ΑΛΕΞΑΝΔΡΕΙΟ ΤΕΙ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΑΘΗΜΑ: ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΚΑΘΗΓΗΤΕΣ : ΚΩΣΤΑΣ ΔΙΑΜΑΝΤΑΡΑΣ, ΚΩΣΤΑΣ ΓΟΥΛΙΑΝΑΣ

ΕΡΓΑΣΤΗΡΙΑΚΗ ΑΣΚΗΣΗ 7

ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΤΑ ΜΟΝΤΕΛΑ SVR ΚΑΙ MLP

Σκοπός της άσκησης: Η αποτίμηση της επίδοσης δύο κλασικών μοντέλων για την πρόβλεψη της τιμής ενός μεγέθους που παίρνει συνεχείς τιμές. Ένα τέτοιο πρόβλημα καλείται παλινδρόμηση (regression). Τα μοντέλα που θα χρησιμοποιηθούν είναι **(α) Support Vector Regression (SVR)** και **(β) Multi-Layer Perceptron (MLP)**. Θα εφαρμοστεί η μέθοδος της διασταύρωσης (Cross-Validation) και τα κριτήρια επίδοσης:

1. Mean Squared Error (MSE) – Μέσο Τετραγωνικό Σφάλμα
2. Mean Absolute Error (MAE) – Μέσο Απόλυτο Σφάλμα

Βήματα υλοποίησης:

1. Κατεβάστε το σύνολο δεδομένων (data set) **Housing** από την παρακάτω ιστοσελίδα:

   https://archive.ics.uci.edu/ml/datasets/Housing

   Αφορά τις τιμές των κατοικιών στα περίχωρα της Βοστώνης. Αποτελείται από 506 πρότυπα όπου κάθε πρότυπο περιέχει 13 χαρακτηριστικά της εκάστοτε κατοικίας. Ο στόχος είναι να προβλεφθεί η τιμή της κατοικίας με βάση αυτά τα χαρακτηριστικά.

   Το data set αποτελείται από δύο αρχεία:

   i. housing.data : περιέχει τα δεδομένα. Αποτελείται από 506 γραμμές, όπου κάθε γραμμή αντιστοιχεί σε ένα πρότυπο. Κάθε πρότυπο περιέχει 13 χαρακτηριστικά συν την τιμή στόχου. Στόχος είναι η τιμή: «MEDV (Median value of owner-occupied homes in $1000's)» που βρίσκεται στην τελευταία ($14^{η}$) στήλη κάθε γραμμής.

   ii. housing.names : ενημερωτικό κείμενο το οποίο περιέχει την περιγραφή των δεδομένων.

2. Διαβάστε το αρχείο δεδομένων housing.data χρησιμοποιώντας την βιβλιοθήκη pandas και την συνάρτηση read_csv()
3. Βρείτε το Πλήθος των attributes: NumberOfAttributes (στη συγκεκριμένη περίπτωση = 13), το Πλήθος των προτύπων: NumberOfPatterns (στη συγκεκριμένη περίπτωση = 506). Κατόπιν
   - Αρχικοποιήστε τον πίνακα των προτύπων x σε μηδέν. Πρέπει να έχει διαστάσεις (NumberOfAttributes-1) × NumberOfPatterns
   - Αρχικοποιήστε το διάνυσμα στόχων t σε μηδέν. Πρέπει να έχει διαστάσεις 1 × NumberOfPatterns
4. Γεμίστε τους πίνακες x, t ως εξής:
   - Ο πίνακας x περιέχει τα 13 πρώτα attributes (για όλα τα pattern)
   - Το διάνυσμα t περιέχει το $14^{ο}$ attribute για κάθε pattern. Οι στόχοι είναι πραγματικοί αριθμοί και δεν χρειάζονται περαιτέρω επεξεργασία.

5. Θα εφαρμοστεί η μέθοδος train_test_split() για K=9 folds.
6. Δώστε τις παραμέτρους gamma και C σε ένα διπλό loop. Πχ. δοκιμάστε gamma=[0.0001, 0.001, 0.01, 0.1] και C=[1, 10, 100, 1000]. Μέσα στο διπλό loop θα κάνετε Cross-Validation.

Για κάθε *gamma*

Για κάθε *C*

7. Στο Cross-Validation loop θα πρέπει να κάνετε τα εξής:

