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. Έστω $\varphi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ η σ.π.π. της τυποποιημένης κανονικής κατανομής. Θεωρήστε το ολοκλήρωμα

$$J = \int_{-\infty}^{+\infty} (x+a)^2 \varphi(x) dx = 1 + a^2.$$

(α) Εκτιμήστε το J με χρήση *Monte Carlo* ολοκλήρωσης, προσομοιώνοντας τιμές από την τυποποιημένη κανονική κατανομή. Χρησιμοποιήστε 100 και 1000 προσομοιωμένες τιμές και θεωρήστε ότι το a = 0, 1, 2, 3, 4.

(β) Αποδείξτε θεωρητικά ότι ο παραπάνω *Monte Carlo* εκτιμητής είναι αμερόληπτος και βρείτε την θεωρητική του τυπική απόκλιση.

(γ) Εφαρμόστε δειγματοληψία σπουδαιότητας με χρήση της συνάρτησης $g(x) = \varphi(x-a)$. Επαναλάβετε τα ερωτήματα (α) και (β) και προβείτε σε συγκρίσεις των τυπικών σφαλμάτων των δύο εκτιμητών.

(δ) Θεωρήστε 1000 προσομοιωμένες τιμές, a = 4 και τον εκτιμητή του ερωτήματος (α). Χρησιμοποιώντας την τεχνική *Bootstrap*, με χρήση δικής σας συνάρτησης στην R, εκτιμήστε το τυπικό σφάλμα του εκτιμητή και συγκρίνετέ το με το θεωρητικό.

2. Έστω ότι θέλετε να προσομοιώσετε 1000 τιμές από την σ.π.π.

$$f(x) = \frac{1}{e^3 - 1} e^x, \; x \in [0,3].$$

(α) Χρησιμοποιήστε τη μέθοδο αντιστροφής για την προσομοίωση, με χρήση δικού σας κώδικα στην R. Συγκρίνετε το διάγραμμα της f(x) με το ιστόγραμμα των προσομοιωμένων τιμών.

(β) Χρησιμοποιήστε την μέθοδο απόρριψης για την προσομοίωση, με χρήση δικού σας κώδικα στην R. Συγκρίνετε το διάγραμμα της f(x) με το ιστόγραμμα των προσομοιωμένων τιμών.

(γ) Προσομοιώστε 100 τιμές από την f(x) μέσω της μεθόδου αντιστροφής. Χρησιμοποιήστε *Epanechnikov* πυρήνα και τα εν λόγω δεδομένα για να εκτιμήσετε την f(x). Για την εύρεση του "βέλτιστου" πλάτους h μεγιστοποιήστε την *cross-validated* πιθανοφάνεια, με χρήση δικού σας κώδικα στην R. Προβείτε σε ένα διάγραμμα της εκτιμώμενης f(x) για το h που βρήκατε και σχολιάστε το αποτέλεσμα που πήρατε.

(δ) Προσομοιώστε 10 τιμές από την f(x) μέσω της μεθόδου αντιστροφής. Χρησιμοποιώντας τις εν λόγω τιμές προβείτε σε έναν *Bootstrap* έλεγχο υπόθεσης (χρησιμοποιώντας δική σας συνάρτηση και όχι κάποια έτοιμη της R) της μηδενικής υπόθεσης $\mu = 2$ έναντι της εναλλακτικής $\mu \neq 2$, σε επίπεδο σημαντικότητας 5%, όπου το μ δηλώνει την (υποθετικά) άγνωστη μέση τιμή της κατανομής f(x). Απαντήστε στο ερευνητικό σας ερώτημα και με τη βοήθεια ενός 95% *Bootstrap* διαστήματος εμπιστοσύνης (χρησιμοποιώντας δική σας συνάρτηση και όχι κάποια έτοιμη της R), βασισμένου σε ποσοστιαία σημεία. Για τον έλεγχο υπόθεσης και για το διάστημα εμπιστοσύνης χρησιμοποιήστε 1000 *Bootstrap* δείγματα. Βρείτε την πραγματική μέση τιμή της f(x) και σχολιάστε τα αποτελέσματα του ελέγχου και του διαστήματος εμπιστοσύνης,

3. (α) Θεωρήστε την Γάμμα κατανομή με σ.π.π.

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \text{ όπου } \alpha, \beta > 0 \text{ άγνωστες παράμετροι.}$$

Έστω ότι διαθέτετε τυχαίο δείγμα μεγέθους n από την παραπάνω κατανομή. Ποια είναι η επαρκής στατιστική συνάρτηση και τι διάσταση έχει; Αναπτύξτε θεωρητικά (μόνο) τα βήματα του *Newton Raphson* αλγόριθμου για την μεγιστοποίηση της λογαριθμικής πιθανοφάνειας ως προς α και β.

