ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Πληροφορικής



Εργασία Μαθήματος «Αναγνώριση Προτύπων»

|  |  |
| --- | --- |
|  |  |
| Όνομα φοιτητή – Αρ. Μητρώου  (όλων σε περίπτωση ομαδικής εργασίας) | ΕΛΕΥΘΕΡΙΟΣ ΚΟΝΤΟΥΡΗΣ Π19077 |
| ΣΤΑΥΡΟΣ ΤΖΙΟΥΝΗΣ |
| ΝΙΚΟΛΑΟΣ ΣΤΥΛΙΔΙΩΤΗΣ Π19166 |
|  |
| Ημερομηνία παράδοσης | 5/3/2022 |

**Εκφώνηση της άσκησης**

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 𝒈: ℝ𝒍 → 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 𝒈: ℝ𝒍 → 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

𝒈: ℝ𝒍 → 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

Περιεχόμενα

[1 Προ-επεξεργασία δεδομένων 7](#_Toc97327701)

[2 Οπτικοποίηση Δεδομένων 9](#_Toc97327702)

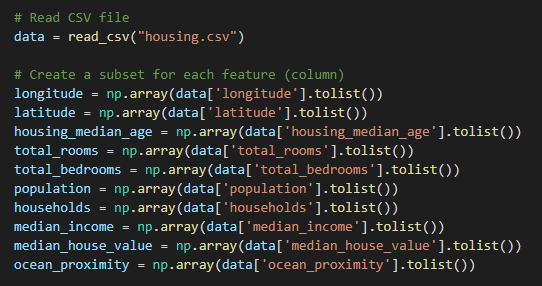
[3 Παλινδρόμηση Δεδομένων 10](#_Toc97327703)

[3.1 Αλγόριθμος Least Mean Squares 10](#_Toc97327704)

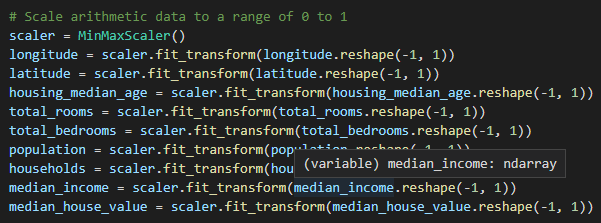
[3.2 Πολυστρωματικό νευρωνικό δίκτυο 11](#_Toc97327705)

[4 Βιβλιογραφικές Πηγές 12](#_Toc97327706)

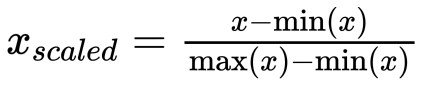
1. Προ-επεξεργασία δεδομένων



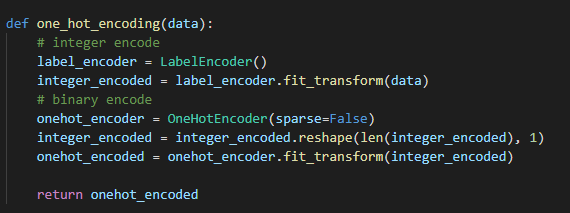
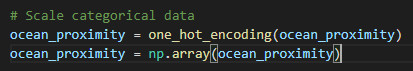
Αρχικά διαβάζουμε το αρχείο csv και μετατρέπουμε σε πίνακα καθεμία από τις στήλες του, ώστε να δημιουργήσουμε υποσύνολα για κάθε ένα χαρακτηριστικό.



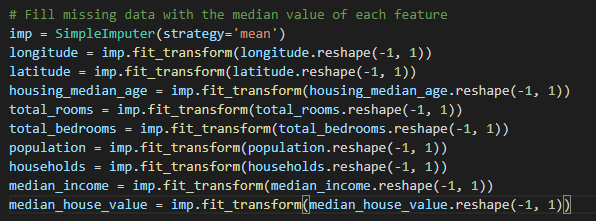
Έπειτα προχωράμε σε scale των δεδομένων χρησιμοποιώντας την MinMax τεχνική. Κατά τη μέθοδο αυτή, όλες οι τιμές μετατρέπονται σε διαφορετική κλίμακα και πιο συγκεκριμένα στο διάστημα [0,1]. Αυτό γίνεται με τον παρακάτω μαθηματικό τύπο:



Σημαντικό ρόλο σε αυτην την διαδικάσια παίζουν οι συναρτήσεις fit() και transform(), καθώς όμως και η fit\_transform(). Η πρώτη βρίσκει τους συντελεστές της εξισώσης που χρησιμοποιεί ο αντίστοιχος αλγόριθμος και η δεύτερη τους εφαρμόζει στο χαρακτηριστικό. Η τρίτη απλά συνδιάζει τα προηγούμενα δύο και χρησιμοποιείτε κυρίως κατά την φάση της δοκιμής. Στην προκείμενη περίπτωση χρησιμοποιούμε την παραπάνω εξίσωση και μετά κάνουμε fit\_transform() σε κάθε χαρακακτηριστικό ώστε να γίνει ‘scaled’.



Για το κατηγορικό δεδομένο ‘ocean\_proximity’ χρησιμοποιείτε η τεχνική του One-Hot-Vector encoding. Αρχικά παιρνάμε σε κάθε διαφορετική κατηγορία δεδομένων μια ετικέτα και μετά σύμφωνα με αυτές τιν ετικέτες εφαρμόζουμε τον αλγόριθμο.



