομοιωμένων τιμών με τον μέσο και την τυπική απόκλιση της θεωρητικής κατανομής.

(β) Υπενθυμίζουμε ότι η αρνητική διωνυμική κατανομή NB(r, p), που ακολουθεί μία τ.μ. X, μπορεί να προκύψει ως η ακόλουθη μίξη κατανομών:

$$\Lambda \sim \text{Gamma}\left(r, \frac{1-p}{p}\right) \quad \text{και} \quad X \mid \Lambda = \lambda \sim \text{Poisson}(\lambda).$$

Ο αρχικός (κλασικός) εκτιμητής της μέσης τιμής της τ.μ. X είναι ο δειγματικός μέσος. Υπολογίστε τη Rao-Blackwellized εκδοχή του, με χρήση της παραπάνω μίξης, και ακολούθως τη διασπορά του (αναλυτικά) και συγκρίνετέ τη με τη θεωρητική διασπορά του αρχικού εκτιμητή. Γράψτε έναν κώδικα στην R που να επιβεβαιώνει τα θεωρητικά σας αποτελέσματα (n = 5000, r = 1 και p = 0.5).

(γ) Έστω $X_1, \ldots, X_n$ τυχαίο δείγμα και θεωρήστε τη στατιστική συνάρτηση

$$T = \frac{\left(X_1 + ... + X_n\right)^2}{n}.$$

Αν τα $X_i \sim U(0, 1)$ $(i = 1, \ldots, n)$ να προσομοιώσετε, με τη βοήθεια της R, 10000 τιμές από την T (n = 80) και ακολούθως να κατασκευάσετε το ιστόγραμμα των προσομοιωμένων τιμών και να υπολογίσετε τον μέσο και την τυπική τους απόκλιση. Συγκρίνετε το μέσο των προσομοιωμένων τιμών με τη θεωρητική μέση τιμή της τ.μ. T.

(δ) Στον παρακάτω σύνδεσμο

http://www.math.ntua.gr/~fouskakis/Computational_Stats/data1.rds

θα βρείτε τις παρατηρήσεις x_i $(i = 1,\ldots 80)$. Εκτιμήστε το τυπικό σφάλμα της στατιστικής συνάρτησης T, του παραπάνω ερωτήματος, μέσω της τεχνικής Bootstrap (B = 10000) με χρήση αποκλειστικά δικού σας κώδικα στην R. Εκτιμήστε το τυπικό σφάλμα της T χρησιμοποιώντας Jackknife και δικό σας κώδικα στην R. Αν κατόπιν μαθαίνατε ότι οι αρχικές παρατηρήσεις έχουν προέλθει από την κατανομή U(0, 1) να συγκρίνετε το ιστόγραμμα των προσομοιωμένων τιμών για τη στατιστική συνάρτηση T σε αυτό το ερώτημα με αυτό των προσομοιωμένων τιμών του προηγούμενου ερωτήματος.

2. Θεωρήστε τα "Old Faithful geyser data", τα οποία μας δίνουν 272 χρόνους (σε λεπτά) μεταξύ διαδοχικών εκρήξεων του ηφαιστείου "Old Faithful geyser - Yellowstone National Park, Wyoming, USA". Πληκτρολογώντας `faithful$eruptions` στην R παίρνετε τα δεδομένα.

(α) Έστω ότι θέλετε να εκτιμήσετε την σ.π.π. f(x) από όπου προέρχονται τα εν λόγω δεδομένα. Έστω ότι θέλετε να χρησιμοποιήσετε *Epanechnikov* πυρήνα. Βρείτε το βέλτιστο πλάτος h μεγιστοποιώντας την *cross-validated* πιθανοφάνεια, με χρήση δικού σας κώδικα στην R. Προβείτε σε ένα διάγραμμα της εκτιμώμενης f(x) για το h που βρήκατε, χρησιμοποιώντας την έτοιμη συνάρτηση `density` με *Epanechnikov* πυρήνα και σχολιάστε το αποτέλεσμα που πήρατε.

