

ΜΟΝΤΕΛΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Διαγώνισμα 1

ΘΕΜΑ 1

Στον παρακάτω πίνακα δίνεται το training set από ένα μικρό πρόβλημα πρόβλεψης μιας λοίμωξης. Οι μεταβλητές X_1 και X_2 αντιστοιχούν σε τυποποιημένες μετρήσεις πυρετού και συχνότητας βήχα, ενώ η δίτιμη μεταβλητή Y δείχνει την ύπαρξη ή όχι της λοίμωξης (1 και 0, αντίστοιχα)

X_1	X_2	Y
2	2	0
2	4	0
2	5	1
3	3	1
3	4	1
3	2	0
4	3	0

Για την πρόβλεψη της τιμής του Y για μια νέα παρατήρηση χρησιμοποιούμε τη μέθοδο k -nearest-neighbor με παράμετρο k .

Έστω μια νέα παρατήρηση ενός ατόμου με τιμές $X_1=3$ και $X_2=5$

(α) Να γίνει η πρόβλεψη αν το άτομο έχει τη λοίμωξη, για τιμές του $k=1, 2, 7$.

(β) Πώς συμπεριφέρεται ο αλγόριθμος πρόβλεψης όταν $k=7$; Εξηγήστε.

(γ) Σχεδιάστε ποια μορφή αναμένεται (ποιοτικά) να έχει το διάγραμμα bias-variance για τιμές του k από 1 έως 7 και εξηγήστε γιατί.

ΘΕΜΑ 2

Το αρχείο prob2.Rdata περιέχει τα dataframes $p2train$, $p2test$ τα οποία είναι τα training και test sets, αντίστοιχα, από μετρήσεις μιας ανεξάρτητης μεταβλητής x και μιας εξαρτημένης μεταβλητής y .

(α) Εφαρμόστε μοντέλα πολυωνυμικής παλινδρόμησης βαθμού 2,3,6,9 στο training set. Βρείτε το μοντέλο με το μικρότερο mean square error μέσω cross-validation στο training set.

(β) Για το μοντέλο που καταλήξατε στο (α), υπολογίστε το mean square prediction error από το test set.

(γ) Εφαρμόστε το μοντέλο natural splines στο training set, με αριθμό κόμβων 2.

(δ) Για το μοντέλο που καταλήξατε στο (γ), υπολογίστε το mean square prediction error από το test set.

ΘΕΜΑ 3

Το αρχείο prob3.Rdata περιέχει τα dataframes $p3train$, $p3test$ τα οποία είναι τα training και test sets, αντίστοιχα, από δεδομένα λοίμωξης παιδιών. Οι μεταβλητές fever_hours, age, sex, wcc, prevAB, pct, cpr είναι οι πιθανοί παράγοντες ενώ η μεταβλητή Infection δείχνει την ύπαρξη λοίμωξης (0=αρνητικό, 1=θετικό).

(α) Δημιουργήστε ένα δέντρο απόφασης με βάση το training set για πρόβλεψη της λοίμωξης βασισμένο μόνο στις μεταβλητές wcc και pct. Σχεδιάστε το δέντρο στον

υπολογιστή. Σχεδιάστε στην κόλλα σας τις περιοχές πρόβλεψης 0 ή 1 στο καρτεσιανό επίπεδο των wcc, pct, με βάση τα αποτελέσματα του δέντρου.

(β) Εφαρμόστε το δέντρο που βρήκατε στο (α) στο test set, και υπολογίστε τα ποσοστά λάθους ανά κατηγορία (ψευδώς θετικά και ψευδώς αρνητικά).

(γ) Δημιουργήστε ένα μοντέλο πρόβλεψης βασισμένο στον αλγόριθμο random forest και το training set, με όλες τις ανεξάρτητες μεταβλητές και τιμές της παραμέτρου $m=2$ και 3.

Εφαρμόστε τα μοντέλα που βρήκατε στο test set και επιλέξτε αυτό που έχει το μικρότερο misclassification error. Για το μοντέλο αυτό υπολογίστε τα ποσοστά λάθους ανά κατηγορία (ψευδώς θετικά και ψευδώς αρνητικά).

ΘΕΜΑ 4

Το αρχείο prob4.Rdata περιέχει τα dataframes p4train, p4test τα οποία είναι τα training και test sets, αντίστοιχα, από δεδομένα σωματικών μετρήσεων και ποσοστού σωματικού λίπους ανδρών. Η μεταβλητή brozek περιέχει το ποσοστό λίπους σώματος υπολογισμένου μέσω της εξίσωσης Brozek που βασίζεται στη μέτρηση της πυκνότητας λίπους.

Σκοπός της άσκησης είναι να αναπτυχθεί ένα προβλεπτικό μοντέλο για το ποσοστό λίπους με βάση σωματικά χαρακτηριστικά που περιγράφονται από τις μεταβλητές :

Age: Ηλικία

weight: Βάρος (lbs)

height – Ύψος (inches)

BMI – Δείκτης μάζας σώματος = $\text{Weight}/\text{Height}^2$ (kg/m²)

neck – Περίμετρος λαιμού (cm)

chest – Περίμετρος στήθους (cm)

abdom – Περίμετρος κοιλιάς (cm)

hip – Περίμετρος γοφών (cm)

dthigh – Περίμετρος μηρού (cm)

knee – Περίμετρος γόνατος (cm)

ankle – Περίμετρος αστραγάλου (cm)

biceps – Περίμετρος δικεφάλων (cm)

forearm – Περίμετρος βραχίονα (cm)

wrist – Περίμετρος καρπού (cm)

Αναπτύξτε ένα προβλεπτικό μοντέλο για το ποσοστό λίπους, χρησιμοποιώντας όποια μέθοδο ή μεθόδους προτιμάτε. Περιγράψτε τη διαδικασία ανάπτυξης του μοντέλου σας (π.χ. ποιες μεθόδους δοκιμάσατε, πώς τις συγκρίνατε, γιατί καταλήξατε στο συγκεκριμένο μοντέλο).