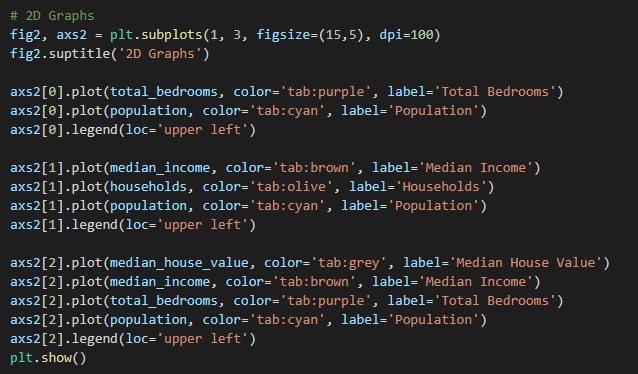
Έπειτα συμπληρώνουμε τις ελλειπείς τιμές χρησιμοποιώντας την συνάρτηση SimpleImputer() και την μέση τιμή του κάθε χαρακτηριστικού. Και σε αυτην την περίπτωση χρησιμοποιούμε το fit\_transform για να εφαρμόσουμε τις αλλαγές στα δεδομένα.



1. Οπτικοποίηση Δεδομένων



Για την οπτικοποίηση των δεδομένων φτιάχνουμε ένα παράθυρο με 10 ιστογράμματα (το 1 κενό) και περνάμε τα δεδομένα.



Για τα γραφήματα με τον συνδυασμό δεδομένων φτιάχνουμε 3 σχήματα.

1. Παλινδρόμηση Δεδομένων

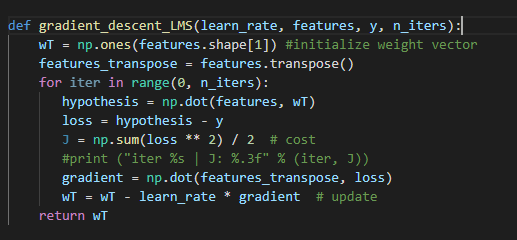
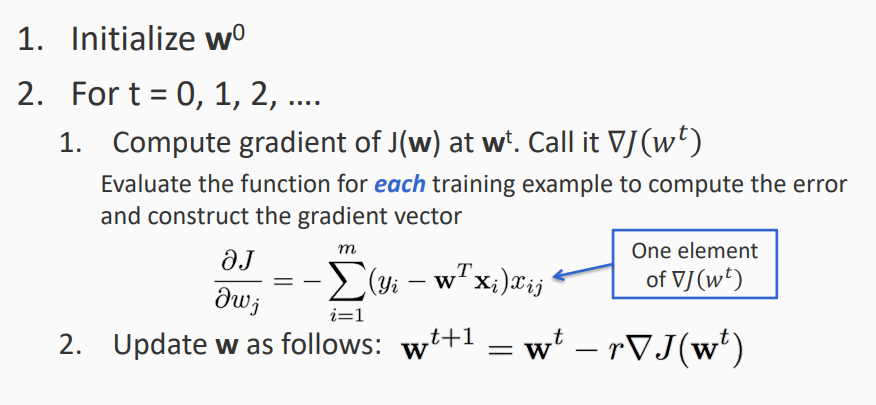
### 3.1 Αλγόριθμος Least Mean Squares

Πριν γίνει η παλινδρόμηση των δεδομένων, δημιουργούμε τυχαία υποσύνολα εκπαίδευσης από τα αρχικά δεδομένα και τα δίνουμε ως είσοδο στον αλγόριθμο. Στόχος μας είναι να προσεγγίσουμε την τιμή ‘house\_median\_value’ μέσω του γραμμικού μοντέλου:

**house\_median\_value** = w0 + w1\*longitude + w2\*latitude + w3\* housing\_median\_age + w4\* total\_rooms + w5\* total\_bedrooms + w6\* population + w7\* households + w8\* median\_income + w9\* ocean\_proximity

και χρησιμοποιώντας δεδομένα εκπαίδευσης, να βρούμε την καλύτερη τιμή του διανύσματος w. Με αυτό και δεδομένων κάποιων χαρακτηριστικών, θα μπορεί να δίνεται μια πρόβλεψη του ‘house\_median\_value’

Για να υπολογίσουμε την καλύτερη αυτή τιμή θα πρέπει να βρούμε το ελάχιστο τετραγωνικό σφάλμα (κόστος) που προκύπτει. Αυτό μπορεί να επιτευχθεί χρησιμοποιώντας τον αλγόριθμο Batch Gradient Descent:



Ο αλγόριθμος υπολογίζει για κάθε δεδομένο εκπαίδευσης το σφάλμα και κατασκευάζει το διάνυσμα της παραγώγου. Έπειτα με τα αποτελέσματα αυτά και τον βαθμό παλινδρόμισης υπολογίζει το νέο βάρος. Η διαδικασία επαναλαμβάνεται για συγκεκριμένο αριθμό επαναλήψεων. Ο βαθμός παλινδρόμισης είναι μια σταθερή παράμετρος που καθορίζει το βήμα του αλγόριθμου σε κάθε επανάληψη.

### 3.2 Πολυστρωματικό νευρωνικό δίκτυο

1. Βιβλιογραφικές Πηγές

<https://svivek.com/teaching/lectures/slides/linear-models/lms-regression.pdf>

<https://www.cs.princeton.edu/courses/archive/spring13/cos511/scribe_notes/0411.pdf>

<https://vitalflux.com/difference-between-online-batch-learning/>

<https://www.bogotobogo.com/python/python_numpy_batch_gradient_descent_algorithm.php>