(β) Γράψτε την δική σας συνάρτηση για την εκτίμηση της σ.π.π. f(x) με χρήση του *Epanechnikov* πυρήνα με βάση το h που βρήκατε στο παραπάνω ερώτημα. Ως x θεωρήστε τις τιμές που λαμβάνει η συνάρτηση `density`. Η συνάρτησή σας θα πρέπει να επιστρέφει την εκτιμώμενη f για τα εν λόγω x. Προβείτε σε ένα διάγραμμα της εκτιμώμενης f(x) που λάβατε και σχολιάστε.

(γ) Με τη βοήθεια της συνάρτησης `integrate` εκτιμήστε, χρησιμοποιώντας τη συνάρτηση της προηγούμενης ερώτησης, την πιθανότητα ο χρόνος μεταξύ διαδοχικών εκρήξεων να είναι μεγαλύτερος των 3.5 λεπτών.

(δ) Προσομοιώστε 250 τιμές από την εκτιμώμενη f(x), με βάση τον *Epanechnikov* πυρήνα και το h που βρήκατε στο (α) ερώτημα, και εκτιμήστε εκ νέου την ζητούμενη πιθανότητα του ερωτήματος (γ).

3. Έστω ότι διαθέτετε δύο ανεξάρτητους πληθυσμούς (ομάδες) και τις εξής 4 παρατηρήσεις: x = (2, 7, 3, 9). Αρχικά θεωρήστε πως γνωρίζετε πως η $1^{η}$ και η $3^{η}$ παρατήρηση ανήκουν στην πρώτη ομάδα, ενώ η $2^{η}$ και η $4^{η}$ στη δεύτερη ομάδα. Έστω ότι η μεταβλητή ενδιαφέροντος κάτω από την πρώτη ομάδα ακολουθεί την κατανομή Poisson με άγνωστη παράμετρο $\lambda_1 > 0$ και κάτω από τη δεύτερη ομάδα την κατανομή Poisson με άγνωστη παράμετρο $\lambda_2 > 0$. Τέλος έστω π_1 η πιθανότητα μια παρατήρηση να ανήκει στην πρώτη ομάδα και $\pi_2 = (1 - \pi_1)$ η πιθανότητα να ανήκει στη δεύτερη ομάδα.

(α) Με βάση την παραπάνω πληροφορία εκτιμήστε με τη μέθοδο μέγιστης πιθανοφάνειας τις παραμέτρους λ_1, λ_2, π_1, π_2.

(β) Έστω τώρα πως δεν γνωρίζετε σε ποια από τις δύο ομάδες ανήκει κάθε παρατήρηση. Θεωρήστε τον αλγόριθμο EM για την εκτίμηση των αγνώστων παραμέτρων λ_1, λ_2, π_1, π_2. Αναπτύξτε (θεωρητικά) πλήρως τα βήματα του αλγορίθμου και εν συνεχεία δημιουργήστε μια δική σας συνάρτηση στην R που θα υλοποιεί τον αλγόριθμο. Ως κριτήριο τερματισμού, για δύο διαδοχικές επαναλήψεις (r) και (r+1), χρησιμοποιήστε το παρακάτω: $\left(\lambda_1^{(r+1)} - \lambda_1^{(r)}\right)^2 + \left(\lambda_2^{(r+1)} - \lambda_2^{(r)}\right)^2 \leq 10^{-10}$.

4. Στην βιβλιοθήκη `lars` της R θα βρείτε τα δεδομένα `diabetes`. Αρχικά κατεβάστε και φορτώστε τη βιβλιοθήκη και εν συνεχεία με χρήση της εντολής `data(diabetes)` φορτώστε τα δεδομένα. Τα δεδομένα έχουν πληροφορία για n = 442 ασθενείς με διαβήτη, σχετικά με ένα ποσοτικό χαρακτηριστικό `diabetes$y`, που μετρά την εξέλιξη της νόσου (μεταβλητή απόκρισης) και p = 10 επεξηγηματικές μεταβλητές (`diabetes$x`) οι οποίες είναι η ηλικία, το φύλο, ο δείκτης μάζας σώματος, η μέση αρτηριακή πίεση και έξι μετρήσεις ουρίας αίματος. Οι τιμές των επεξηγηματικών μεταβλητών έχουν τυποποιηθεί, ώστε να έχουν άθροισμα 0 και άθροισμα τετραγώνων ίσο με 1.

