ΔΔΠΜΣ "Προηγμένα Συστήματα Υπολογιστών και Επικοινωνιών", 2019-2020
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

1ο Σύνολο ασκήσεων για το μάθημα "Τεχνικές Μηχανικής Μάθησης"

Έκδοση 1.0

Εισαγωγικά

Οι παρακάτω ασκήσεις είναι προαιρετικές και λειτουργούν προσθετικά στη βαθμολογία του μαθήματος, μέχρι 1.5 βαθμό.

Παραδοτέα: Συνοπτική αναφορά με απαντήσεις στα ζητούμενα της εργασίας, καθώς και ο σχετικός κώδικας σε γλώσσα python (με επαρκή σχόλια παρακαλώ).

Άσκηση 1: Στην άσκηση αυτή θα εφαρμόσουμε τη μέθοδο γραμμικής παλινδρόμησης στο πρόβλημα της μοντελοποίησης της αντοχής του σκυροδέματος σε συμπίεση. Το σκυρόδεμα (τσιμέντο) είναι το πιο σημαντικό υλικό στην κατασκευή κτιρίων, και η αντοχή του στη συμπίεση είναι μη γραμμική συνάρτηση της ηλικίας του και των συστατικών του. Στο πρόβλημα αυτό προσπαθούμε να εκπαιδεύσουμε ένα μοντέλο της αντοχής του σκυροδέματος στη συμπίεση σαν συνάρτηση των χαρακτηριστικών του.

Σας δίνεται το σύνολο δεδομένων "Concrete_Data.csv", σε μορφή Comma Separated Values (CSV). Επίσης δίνεται αρχείο "Concrete_Readme.txt" με πληροφορίες για το σύνολο δεδομένων καθώς και το αρχείο "Concrete_Data.xls" που έχει το όνομα του κάθε χαρακτηριστικού (έχει αφαιρεθεί από το Concrete_Data.csv για ευκολία). Κάθε γραμμή του αρχείου αντιστοιχεί σε ένα δείγμα σκυροδέματος και οι στήλες αντιστοιχούν σε μετρήσεις του δείγματος (χαρακτηριστικά).

Η μεταβλητή που προσπαθούμε να μοντελοποιήσουμε αντιστοιχεί στην τελευταία στήλη, και είναι η "Concrete compressive strength", που αντιστοιχεί στην αντοχή του σκυροδέματος στην πίεση και δίνεται σε MPa. Αρχικά θα χρησιμοποιήσουμε το 70% του συνόλου δεδομένων για εκπαίδευση (με τη σειρά που δίνεται) και το υπόλοιπο 30% για αξιολόγηση.

Ζητούνται τα παρακάτω:

1. Αξιολογήστε την επίδοση της γραμμικής παλινδρόμησης ελαχίστων τετραγώνων (Ordinary Least Squares regression), καθώς και της γραμμικής παλινδρόμησης Ridge και LASSO. Πειραματιστείτε με διαφορετικές τιμές του βάρους ομαλοποίησης. Παρουσιάστε συνοπτικά τα αποτελέσματα της αξιολόγησης με βάση το μέσο τετραγωνικό σφάλμα (MSE),

το μέσο απόλυτο σφάλμα (MAE) και το μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE).

$$MSE = \frac{1}{N}\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$$

$$MAE = \frac{1}{N}\|\mathbf{y} - \hat{\mathbf{y}}\|_1$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}|\frac{y(i) - \hat{y}(i)}{y(i)}|$$

2. Στο προηγούμενο ερώτημα επιλέξατε το βάρος ομαλοποίησης (υπερπαράμετρος `alpha` στο scikit-learn) εξετάζοντας τα αποτελέσματα στο σύνολο αξιολόγησης. Ποιο μειονέκτημα έχει αυτή η προσέγγιση; Μπορείτε να προτείνετε άλλες στρατηγικές επιλογής της υπερπαραμέτρου;

3. Επαναλάβετε το βήμα 1, με τη διαφορά ότι η αξιολόγηση θα γίνει ως εξής: Επιλέγετε με τυχαίο τρόπο το 70% του συνόλου δεδομένων για εκπαίδευση και το υπόλοιπο 30% για αξιολόγηση και υπολογίζετε τις μετρικές αξιολόγησης. Η διαδικασία αυτή επαναλαμβάνεται 10 φορές και ως αποτέλεσμα δίνετε το μέσο όρο και την τυπική απόκλιση της κάθε μετρικής. Συμφωνούν τα αποτελέσματα με αυτά του βήματος 2;

4. Δεδομένης της μη γραμμικότητας της συνάρτησης που προσπαθούμε να μοντελοποιήσουμε, αξίζει να αξιολογήσουμε και πιο εκφραστικά μοντέλα γραμμικής παλινδρόμησης με πολυωνυμικούς όρους των χαρακτηριστικών. Υλοποιήστε συνάρτηση

