ΜΑΣ 452

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ, ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ

ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΙΙ

ΔΙΔΑΣΚΩΝ: ΣΕΡΓΙΟΣ ΑΓΑΠΙΟΥ

ΕΡΓΑΣΙΑ 1 ΦΕΒΡΟΥΑΡΙΟΣ 2019

Ημερομηνία Παράδοσης: 26/02/2019

Λύστε όλες τις ασκήσεις. Για την επίλυση της ερώτησης 3, απαιτείται η χρήση της γλώσσας **R**. Μαζί με τις χειρόγραφες απαντήσεις σας, θα πρέπει να παραδώσετε και ηλεκτρονικά αρχείο **R** με τον σχετικό κώδικα.

---

1. Έστω $P \in \mathbb{R}^{n\times n}$ συμμετρικός πίνακας.

i) Να χρησιμοποιήσετε το Φασματικό Θεώρημα [1] για να δείξετε ότι ο P είναι ταυτοδύναμος με $\text{rank}(P) = r$ αν και μόνο αν έχει r ιδιοτιμές ίσες με 1 και $n - r$ ίσες με 0. **[1]**

ii) Με βάση το μέρος (i), να δείξετε ότι αν P πίνακας προβολής (δηλαδή συμμετρικός και ταυτοδύναμος) τότε $\text{Tr}\,(P) = \text{rank}(P) = \dim(V)$ όπου V ο χώρος στον οποίο προβάλει ο P. **[1]**

---

2. i) Έστω $X \sim N_d(m_1, I_d)$ και $Y \sim N_d(m_2, I_d)$ με αντίστοιχες συναρτήσεις πυκνότητας πιθανότητας g_1, g_2 αντίστοιχα. Να υπολογίσετε την απόκλιση $KL(g_1|g_2)$ και να εξετάσετε αν είναι συμμετρική. **[1]**

ii) Έστω $X \sim N(\mu_1, \sigma_1)$ και $Y \sim N(\mu_2, \sigma_2)$ με αντίστοιχες συναρτήσεις πυκνότητας πιθανότητας f_1, f_2 αντίστοιχα. Να υπολογίσετε την απόκλιση $KL(f_1|f_2)$ και να εξετάσετε αν και πότε είναι συμμετρική. **[1]**

---

3. Στο αρχείο **happiness.csv** σας δίνονται δεδομένα για το δείκτη ευτυχίας, μεταβλητή **Score**, 158 χωρών ως συνάρτηση 6 επεξηγηματικών μεταβλητών που είναι δείκτες που αντιστοιχούν στην επίδοση των χωρών σε 6 διαφορετικούς τομείς όπως το μέσο κατά κεφαλή εισόδημα ή το μέσο προσδόκημο ζωής[2]. Θεωρήστε το γραμμικό μοντέλο παλινδρόμησης.

[1] Φασματικό Θεώρημα: οι συμμετρικοί πίνακες διαγωνιοποιούνται από ορθογώνιους πίνακες

[2] GDP, Family, LifeExpectancy, Freedom, TrustGovernment, Generosity

συνεχίζεται

a) Να παρουσιάσετε τον πίνακα δειγματικής συσχέτισης των διαφόρων μεταβλητών στα δεδομένα. Ποια επεξηγηματική μεταβλητή φαίνεται να έχει την υψηλότερη συσχέτιση με το Score; Ποια τη χαμηλότερη; Να αναφέρετε 2 ζεύγη επεξηγηματικών μεταβλητών που φαίνονται να είναι σχετικά ισχυρά συσχετισμένες μεταξύ τους.

b) Να κάνετε τους εξής ελέγχους σε επίπεδο σημαντικότητας 0.05, προσαρμόζοντας τα κατάλληλα γραμμικά μοντέλα.

   i) Ποιες μεταβλητές φαίνονται να είναι σημαντικές με δεδομένο ότι όλες οι υπόλοιπες βρίσκονται στο μοντέλο;

   ii) Αφαιρέστε τη μεταβλητή Freedom και επαναλάβετε την προσαρμογή. Τι παρατηρείτε;

c) Να εφαρμόσετε τη μέθοδο προς τα εμπρός επιλογής υποσυνόλου των μεταβλητών με κριτήριο και να απαντήσετε τα πιο κάτω ερωτήματα:

   i) Η μέθοδος επιστρέφει τα καλύτερα υποσύνολα για αριθμό μεταβλητών $p = 1$ μέχρι $p = 6$. Είναι εμφωλιασμένα αυτά τα καλύτερα υποσύνολα; Περιμένατε να είναι εμφωλιασμένα ή όχι; Εξηγήστε σύντομα.

   ii) Πόσες και ποιες μεταβλητές έχουν τα βέλτιστα υποσύνολα σύμφωνα με τα κριτήρια R^2_{adj}, C_p και BIC;

d) Να εφαρμώσετε τη μέθοδο βελτίστου υποσυνόλου με χρήση Cross-Validation και Bootstrap. Πόσες και ποιες μεταβλητές περιέχει το βέλτιστο μοντέλο κάθε φορά; Να εμφανίσετε τις αντίστοιχες σταθερές παλινδρόμησης των βέλτιστων μοντέλων.

[6]

END