Για κάθε *fold*

- Δημιουργήστε τους πίνακες προτύπων xtrain και xtest (χωρίς επαύξηση) καθώς και τα διανύσματα στόχων ttrain και ttest.
- Δημιουργήστε ένα δίκτυο SVR με πυρήνα RBF χρησιμοποιώντας την κλάση SVR (http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html):
- SVR(*C*, *kernel='rbf'*, *gamma*, *degree*, *coef0*, ...) όπου:
  - *C = παράμετρος C του SVR*
  - *kernel = επιλογή συνάρτησης πυρήνα μεταξύ* 'linear', 'poly', 'rbf', 'sigmoid'. Επιλέξτε 'rbf'
  - *gamma = η παράμετρος* γ *του πυρήνα RBF* $k(x,y)=\exp\left(-\gamma\|x-y\|^2\right)$
  - *degree = η παράμετρος d του πολυωνυμικού πυρήνα* $k(x,y)=\left(\gamma\, x^T y+r\right)^d$
  - *coef0 = είτε η παράμετρος r του πολυωνυμικού πυρήνα* $k(x,y)=\left(\gamma\, x^T y+r\right)^d$ *, είτε η παράμετρος r του σιγμοειδούς πυρήνα* $k(x,y)=\tanh\left(\gamma\, x^T y+r\right)$
  - ...και άλλες λοιπές παράμετροι (βλ. Documentation)
- Εκπαιδεύστε το δίκτυο που φτιάξατε χρησιμοποιώντας τη συνάρτηση fit() με εισόδους το μοντέλο, τον πίνακα των προτύπων εκπαίδευσης (xtrain), και το διάνυσμα των στόχων εκπαίδευσης (ttrain)
- Αφού εκπαιδεύσετε το μοντέλο κάνετε ανάκληση χρησιμοποιώντας τη συνάρτηση predict() με είσοδο τον πίνακα των προτύπων ελέγχου (xtest)
- Το διάνυσμα που πήρατε στην έξοδο είναι το $predict_{test}$.
- Υλοποιήστε τη συνάρτηση regrevaluate() με τρεις εισόδους και μια έξοδο ως εξής:

```
def regrevaluate( t, predict, criterion ):
# Είσοδοι:
#   t : διάνυσμα με τους πραγματικούς στόχους (πραγματικοί αριθμοί)
#   predict : διάνυσμα με τους εκτιμώμενους στόχους (πραγματικοί αριθμοί)
#   criterion : text-string με τις εξής πιθανές τιμές:
#       'mse'
#       'mae'
# Έξοδος value : η τιμή του κριτηρίου που επιλέξαμε.
```

- Αν criterion='mse' τότε

$$\text{value}=\frac{1}{n}\sum_{i=1}^{n}\left[t(i)-predict(i)\right]^2$$

- Αν criterion='mae' τότε

$$\text{value} = \frac{1}{n}\sum_{i=1}^{n}\left|t(i) - predict(i)\right|$$

end for fold

8. Μετά το τέλος του Cross-Validation loop υπολογίστε και σώστε σε array τα εξής:
    1. τη μέση τιμή του MSE για όλα τα folds
    2. τη μέση τιμή του MAE για όλα τα folds

end for C

end for gamma

9. Βρείτε τις τιμές του ζεύγους (gamma,C) που δίνει το μικρότερο μέσο MSE και τυπώστε κατάλληλο μήνυμα στην οθόνη.
10. Παρομοίως για την τιμή του ζεύγους (gamma,C) που δίνει το μικρότερο μέσο MAE.
11. Εκτελέστε ξανά το fold=1 με τις βέλτιστες τιμές των παραμέτρων gamma και C οι οποίες δίνουν τη μικρότερη τιμή μέσου MSE. Στο figure(1) τυπώστε το εξής γράφημα:
    a. δείξτε με μπλε γραμμή τους πραγματικούς στόχους $t_{test}(i)$ για όλα τα πρότυπα του test set
    b. δείξτε με κόκκινες τελείες τους εκτιμώμενους στόχους $predict_{test}(i)$ για όλα τα πρότυπα του test set
12. Επαναλάβετε όλα τα παραπάνω βήματα όμως αντί για SVR χρησιμοποιήστε νευρωνικό δίκτυο MLP δύο στρωμάτων. Στο βήμα 7, αντί για ένα διπλό loop χρησιμοποιήστε ένα μονό loop με το πλήθος των κρυφών νευρώνων N:

    Για κάθε *N*=[….]

    Πχ χρησιμοποιήστε τιμές N=[5,10,20,30,40,50]. Μέσα στο Cross-Validation loop δημιουργήστε το νευρωνικό μοντέλο με την κλάση MLPRegressor, η οποία είναι εντελώς ανάλογη με την MLPClassifier που χρησιμοποιήθηκε στο εργαστήριο 4. Κατόπιν εκπαιδεύστε το μοντέλο με την fit() και τέλος κάνετε ανάκληση με την predict() όπως και στο εργαστήριο 4. Χρησιμοποιήστε την regrevaluate() για να κάνετε αποτίμηση της επίδοσης του μοντέλου.
13. Βρείτε τις τιμές του N που δίνουν το μικρότερο μέσο MSE και το μικρότερο μέσο MAE και τυπώστε κατάλληλα μηνύματα στην οθόνη. Συγκρίνετε τα αποτελέσματα με τα αντίστοιχα του SVR (βήματα 10, 11).
14. Εκτελέστε ξανά το fold=1 με τη βέλτιστη τιμή της παραμέτρου N η οποία δίνει τη μικρότερη τιμή μέσου MSE. Στο figure(2) τυπώστε το εξής γράφημα:
    a. δείξτε με μπλε γραμμή τους πραγματικούς στόχους $t_{test}(i)$ για όλα τα πρότυπα του test set
    b. δείξτε με κόκκινες τελείες τους εκτιμώμενους στόχους $predict_{test}(i)$ για όλα τα πρότυπα του test set