Υπολογιστική Εργασία του Μαθήματος Αναγνώριση Προτύπων
Ακαδημαϊκό Έτος 2021 – 2022
Γ.Τσιχριντζής – Δ. Σωτηρόπουλος

Προσέγγιση Τιμών Ακινήτων με χρήση Αλγορίθμων Μηχανικής Μάθησης

Στόχος της συγκεκριμένης εργασίας είναι η ανάπτυξη αλγορίθμων μηχανικής μάθησης για την προσέγγιση της διάμεσης τιμής ενός ακινήτου σε μια ευρύτερη γεωγραφική περιοχή της πολιτείας της Καλιφόρνια βάσει ενός συνόλου αντικειμενικών χαρακτηριστικών των ακινήτων στην εν λόγω περιοχή. Κάθε τέτοια περιοχή αποτελεί στην πραγματικότητα την μικρότερη γεωγραφική οντότητα που καταγράφηκε στην σχετική απογραφή του 1990 με πληθυσμιακό εύρος μεταξύ των 600 και 3000 κατοίκων. Το σχετικό σύνολο των δεδομένων είναι αποθηκευμένο στο συνοδευτικό αρχείο "**housing.csv**".

Η διαδικασία εκπαίδευσης των εμπλεκόμενων μηχανισμών μηχανικής μάθησης θα πρέπει να βασιστεί σε ένα σύνολο αντικειμενικών γνωρισμάτων των ακινήτων που υπάρχουν σε κάθε γεωγραφική περιοχή το οποίο περιλαμβάνει τα παρακάτω χαρακτηριστικά:

i. το γεωγραφικό μήκος (**longitude**) του κέντρου της περιοχής
ii. το γεωγραφικό πλάτος (**latitude**) του κέντρου της περιοχής
iii. την διάμεση ηλικία των ακινήτων (**housing_median_age**) της περιοχής
iv. το συνολικό πλήθος δωματίων (**total_rooms**) των ακινήτων της περιοχής
v. το συνολικό πλήθος υπνοδωματίων (**total_bedrooms**) των ακινήτων της περιοχής
vi. τον πληθυσμό (**population**) της περιοχής
vii. το πλήθος των νοικοκυριών (**households**) της περιοχής
viii. το διάμεσο εισόδημα (**median_income**) των κατοίκων της περιοχής
ix. την εγγύτητα προς τον ωκεανό (**ocean_proximity**) της περιοχής

ώστε να προσεγγιστεί:

x. η διάμεση τιμή (**median_house_value**) των ακινήτων της περιοχής.

Προ-επεξεργασία Δεδομένων

1. Θα πρέπει να αναγνωρίσετε τα υποσύνολα των αριθμητικών και των κατηγορικών χαρακτηριστικών.
2. Για το υποσύνολο των αριθμητικών χαρακτηριστικών θα πρέπει να πειραματιστείτε με διαφορετικές τεχνικές κλιμάκωσης (**scaling**) των δεδομένων ώστε όλα τα αριθμητικά χαρακτηριστικά να αναπαρίστανται στην ίδια κλίμακα.
3. Για το υποσύνολο των κατηγορικών χαρακτηριστικών μπορείτε να χρησιμοποιήσετε την One Hot Vector κωδικοποίηση ώστε τα εν λόγω δεδομένα να λάβουν διανυσματική αναπαράσταση.
4. Θα πρέπει να αναγνωρίσετε αν υπάρχουν αριθμητικά χαρακτηριστικά με ελλιπείς τιμές. Για τις συγκεκριμένες εγγραφές μπορείτε να συμπληρώσετε τις τιμές που απουσιάζουν με την διάμεση τιμή του χαρακτηριστικού.

Οπτικοποίηση Δεδομένων

1. Να αναπαραστήσετε γραφικά τα ιστογράμματα συχνοτήτων (που αντιστοιχούν στις συναρτήσεις πυκνότητας πιθανότητας) για κάθε μία από τις 10 μεταβλητές που εμπλέκονται στο πρόβλημα.
2. Προσπαθήστε να δημιουργήσετε δισδιάστατα γραφήματα των δεδομένων στα οποία να αναπαρίστανται με ευδιάκριτο τρόπο συνδυασμοί 2, 3 ή και 4 μεταβλητών.

Παλινδρόμηση Δεδομένων

1. Να υλοποιήσετε τον **Αλγόριθμο Ελάχιστου Μέσου Τετραγωνικού Σφάλματος** (**Least Mean Squares**), ώστε ο εκπαιδευμένος μηχανισμός μάθησης να υλοποιεί μία γραμμική παλινδρόμηση της μορφής $\boldsymbol{g}$: $\mathbb{R}^{\boldsymbol{l}} \rightarrow \mathbb{R}$**,** όπου l είναι η διάσταση του τελικού χώρου των χαρακτηριστικών.
2. Να υλοποιήσετε τον **Αλγόριθμο Ελάχιστου Τετραγωνικού Σφάλματος** (**Least Squares**), ώστε ο εκπαιδευμένος μηχανισμός μάθησης να υλοποιεί μία γραμμική παλινδρόμηση της μορφής $\boldsymbol{g}$: $\mathbb{R}^{\boldsymbol{l}} \rightarrow \mathbb{R}$**,** όπου l είναι η διάσταση του τελικού χώρου των χαρακτηριστικών.
3. Να υλοποιήσετε ένα πολυστρωματικό νευρωνικό δίκτυο, ώστε ο εκπαιδευμένος μηχανισμός μάθησης να υλοποιεί μία μη-γραμμική παλινδρόμηση της μορφής $\boldsymbol{g}$: $\mathbb{R}^{\boldsymbol{l}} \rightarrow \mathbb{R}$**,** όπου l είναι η διάσταση του τελικού χώρου των χαρακτηριστικών.

Παρατηρήσεις:

I. Για κάθε μηχανισμό μηχανικής μάθησης που θα υλοποιήσετε θα πρέπει να αναφέρετε την ακρίβεια παλινδρόμησης που επιτυγχάνει σε όρους **Μέσου Τετραγωνικού Σφάλματος** και **Μέσου Απόλυτου Σφάλματος** τόσο κατά την φάση της εκπαίδευσης όσο και κατά την φάση του ελέγχου σύμφωνα με την μέθοδο της 10-πλής διεπικύρωσης (**10 fold cross validation).**
II. Η διαδικασία κατάτμησης των δεδομένων σε υποσύνολα εκπαίδευσης και ελέγχου σύμφωνα με την μέθοδο της 10-πλής διεπικύρωσης θα πρέπει να λάβει υπόψη την ιδιαίτερη κατανομή των τιμών του χαρακτηριστικού που θα παίξει τον ρόλο της εξαρτημένης μεταβλητής παλινδρόμησης.
III. Παραδοτέα της εργασίας αποτελούν ο **κώδικας** της υλοποίησής σας σε MATLAB ή Python καθώς και ένα συνοδευτικό **κείμενο τεκμηρίωσης**.
IV. Μπορείτε να εργασθείτε σε ομάδες των **δύο** ή **τριών ατόμων**.
V. Καταληκτική ημερομηνία παράδοσης της εργασίας είναι η **τελευταία μέρα** της εξεταστικής περιόδου.

ΚΑΛΗ ΕΠΙΤΥΧΙΑ!