Εργασία 1 – Επιβλεπόμενη Μάθηση

Μέρος Ι - Λογιστική Παλινδρόμηση

Ιδιωτική κλινική θέλει να κατασκευάσει σύστημα πρόβλεψης διαβήτη, ώστε να αυτοματοποιήσει την ανίχνευση διαβητικών ασθενών. Επιπλέον, διαθέτει στη διάθεση του ένα σύνολο εξετάσεων 768 ασθενών, που περιλαμβάνουν:

1. **Pregnancies**: Εγκυμοσύνες
2. **Glucose:** Συγκέντρωση γλυκόζης
3. **Blood Pressure:** Αρτηριακή πίεση
4. **Skin Thickness:** Πάχος δέρματος στους τρικέφαλους
5. **Insulin:** Συγκέντρωση ινσουλίνης
6. **BMI:** Δείκτης μάζας-σώματος
7. **Diabetes Pedigree Function:** Δείκτης που εκφράζει πιθανότητα εμφάνισης διαβήτη με βάση το οικογενειακό ιστορικό
8. **Age:** Ηλικία
9. **Outcome:** Αποτέλεσμα εξετάσεων (0: Αρνητικός, 1: Θετικός)

Ερωτήματα

1. Φορτώστε το σύνολο δεδομένων diabetes.csv σε ένα DataFrame μέσω της βιβλιοθήκης pandas. Στη συνέχεια, περιγράψτε το κάθε χαρακτηριστικό με μέση τιμή, τυπική απόκλιση, ελάχιστη και μέγιστη τιμή (df.describe()). Τέλος, δημιουργήστε το ιστόγραμμα (Histogram) για κάθε χαρακτηριστικό. Για το Outcome, να δημιουργηθεί ραβδόγραμμα (Bar Plot), εφόσον έχει μόνο 2 τιμές.
2. Θεωρείτε πως η ποιότητα των δεδομένων είναι καλή ή κακή? Αιτιολογείστε, αξιοποιώντας τις πληροφορίες από ερώτημα (1). Δώστε τουλάχιστον 2 επιχειρήματα.
3. Τι κατανομή ακολουθεί η μεταβλητή Age? Είναι καλή η κατανομή αυτή για την κατασκευή της συγκεκριμένης εφαρμογής? Αιτιολογείστε.
4. Σύμφωνα με κλινικές μελέτες, αν κάποιος ασθενής έχει υψηλά επίπεδα γλυκόζης, είναι πολύ πιθανό να εμφανίσει διαβήτη. Να Δείξετε πως διαπιστώνεται αυτό από τα δεδομένα.
5. Δημιουργήστε Numpy arrays με inputs (x) και targets (y), όπου στο x περιλαμβάνονται όλα τα χαρακτηριστικά (εκτός του outcome) και y το outcome. Στη συνέχεια, χωρίστε το σύνολο δεδομένων σε σύνολα εκπαίδευσης-επικύρωσης (train-validation) με ποσοστό 70-30%, χρησιμοποιώντας ως seed (random state) το 0. Αναφέρετε το πλήθος των παραδειγμάτων εκπαίδευσης και επικύρωσης. Θα χρειαστείτε να φορτώσετε τη συνάρτηση train_test_split της scikit-learn.
6. Εκπαιδεύστε ταξινομητή (Classifier) Λογιστικής Παλινδρόμησης (Logistic Regression) στο train set και μετρήστε την ακρίβεια (accuracy score) του στο test set. Χρησιμοποιήστε ως seed το 0. Μπορείτε να συμβουλευτείτε το documentation της scikit-learn: https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html.
7. Τι θεωρείτε πως είναι χειρότερο για τον ταξινομητή σας στη συγκεκριμένη εφαρμογή, να προβλέπει ότι κάποιος ασθενής έχει διαβήτη, χωρίς να έχει ή ότι κάποιος ασθενής δεν έχει διαβήτη, ενώ έχει? Αιτιολογήστε.
8. Δουλεύει καλά ο ταξινομητής σας για όλες τις ηλικίες? Παρουσιάστε ραβδόγραμμα με την ακρίβεια ανά ομάδα ηλικιών 0 ως 25, 25 ως 50 και >50 του test set.

