Εργασία 5 – Feature Importance, Dimensionality Reduction, KNN

Μέρος 1 – Feature Importance

1. Φορτώστε το σύνολο δεδομένων wine-full.csv και χωρίστε το σε train-test (80-20% και random_state=0). Διαχωρίστε τα δεδομένα σε input-targets, όπου target η μεταβλητή Quality.
2. Στη συνέχεια, εκπαιδεύστε Lasso (Linear+L1) και Random Forest Regressor (random_state=0) και υπολογίστε τις μετρικές MAE, R^2 στο test set για κάθε Regressor. Να συγκρίνετε την απόδοση αυτών των 2.
3. Εμφανίστε ραβδόγραμμα με τα 5 πιο σημαντικά features για κάθε Regressor.
4. Εξετάστε πως επηρεάζει κάθε feature την απόδοση ενός μοντέλου χρησιμοποιώντας τη συνάρτηση *permutation_importance* της scikit-learn: https://scikit-learn.org/1.5/modules/permutation_importance.html. Ορίστε ως estimator το Linear Regression. Στη συνέχεια, να εμφανίσετε τη μέση τιμή και τυπική απόκλιση της σημαντικότητας κάθε σκορ με βάση τη παραπάνω μέθοδο.
5. Ας υποθέσουμε πως η ανάλυση κοστίζει ακριβά, οπότε μπορούμε να αναλύσουμε μέχρι 3 χαρακτηριστικά ενός δείγματος. Ποιά είναι τα 3 πιο σημαντικά χαρακτηριστικά με βάση τα permutation_importance, lasso και random forest? Αν υπάρχει ασυμφωνία μεταξύ των μεθόδων, ποιά άλλη μέθοδο θα μπορούσατε να εφαρμόσετε, ώστε να καταλήξετε στα 3 πιο σημαντικά χαρακτηριστικά?

Μέρος 2 – Dimensionality Reduction, KNN

1. Φορτώστε το dataset 70,000 ασπρόμαυρων χειρόγραφων ψηφίων (0 ως 9) μεγέθους 28x28 της MNIST:

2. Εμφανίστε σε 1 πλοτ 5 τυχαία ψηφία από κάθε κλάση (Σύνολο 5*10=50 εικόνες).
3. Θέλουμε να κανονικοποιήσουμε τις εισόδους, ώστε να πετύχουμε καλύτερη ακρίβεια. Ποια μέθοδο κανονικοποίησης για να εκπαιδεύσετε KNN? Αιτιολογήστε. Στη συνέχεια, εφαρμόστε τη μέθοδο που επιλέξατε και μετατρέψτε τα inputs σε διανύσματα.
4. Εκπαιδεύστε 4 ταξινομητές KNN με $k = 5,15,51,\sqrt{60000}$ στο train set και υπολογίστε την ακρίβεια (accuracy score) τους στο test set. Στη συνέχεια, να δημιουργήσετε line plot για σύγκριση του K σε σχέση με την ακρίβεια.
5. Είναι σημαντικά όλα τα pixel της εικόνας ενός ψηφίου? Εναλλακτικά, όλα τα pixels των εικόνων βοηθούν τον ταξινομητή? Αιτιολογήστε με βάση το διάγραμμα του ερωτήματος 2.
6. Εφαρμόστε PCA, ώστε να μειώσετε τη διάσταση των δεδομένων σε 300 χαρακτηριστικά (n_components=300). Στη συνέχεια, επαναλάβετε το ερώτημα 4. **Προσοχή**! **Θέλουμε να καλέσουμε τη συνάρτηση *pca.fit_transform* για το train set, αλλά τη συνάρτηση *pca.transform* για το test set. Τι θα γινόταν αν εφαρμόζαμε την fit transform στα x_train και x_test ξεχωριστά ή σε όλα μαζί?**
7. Τι παρατηρείτε ως προς τον χρόνο και την ακρίβεια του KNN ύστερα από την εφαρμογή του PCA?

8. Εφαρμόστε LDA μειώνοντας τα χαρακτηριστικά σε 2. Στη συνέχεια:
    - Να εμφανίσετε τις μετρικές με τη χρήση του *classification_report* της scikit-learn.
    - Να δημιουργήσετε 2D scatter plot με τα x_train ύστερα από σχηματισμό LDA. Δώστε διαφορετικό χρώμα σε κάθε ψηφίο. Τι παρατηρείτε για τον διαχωρισμό τους?