Μηχανική Μάθηση Διαγώνισμα 4/4/2022

ΟΝΟΜΑΤΕΠΩΝΥΜΟ:

ΟΔΗΓΙΕΣ

Κατεβάστε το αρχείο examdatasets.zip από το φάκελο Ακαδημαϊκό Έτος 2021-22/Αρχεία Διαγωνίσματος Φεβρουαρίου 2022. Αποσυμπιέστε τα δύο αρχεία δεδομένων A2.Rdata και cluster.Rdata, που χρειάζονται στις ασκήσεις.

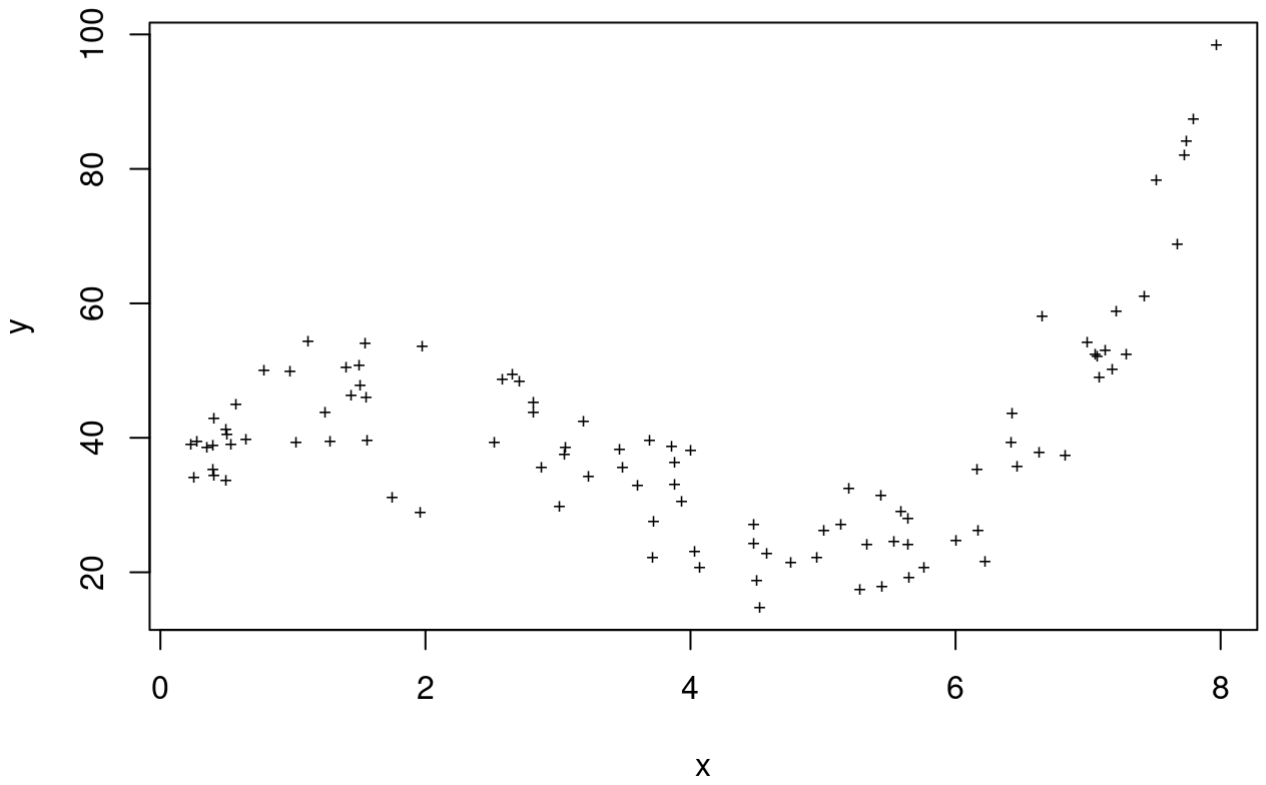
Σε όσες ασκήσεις χρησιμοποιήσετε R θα πρέπει ο κώδικας που χρησιμοποιήσατε να φαίνεται σε ένα script της μορφής onomatepwnimo-askisi-1.R. Οι τελικές απαντήσεις θα πρέπει να δίνονται στο γραπτό μαζί με την αιτιολόγηση όπου χρειάζεται. Στον κώδικα καλό θα είναι να υπάρχουν και σχόλια (σε γραμμές που αρχίζουν με #) σε όποια σημεία νομίζετε ότι χρειάζεται περαιτέρω επεξήγηση.

Στο τέλος θα συμπιέσετε όλα τα αρχεία που θα παραδώσετε σε ένα αρχείο onomatepwnimo.zip και είτε θα το αναρτήσετε στο eclass στην εργασία που έχει δημιουργηθεί είτε θα το παραδώσετε σε μένα σε flash drive που θα σας δώσω.

Καλή επιτυχία!