9. Ενδεχομένως, το μοντέλο σας να χάνει ακρίβεια λόγω υπέρ-προσαρμογής στο train set. Επαναλάβετε το ερώτημα 5 χρησιμοποιώντας τεχνικές Regularization (**penalty**): l1, l2 και elastic net. Εμφανίστε πινακάκι με την ακρίβεια (accuracy_score) για κάθε περίπτωση.
10. Επαναλάβετε τις διαδικασίες 5-6 (με μία for loop), χρησιμοποιώντας seed από 0 ως 9. Υπολογίστε μέσο όρο και τυπική απόκλιση της ακρίβειας σας.

Μέρος ΙΙ - Γραμμική Παλινδρόμηση

Η ίδια ιδιωτική κλινική θέλει να εντοπίζει πιθανούς διαβητικούς μέσω μιας έξυπνης εφαρμογής τηλεφώνου. Οι χρήστες θα μπορούν να εισάγουν δεδομένα που απαιτούν ελάχιστο ιατρικό εξοπλισμό, όπως **εγκυμοσύνες, αρτηριακή πίεση, BMI και ηλικία** και η εφαρμογή θα προβλέπει το επίπεδο ινσουλίνης στο αίμα. Αν είναι πάνω από 170, τότε θα τους προτείνει να προσέλθουν για εξετάσεις.

Ερωτήματα

1. Φορτώστε το σύνολο δεδομένων diabetes.csv. Δημιουργήστε Numpy arrays με inputs (x) και targets (y), όπου x: (Pregnancies, Blood Pressure, BMI, Age) και y η μεταβλητή Insulin. Χωρίστε το σύνολο δεδομένων σε σύνολα εκπαίδευσης-επικύρωσης (train-validation) με ποσοστό 70-30% με 0 seed.
2. Χρησιμοποιείστε Γραμμική παλινδρόμηση της scikit-learn: https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.LinearRegression.html, ώστε να προβλέψετε την ποσότητα ινσουλίνης στο test set και μετρήστε την ακρίβεια με κατάλληλη μετρική. Ποια μετρική είναι καταλληλότερη για την αξιολόγηση του μοντέλου: MSE ή MAE? Αιτιολογείστε.
3. Θέλουμε να ερευνήσουμε πιο χαρακτηριστικό από τα 4 επηρεάζει περισσότερο την ινσουλίνη. Περιγράψτε πως μπορούμε να χρησιμοποιήσουμε τα βάρη (coefficients) της παλινδρόμησης για να εξετάσουμε τη σημαντικότητα κάθε χαρακτηριστικού, καθώς και τα ενδεχόμενα προβλήματα αυτής της μεθόδου.
4. Επαναλάβετε το ερώτημα (3) χρησιμοποιώντας Lasso Regression: https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.Lasso.html (L1). Δοκιμάστε τιμές (0.2, 0.4, 0.6, 0.8, 1.0) για το βάρος λ (alpha στην scikit-learn) και κατασκευάστε πινακάκι με την ακρίβεια, χρησιμοποιώντας τη μετρική του ερωτήματος 2.
5. Χρησιμοποιήστε το καλύτερο από τα 5 μοντέλα Lasso του ερωτήματος (4) και συγκρίνετε τα coefficients του Linear με του Lasso που επιλέξατε. Τελικά, ποιο είναι το σημαντικότερο χαρακτηριστικό?
6. Επαναλάβετε το ερώτημα 2, αφαιρώντας από το X το χαρακτηριστικό με τη χαμηλότερη βαρύτητα που βρήκατε και συγκρίνετε την επίδοση του μοντέλου σας με το (2).

Οδηγίες

- Χρησιμοποιείστε την πλατφόρμα Google Colab για την υλοποίηση της άσκησης.
- Τα plots/πινακάκια να εμφανίζονται επάνω στο Colab. Επίσης, μπορείτε να εισάγετε κελιά για κείμενο και πινακάκια όπως φαίνεται στον παρακάτω σύνδεσμο: https://colab.research.google.com/notebooks/markdown_guide.ipynb
- Δώστε έμφαση στην παρουσίαση της εργασίας. Copy-Paste από το ChatGPT θα αγνοούνται.
- **Στο elearning θα υποβάλλετε το link της εργασίας σας στο Google Colab.**