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. Στα παρακάτω ερωτήματα μην χρησιμοποιήσετε καμία έτοιμη συνάρτηση προσομοίωσης τιμών της R πέραν της *runif*.

α) Χρησιμοποιώντας τη μέθοδο αντιστροφής προσομοιώστε 10000 τιμές από την σ.π.π. $f(x) = 3x^2, \quad 0 \le x \le 1$. Συγκρίνετε το ιστόγραμμα των προσομοιωμένων τιμών με το γράφημα της σ.π.π. Επιπλέον συγκρίνετε τον μέσο και την τυπική απόκλιση των προσομοιωμένων τιμών με τον μέσο και τυπική απόκλιση της θεωρητικής κατανομής.

β) Επαναλάβετε το παραπάνω ερώτημα χρησιμοποιώντας τώρα την μέθοδο απόρριψης. Εκτιμήστε την ολική πιθανότητα αποδοχής του αλγορίθμου σας και συγκρίνετέ την με την θεωρητική. Συγκρίνετε το ιστόγραμμα των προσομοιωμένων τιμών με το γράφημα της σ.π.π. Επιπλέον συγκρίνετε τον μέσο και την τυπική απόκλιση των προσομοιωμένων τιμών με τον μέσο και τυπική απόκλιση της θεωρητικής κατανομής.

γ) Έστω ότι θέλετε να υπολογίσετε το ολοκλήρωμα $I = \int_{38}^{\infty} f(x)dx = \int_{38}^{\infty} e^{-x}x^2 dx$. Βρείτε την τιμή του εν λόγω ολοκληρώματος. Εν συνεχεία σκοπός σας είναι να εκτιμήσετε την τιμή του I, χρησιμοποιώντας δειγματοληψία σπουδαιότητας, προσομοιώνοντας 1000 τιμές από δύο διαφορετικές κατανομές σπουδαιότητας. Εφαρμόστε την κάθε διαδικασία 100 φορές και στο τέλος συγκρίνετε την μέση τιμή (στις 100 επαναλήψεις) των εκτιμητών που παίρνετε με κάθε μέθοδο με την τιμή του I και τις διασπορές (στις 100 επαναλήψεις) των εκτιμητών που παίρνετε με κάθε μέθοδο μεταξύ τους. Ως κατανομή σπουδαιότητας χρησιμοποιήστε (α) την περικομμένη

εκθετική κατανομή με σ.π.π. $g(x) = e^{-(x-38)}$, $x \geq 38$. (β) την κατανομή Pareto με σ.π.π. $g(x) = 4 \cdot 38^4 x^{-5}$, $x \geq 38$. Για να προσομοιώσετε από τις δύο κατανομές σπουδαιότητας χρησιμοποιείστε την μέθοδο αντιστροφής.

2. Προσομοιώστε 50 τιμές από την τυποποιημένη κανονική κατανομή με χρήση της εντολής *rnorm* στην R. Χρησιμοποιήστε αυτές τις 50 παρατηρήσεις με τον *Epanechnikov* πυρήνα και υπολογίστε την *cross-validated* πιθανοφάνεια για τιμές του πλάτους h = 0.01, 0.02,…, 4.99, 5.00. Για ποια τιμή του h μεγιστοποιείται η *cross-validated* πιθανοφάνεια; Για την τιμή του πλάτους που επιλέξατε προβείτε στο διάγραμμα της εκτίμησης της σ.π.π. με χρήση του *Epanechnikov* πυρήνα και σχολιάστε.

3. Έστω ότι μελετάμε τον χρόνο επιβίωσης πειραματόζωων σε ένα νέο φάρμακο, σε μήνες. Σε δείγμα μεγέθους n = 20, πήραμε τις ακόλουθες παρατηρήσεις για m = 16 τ.μ.

0.05511269, 0.07437246, 0.11098159, 0.13552999,
0.20371014, 0.22090690, 0.23699706, 0.27435875,
0.28669966, 0.47155521, 0.96822690, 0.97200651,
1.04514368, 1.37989648, 1.49109121, 1.53370336,

