

ΜΟΝΤΕΛΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Διαγώνισμα 2

ΘΕΜΑ 1

Έστω ότι θέλουμε να αναπτύξουμε ένα μοντέλο μηχανικής μάθησης για πρόβλεψη μιας εξαρτημένης μεταβλητής, όπου το σύνολο δεδομένων εκπαίδευσης (training set) αποτελείται από n παρατηρήσεις και p ανεξάρτητες μεταβλητές.

(α) Μεταξύ ενός μοντέλου παλινδρόμησης με k όρους και ενός μοντέλου παλινδρόμησης με $2k$ όρους, ποιο έχει τη μεγαλύτερη ευελιξία και γιατί;

(β) Για κάθε μια από τις παρακάτω περιπτώσεις αναφέρετε αν ένα μοντέλο μικρότερης ή μεγαλύτερης ευελιξίας αναμένεται να έχει μικρότερο σφάλμα πρόβλεψης (testing error) και γιατί:

(β1) Το n είναι πολύ μεγάλο και το p μικρό.

(β2) Το n είναι μικρό και το p μεγάλο.

(β3) Η πραγματική σχέση μεταξύ των ανεξάρτητων και της εξαρτημένης μεταβλητής είναι ισχυρά μη γραμμική.

(β4) Η διασπορά των αποκλίσεων $\sigma^2 = Var(\varepsilon)$ είναι πολύ μεγάλη.

ΘΕΜΑ 2

Το αρχείο prob2.Rdata περιέχει το dataframe prostate, το οποίο περιλαμβάνει δεδομένα σχετικά με την τιμή του PSA και άλλες κλινικές μετρήσεις ανδρών με καρκίνο του προστάτη. Σκοπός της μελέτης είναι να αναπτυχθεί ένα μοντέλο πρόβλεψης του PSA με βάση τις υπόλοιπες κλινικές μετρήσεις. Οι μεταβλητές αντιστοιχούν στα εξής:

lcavol	log μέγεθος όγκου
lweight	log βάρος προστάτη
age	ηλικία σε έτη
lbph	log του μεγέθους καλοήθους υπερπλασίας
lcp	log διείσδυσης στην κάψα
gleason	Gleason score (6-10)
pgg45	% όγκου με Gleason score 4 ή 5
lpsa	response
train	διαχωρισμός παρατηρήσεων T: Training set, F: test set

(α) Εφαρμόστε στο training set ένα μοντέλο παλινδρόμησης με εξαρτημένη μεταβλητή την lpsa και ανεξάρτητες τις υπόλοιπες. Υπολογίστε το μέσο τετραγωνικό σφάλμα πρόβλεψης στο test set.

(β) Εφαρμόστε το μοντέλο ridge regression στο training set επιλέγοντας την τιμή του λ μέσω cross validation. Υπολογίστε το μέσο τετραγωνικό σφάλμα πρόβλεψης στο test set.

(γ) Εφαρμόστε το μοντέλο lasso regression στο training set επιλέγοντας την τιμή του λ μέσω cross validation. Υπολογίστε το μέσο τετραγωνικό σφάλμα πρόβλεψης στο test set.

ΘΕΜΑ 3

Το αρχείο prob3.Rdata περιέχει τα dataframes p3train, p3test τα οποία είναι τα training και test sets, αντίστοιχα, τα οποία περιλαμβάνουν δεδομένα μελέτης σχετικά με τον κίνδυνο εμφάνισης αρτηριακής καρδιακής νόσου σε άνδρες μιας περιοχής της Νότιας Αφρικής το 1983. Οι μεταβλητές αντιστοιχούν στα εξής:

Sbp	συστολική αρτηριακή πίεση
tobacco	συνολική κατανάλωση προϊόντων καπνού (kg)

ldl	LDL χοληστερόλη
adiposity	ποσοστό λίπους στο σώμα
famhist	οικογενειακό ιστορικό καρδιαγγειακής νόσου (Ναι/Όχι)
typea	δείκτης συμπεριφοράς τύπου A στην κλίμακα Bonkers
obesity	δείκτης παχυσαρκίας
alcohol	κατανάλωση αλκοόλ (τρέχουσα)
age	ηλικία
chd	έκβαση – καρδιαγγειακή νόσος (Ναι/Όχι)

Η μεταβλητή typea αναφέρεται στον χαρακτηρισμό της συμπεριφοράς ενός ατόμου, όπου η συμπεριφορά τύπου A αντιστοιχεί σε επιθετικότητα, ανταγωνιστικότητα, επιμονή και ανυπομονησία κλπ, και έχει βρεθεί ότι σχετίζεται με την εμφάνιση καρδιακής νόσου. Μεγαλύτερη τιμή του δείκτη αντιστοιχεί σε εντονότερη εμφάνιση αυτών των χαρακτηριστικών.

Σκοπός της μελέτης είναι να αναπτυχθεί ένα μοντέλο πρόβλεψης της εμφάνισης καρδιαγγειακής νόσου με βάση τις υπόλοιπες μετρήσεις.

α) Εφαρμόστε ένα μοντέλο λογιστικής παλινδρόμησης με βάση το training set και υπολογίστε τα ποσοστά λάθους ανά κατηγορία στο test set (ψευδώς θετικά και ψευδώς αρνητικά).

β) Εφαρμόστε ένα μοντέλο γραμμικής διακρίνουσας ανάλυσης με βάση το training set και υπολογίστε τα ποσοστά λάθους ανά κατηγορία στο test set (ψευδώς θετικά και ψευδώς αρνητικά).

γ) Δημιουργήστε ένα δέντρο απόφασης με βάση το training set για πρόβλεψη της καρδιαγγειακής νόσου βασισμένο μόνο στις μεταβλητές tobacco και obesity. Σχεδιάστε το δέντρο στον υπολογιστή. Σχεδιάστε στην κόλλα σας τις περιοχές πρόβλεψης 0 ή 1 στο καρτεσιανό επίπεδο των tobacco, obesity, με βάση τα αποτελέσματα του δέντρου.

δ) 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).