Άσκηση 1

Δίνεται ένα training set που φαίνεται στο παρακάτω σχήμα (με μια ανεξάρτητη μεταβλητή X και εξαρτημένη μεταβλητή Y)



Στο παραπάνω σύνολο δεδομένων εφαρμόζεται μια μέθοδος regression με εξομάλυνση, σύμφωνα με την οποία η συνάρτηση πρόβλεψης είναι

$$\hat{f} = \arg\min_f \left(\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int \left[f^{(m)}(x) \right]^2 dx \right)$$

Δώστε προσεγγιστικά διαγράμματα για τη συνάρτηση πρόβλεψης που αντιστοιχεί στις παρακάτω τιμές των λ, m και δικαιολογήστε τις απαντήσεις σας.

(α) $\lambda = \infty, m = 0.$

(β) $\lambda = \infty, m = 1.$

(γ) $\lambda = \infty, m = 2.$

(δ) $\lambda = \infty, m = 3.$

(ε) $\lambda = 0, m = 1.$

(στ) $\lambda = 0, m = 2.$

Άσκηση 2

Υποθέστε ότι έχουμε ένα rapid test για Covid που βασίζεται σε δύο μετρήσεις $X1, X2$ σύμφωνα με τις οποίες κατατάσσει έναν εξεταζόμενο ως θετικό ($Y = 1$) ή αρνητικό ($Y = 0$). Μας ενδιαφέρει να αξιολογήσουμε την ευαισθησία και την ειδικότητα του test, οι οποίες ορίζονται ως το ποσοστό πραγματικά θετικών που το rapid test ταξινομεί ως θετικούς και το ποσοστό των πραγματικά αρνητικών που το rapid test ταξινομεί ως αρνητικούς, αντίστοιχα.

Για την αξιολόγηση έχει επιλεγεί ένα training set στο οποίο εφαρμόστηκε λογιστική παλινδρόμηση και έδωσε τους συντελεστές $b_0 = -36, b_1 = 0.9, b2 = 1$.

Οι τιμές αυτές εφαρμόστηκαν στο test set που περιέχεται στο αρχείο A2.Rdata.

Να υπολογίσετε την ευαισθησία και την ειδικότητα του rapid test με βάση αυτό το test set.

(Οι τιμές του Y στο training και στο test set έχουν προκύψει από PCR και θεωρούνται 100% σωστές).

Άσκηση 3

Σε ένα πρόβλημα ταξινόμησης με δύο κατηγορίες $Y \in \{-1, 1\}$ και δύο ανεξάρτητες μεταβλητές $X1, X2$, δίνεται το παρακάτω training set με N=8 παρατηρήσεις:

Ind	X1	X2	Y
1	3	4	1
2	2	2	1
3	4	4	1
4	1	4	1
5	2	1	-1
6	4	3	-1
7	4	1	-1
8	3	1	-1

(α) Σχεδιάστε το training set στο επίπεδο των $X1, X2$ δείχνοντας με διαφορετικά σύμβολα τις τιμές του Y. Δείξτε ότι οι ευθείες που διέρχονται από τα σημεία { (2,1), (4,3) } και { (2,2), (4,4) } είναι παράλληλες.

(β) Προσδιορίστε το βέλτιστο διαχωριστικό υπερεπίπεδο (ευθεία) και βρείτε την εξίσωσή του.

(γ) Περιγράψτε τον κανόνα ταξινόμησης.

(δ) Προσδιορίστε το margin M και τα support vectors.

Από τις Ασκήσεις 4 και 5 λύστε μόνο μια της επιλογής σας

Άσκηση 4

Το αρχείο cluster.Rdata περιέχει δεδομένα δύο μεταβλητών $X1, X2$. Εφαρμόστε 3 βήματα της μεθόδου k-means clustering με 3 ομάδες χρησιμοποιώντας ως αρχικά κέντρα τα σημεία (20,20), (20,10) και (30,50). Δώστε τα κέντρα των ομάδων μετά τις 3 επαναλήψεις και σχεδιάστε τα σημεία των 3 ομάδων με διαφορετικά χρώματα. (Αν θέλετε μπορείτε να εφαρμόσετε τη μέθοδο k-means μέχρι το τέλος, αλλά δεν είναι υποχρεωτικό).

Άσκηση 5

(Η άσκηση μπορεί να γίνει χωρίς προγραμματισμό σε R.) Δίνεται το παρακάτω αρχείο δεδομένων με δύο μεταβλητές $X1, X2$

Ind	X1	X2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

Εφαρμόστε τη μέθοδο k-means clustering με 2 ομάδες, ξεκινώντας από μια αυθαίρετη ταξινόμηση των παρατηρήσεων. Στο τέλος του αλγορίθμου δημιουργήστε ένα διάγραμμα των παρατηρήσεων με διαφορετικά σύμβολα για κάθε ομάδα.