   `test_poly_regression(X_train, y_train, X_test, y_test, n=2)`

   Η οποία θα δέχεται ως είσοδο ένα σύνολο εκπαίδευσης (`X_train` πίνακας σχεδιασμού του συνόλου εκπαίδευσης και `y_train` η εξαρτημένη μεταβλητή), ένα σύνολο αξιολόγησης (`X_test`, `y_test`), και έναν βαθμό πολυωνύμου $n \geq 1$. Η συνάρτηση θα πρέπει να δημιουργεί ένα νέο σύνολο χαρακτηριστικών που αποτελείται από τα αρχικά χαρακτηριστικά και εκδοχές τους υψωμένες σε δυνάμεις έως n. Συγκεκριμένα αν $\mathbf{X}$ είναι το αρχικό σύνολο, η συνάρτηση δημιουργεί το σύνολο

   $$\mathbf{X}_n = \begin{bmatrix}\mathbf{X} & \mathbf{X}^2 & \dots & \mathbf{X}^n\end{bmatrix}$$

   Αυτό συμβαίνει τόσο για το σύνολο εκπαίδευσης όσο και για το σύνολο αξιολόγησης. Έπειτα, η συνάρτηση εκπαιδεύει και αξιολογεί μοντέλα γραμμικής παλινδρόμησης στα σύνολα που προκύπτουν. Εκτελέστε τη διαδικασία αυτή για $n = 1$ έως $n = 10$. Μπορείτε να χρησιμοποιήσετε οποιονδήποτε από τους τρόπους αξιολόγησης των προηγούμενων υποερωτημάτων (σταθερό σύνολο εκπαίδευσης ή τυχαία επιλογή επαναληπτικά).

   Τι παρατηρείτε; Με βάση αυτά τα αποτελέσματα, ποιο μοντέλο θα επιλέγατε για πρακτική εφαρμογή; Επίσης, ποια είναι τα πλεονεκτήματα και ποια τα μειονεκτήματα των μοντέλων με υψηλότερους βαθμούς πολυωνύμου, n;

Άσκηση 2: Σας δίνεται το σύνολο δεδομένων Cover Type (αρχείο "covtype.data") στο οποίο προσπαθούμε να προβλέψουμε τον τύπο βλάστησης που καλύπτει μία δασική έκταση με βάση τα χαρακτηριστικά της. Το σύνολο δεδομένων περιγράφεται στο αρχείο "covtype.info". Θεωρήστε ότι οι πρώτες 15120 εγγραφές χρησιμοποιούνται για εκπαίδευση και οι υπόλοιπες για αξιολόγηση.

1. Εξετάστε το σύνολο δεδομένων. Ποιες είναι οι κατηγορικές μεταβλητές και πως αναπαρίστανται στο σύνολο δεδομένων; Δώστε συγκεκριμένα παραδείγματα.

2. Κατασκευάστε ένα μοντέλο λογιστικής παλινδρόμηση χρησιμοποιώντας τον αλγόριθμο επίλυσης "LBFGS" με μέγιστο αριθμό 10000 επαναλήψεων, σύγκλιση στο 10^{-3}, ομαλοποίηση L_2 και βάρος $C = 1.0$. Αξιολογήστε την ορθότητα (accuracy) του μοντέλου. Ο αλγόριθμος θα αργήσει να συγκλινει, οπότε θέστε την παράμετρο `verbose=1` ώστε να παρακολουθείτε την πρόοδό του. Επαναλάβετε τη διαδικασία με διαφορετικούς αλγόριθμους επίλυσης και επιλογές παραμέτρων της συνάρτησης `LogisticRegression` του scikit-learn και σχολιάστε σχετικά με το πόσο ευαίσθητη είναι η διαδικασία της εκπαίδευσης στην επιλογή παραμέτρων.

3. Όπως στο υποερώτημα 1, με χρήση Linear Discriminant Analysis (χωρίς την ανάλυση ευαισθησίας στις παραμέτρους).

4. Συγκρίνετε τα αποτελέσματα και την ταχύτητα σύγκλισης των 2 μοντέλων. Επίσης σχολιάστε την ευκολία με την οποία καταλήξατε σε μοντέλα με παρόμοια επίδοση στη μία και στην άλλη περίπτωση. Το παράδειγμα είναι διδακτικό: Στους αλγόριθμους που έχουν περιορισμένο θεωρητικό υπόβαθρο συχνά πρέπει να πειραματιστούμε χωρίς καθοδήγηση για να πετύχουμε κάποιο καλό αποτέλεσμα (από την άλλη όμως συχνά είναι η μόνη μας επιλογή).

Άσκηση 3: Σας δίνεται ένα σύνολο δεδομένων από τη σειρά "Game of Thrones". Κάθε γραμμή του συνόλου δεδομένων αφορά έναν χαρακτήρα και αντιστοιχεί στα χαρακτηριστικά του. Στόχος μας είναι να εκπαιδεύσουμε ένα μοντέλο που από τα χαρακτηριστικά προβλέπει αν ο χαρακτήρας είναι ζωντανός ή όχι.

1. Τι τύπος προβλήματος μηχανικής μάθησης είναι αυτός;

2. Περιγράψτε τη διαδικασία προεπεξεργασίας των δεδομένων που θα ακολουθούσατε πριν χρησιμοποιήσετε τα δεδομένα για την εκπαίδευση κάποιου μοντέλου