(β) Η κατανομή Polya (α, β) αποτελεί μια γενίκευση της Αρνητικής Διωνυμικής. Εξαρτάται από δύο παραμέτρους $\alpha, \beta > 0$ και έχει σ.π.π.

$$f(x) = \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)} \left(\frac{\beta}{1+\beta}\right)^{\alpha} \left(\frac{1}{1+\beta}\right)^{x}.$$

Αρχικά αποδείξτε ότι η τ.μ. X ακολουθεί την κατανομή Polya (α, β) όταν

$$X \mid \theta \sim \text{Poisson}(\theta)$$
$$\theta \sim \text{Gamma}(\alpha, \beta).$$

Επιλέξτε α = 2 και β = 5 και με χρήση της παραπάνω μίξης προσομοιώστε n = 10000 τιμές (x_i) από την Polya (α, β) κατανομή. Χωρίς να αποθηκεύσετε τις θ_i τιμές θωρήστε πως τα "πλήρη" δεδομένα σας είναι τα ζεύγη (x_i, θ_i), i = 1,…,10000, όπου θ_i είναι "ελλιπείς τιμές". Σκοπός σας είναι να εκτιμήσετε τις παραμέτρους α και β. Αν είχατε παρατηρήσει τα θ_i τότε θα έπρεπε απλά να εκτιμήσετε τις παραμέτρους της κατανομής Γάμμα, με την μέθοδο μέγιστης πιθανοφάνειας. Θεωρήστε τον αλγόριθμο EM για την εκτίμηση του α και β. Αναπτύξτε (θεωρητικά) πλήρως τα βήματα του αλγορίθμου, με χρήση των επαρκών στατιστικών, και εν συνεχεία δημιουργήστε μια δική σας συνάρτηση στην R που θα υλοποιεί τον αλγόριθμο. Στο M-step θα χρειαστεί να εφαρμόσετε τον αλγόριθμο *Newton Raphson* του ερωτήματος (α). Επίσης θα χρειαστείτε τις συναρτήσεις `digamma` και `trigamma` της R. Ως αρχικές τιμές θεωρήστε τις α = 1 και β = 1 και ως κριτήριο τερματισμού, για δύο διαδοχικές επαναλήψεις (r) και (r+1), χρησιμοποιείστε: $\left(\alpha^{(r+1)}-\alpha^{(r)}\right)^2+\left(\beta^{(r+1)}-\beta^{(r)}\right)^2 \le 10^{-10}$. Πόσο καλά ο αλγόριθμος εκτίμησε τις τιμές των α και β;

4. Θεωρήστε 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 θεωρήστε το πλήρες πολλαπλό γραμμικό μοντέλο

$$Y=\beta_0+\beta_1X_1+...+\beta_{15}X_{15}+\varepsilon,\ \ \varepsilon \sim N(0,\sigma^2).$$

(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.

(β) 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.

Οδηγίες

- Η εργασία θα πρέπει **να παραδοθεί ηλεκτρονικά** στο email μου, fouskakis@math.ntua.gr, **μέχρι την Πέμπτη 1 Ιουλίου 2021 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**. Ο κώδικας θα πρέπει **υποχρεωτικά**

να είναι σε R. Μπορείτε να χρησιμοποιήσετε οποιαδήποτε έτοιμη συνάρτηση στην R, εκτός και αν στην άσκηση σας ζητείτε διαφορετικά.

- 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.
- Στις **11 Ιουλίου 2021**, θα υπάρξει μια μίνι εξέταση της εργασία μέσω ***teams (χρησιμοποιώντας τον σύνδεσμο του μαθήματος)***. **Η εξέταση θα ξεκινήσει στις 10.00πμ.** και θα καλείται από την πλατφόρμα ο κάθε φοιτητής που παρέδωσε εργασία μεμονωμένα για 10 λεπτά περίπου. Μέσω του *mycourses*, λίγες μέρες πριν την ημέρα εξέτασης, θα ανακοινωθεί ένα πρόγραμμα που περίπου θα τηρήσουμε για την ώρα εξέτασης του καθενός. Όλοι οι φοιτητές εκείνη την μέρα θα πρέπει να είναι διαθέσιμοι και *online, λίγο πριν και λίγο μετά από την προκαθορισμένη ώρα,* ώστε να απαντήσουν στην κλήση που θα λάβουν από εμένα. Η παρουσία όλων εκείνη την μέρα είναι υποχρεωτική. Σε περίπτωση απουσίας η εργασία δεν θα βαθμολογηθεί.

Εύχομαι Επιτυχία