ενώ για τις υπόλοιπες k = 4 τ.μ. έχει απλά καταγραφεί ότι είχαν τιμή μεγαλύτερη από 1.70 (λογοκριμένα - *censored* δεδομένα). Θεωρούμε ότι ο εν λόγω χρόνος επιβίωσης ακολουθεί την εκθετική κατανομή με παράμετρο θ > 0 και σκοπός μας είναι να εκτιμήσουμε την άγνωστη παράμετρο θ. Θεωρήστε τον αλγόριθμο EM για την εκτίμηση του θ. Αναπτύξτε (θεωρητικά) πλήρως τα βήματα του αλγορίθμου και εν συνεχεία δημιουργήστε μια δική σας συνάρτηση στην R που θα υλοποιεί τον αλγόριθμο. Ως αρχική τιμή θεωρήστε τον αντίστροφη τιμή του δειγματικού μέσου των 16 δεδομένων που έχετε πραγματικές τιμές, ενώ ως κριτήριο τερματισμού, για δύο διαδοχικές επαναλήψεις (r) και (r+1), χρησιμοποιείστε το παρακάτω: $\left(\theta^{(r+1)} - \theta^{(r)}\right)^2 \leq 10^{-10}$. Δημιουργήστε ένα διάγραμμα της τιμής της λογαριθμικής πιθανοφάνειας των παρατηρούμενων τιμών σας για κάθε επανάληψη του αλγορίθμου και σχολιάστε.

4. Θεωρείστε το πρόβλημα επιλογής επεξηγηματικών μεταβλητών στην πολλαπλή γραμμική παλινδρόμηση με n = 50 παρατηρήσεις και p = 15 επεξηγηματικές μεταβλητές. Προσομοιώστε με τη βοήθεια της R (με χρήση της *rnorm*) τιμές για τις δέκα πρώτες επεξηγηματικές μεταβλητές από την πολυδιάστατη κανονική κατανομή με μέση τιμή **0** και πίνακα συνδιακύμανσης τον ταυτοτικό, ενώ για τις υπόλοιπες προσομοιώστε τιμές με βάση τη σχέση:

$$X_{ij} \sim N(0.2X_{i1} + 0.4X_{i2} + 0.6X_{i3} + 0.8X_{i4} + 1.1X_{i5}, 1),\ j = 11,...,15 \text{ και } i = 1,...,50.$$

Για τη μεταβλητή απόκρισης, προσομοιώστε τιμές, με τη βοήθεια της R (με χρήση της *rnorm*), με βάση τη σχέση

$$Y_i \sim N(4 + 2X_{i1} - X_{i5} + 2.5X_{i7} + 1.5X_{i11} + 0.5X_{i13}, 1.5^2),\ i = 1,...,50.$$

Εν συνεχεία θεωρείστε το πλήρες πολλαπλό γραμμικό μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{15} X_{15} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

i) 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.
ii) Εφαρμόστε τη μεθοδολογία Lasso με την βοήθεια της βιβλιοθήκης *glmnet* της R και σχολιάστε τα αποτελέσματα. Χρησιμοποιώντας *cross-validation* σχολιάστε την επιλογή της παραμέτρου ποινής λ καθώς και της παραμέτρου συρρίκνωσης s.
iii) Επιλέξτε το πλήρες μοντέλο καθώς και το μοντέλο που γέννησε τα δεδομένα. Χρησιμοποιώντας *5 fold cross-validation* και τη συνάρτηση PRESS (*Prediction Error Sum of Squares*) εξετάστε ποιο μοντέλο προσαρμόζεται καλύτερα στα δεδομένα σας με χρήση δικού σας κώδικα στην R.
iv) Επιλέξτε το μοντέλο που γέννησε τα δεδομένα. Χρησιμοποιώντας 1000 *Bootstrap* δείγματα από τα υπόλοιπα, δώστε εκτιμήτριες και τυπικά σφάλματα για τους συντελεστές του εν λόγω μοντέλου χρησιμοποιώντας έτοιμες συναρτήσεις στην R από την βιβλιοθήκη *bootstrap*. Προβείτε σε συγκρίσεις και σχολιασμό με τα αποτελέσματα που παίρνετε με χρήση της εντολής *lm* στην R.

Οδηγίες

- Η εργασία θα πρέπει **να παραδοθεί ηλεκτρονικά** στο email μου, fouskakis@math.ntua.gr, **μέχρι την Πέμπτη 25 Ιουνίου 2020 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 την γράψετε υποχρεωτικά σε Latex**. 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.
- Κάποια στιγμή μέσα στον Ιούλιο, σε ημερομηνία που θα ανακοινωθεί από το *mycourse*, θα υπάρξει μια μίνι εξέταση της εργασία μέσω ***teams (χρησιμοποιώντας τον σύνδεσμο του μαθήματος)***. **Η εξέταση θα ξεκινήσει στις 09.00πμ.** και θα καλείται από την πλατφόρμα ο κάθε φοιτητής που παρέδωσε εργασία μεμονωμένα για 10 λεπτά περίπου. Όλοι οι

φοιτητές εκείνη την μέρα θα πρέπει να είναι διαθέσιμοι και *online* ώστε να απαντήσουν στην κλήση που θα λάβουν μετά την παραπάνω ώρα. Η παρουσία όλων εκείνη την μέρα είναι υποχρεωτική. Σε περίπτωση απουσίας η εργασία δεν θα βαθμολογηθεί.

Εύχομαι Επιτυχία