(α) Εξερευνώντας πλήρως τον χώρο όλων των πιθανών μοντέλων στο πρόβλημα επιλογής επεξηγηματικών μεταβλητών, με τη βοήθεια δικής σας συνάρτησης στην R, βρείτε το μοντέλο εκείνο που ελαχιστοποιεί την τιμή του κριτηρίου BIC. Καλέστε το εν λόγω μοντέλο M1.

(β) Εφαρμόστε τη μεθοδολογία *Lasso* με την βοήθεια της βιβλιοθήκης `glmnet` της R και σχολιάστε τα αποτελέσματα. Χρησιμοποιώντας *cross-validation* επιλέξτε την παράμετρο ποινής λ, με χρήση της έτοιμης συνάρτησης `cv.glmnet`, που ελαχιστοποιεί το CV-MSE. Χρησιμοποιώντας την εν λόγω τιμή καταλήξτε σε ένα μοντέλο, χωρίς κάποιες από τις επεξηγηματικές μεταβλητές, το οποίο καλέστε M2. Επαναλάβετε την ίδια διαδικασία επιλέγοντας ως λ την τιμή που ελαχιστοποιεί το CV-MSE με σφάλμα εντός μιας τυπικής απόκλισης από την ελάχιστη τιμή. Καλέστε το μοντέλο αυτό M3.

(γ) Χρησιμοποιώντας *5-fold cross-validation* και την (*within fold*) RMSE (*Root Mean Square Error*) συνάρτηση εξετάστε ποιο από τα τρία μοντέλα, M1, M2 και M3, έχει την καλύτερη προβλεπτική ικανότητα, με χρήση δικού σας κώδικα στην R. Για να βρείτε το σφάλμα στα μοντέλα που προέκυψαν από το *Lasso* θα χρησιμοποιήσετε κανονικά τους εκτιμητές των ελαχίστων τετραγώνων.

Οδηγίες

- Η εργασία θα πρέπει **να παραδοθεί ηλεκτρονικά** στο email μου, fouskakis@math.ntua.gr, **μέχρι την Παρασκευή 1 Ιουλίου 2022 στις 13:00μμ. Καμιά εργασία δεν θα γίνει δεκτή μετά την ώρα αυτή.**
- Η εργασία που θα παραδώσετε πρέπει να είναι **σε pdf μορφή αφού πρώτα τη γράψετε υποχρεωτικά σε Latex**. Ο κώδικας θα πρέπει **υποχρεωτικά να είναι σε R**.
- Η εργασία σας μπορεί να είναι είτε στα Ελληνικά είτε στα Αγγλικά.
- Παρακαλώ χρησιμοποιήστε τον **ακόλουθο τίτλο στο pdf αρχείο σας**: Surname-Name.pdf, όπου Surname είναι το επώνυμό σας (με λατινικούς χαρακτήρες) και Name το όνομα σας (με λατινικούς χαρακτήρες). Π.χ. αν παρέδιδα εγώ εργασία θα την ονόμαζα ως εξής: Fouskakis-Dimitris.pdf.
- Παρακαλώ χρησιμοποιήστε **ένα εξώφυλλο στο pdf αρχείο σας**, στο οποίο να υπάρχει κατάλληλος τίτλος και να αναγράφεται **υποχρεωτικά το ονοματεπώνυμο σας, το πρόγραμμα (προπτυχιακό ή μεταπτυχιακό που παρακολουθείτε) καθώς και το email σας και ο αριθμός μητρώου σας.**
- Θα πρέπει να **αποστείλετε ένα μόνο αρχείο**. Η εργασία θα πρέπει να περιλαμβάνει τους κώδικες της R, όχι σε παράρτημα αλλά στην απάντηση του κάθε ερωτήματος, με πλήρη επεξήγηση, γραφήματα και πλήρη περιγραφή των αποτελεσμάτων.
- Θα δοθεί ιδιαίτερη σημασία στην παρουσίαση της εργασίας. Η εργασία πρέπει να είναι κατανοητή και να περιγράφει οτιδήποτε χρησιμοποιήσατε πειστικά για κάποιον που δεν γνωρίζει πάρα πολλά για το αντικείμενο.

Εύχομαι Επιτυχία