Τμήμα Πληροφορικής και Τηλεματικής
Χαροκόπειο Πανεπιστήμιο
ΥΠ23 Τεχνητή Νοημοσύνη

Εργασία 2: Γραμμικά μοντέλα στο scikit-learn

Έκδοση 2024-1.1

Διδάσκων: Χρήστος Δίου

1 Εισαγωγή

Στην άσκηση αυτή θα υλοποιήσουμε και θα δοκιμάσουμε μία απλή εκδοχή ενός μοντέλου γραμμικής ταξινόμησης με το scikit-learn. Η άσκηση σας καθοδηγεί βήμα-βήμα στην υλοποίηση.

Παραδοτέα

Για την υποβολή της εργασίας, θα χρειαστεί να υποβληθεί ο κώδικας σε γλώσσα python, με επαρκή σχόλια και μια συνοπτική αναφορά που να απαντάει στα ερωτήματα της εργασίας, και ενδεχομένως να προσφέρει λεπτομέρειες σχετικά με την υλοποίηση. Ο κώδικας θα πρέπει να μπορεί να εκτελεστεί σε περιβάλλον python3 με πρόσφατες εκδόσεις των βιβλιοθηκών `numpy` και `scikit-learn`. Μπορείτε να ανταλλάσσετε απόψεις και ιδέες μεταξύ σας, αλλά τελικά την άσκηση θα πρέπει να την υλοποιήσετε μόνοι σας.

2 Σύνολο δεδομένων

Σας δίνεται ένα σύνολο δεδομένων που δημοσιεύτηκε το 2019 [1] και το οποίο περιέχει χαρακτηριστικά που αφορούν χρήστες ενός διαδικτυακού καταστήματος. Κάθε διάνυσμα χαρακτηριστικών σχετίζεται με ένα user session. Στόχος μας είναι να εκπαιδεύσουμε ένα μοντέλο το οποίο θα προβλέπει αν ο χρήστης πρόκειται να αγοράσει από το κατάστημα (στήλη "Revenue"). Στον Πίνακα 1 του [1] που σας δίνεται μπορείτε να βρείτε λεπτομέρειες σχετικά με τις μεταβλητές.

Ερώτημα 1: Διερεύνηση δεδομένων (`eda.py`)

Υλοποιήστε ένα Python script το οποίο θα ονομάσετε `eda.py`, το οποίο φορτώνει τα δεδομένα χρησιμοποιώντας τη βιβλιοθήκη Pandas (με τη συνάρτηση `read_csv`) και σας βοηθάει να απαντήσετε στα εξής ερωτήματα:

1. Πόσες είναι οι εγγραφές του συνόλου δεδομένων;
2. Σε τι ποσοστό από αυτές οι χρήστες αγόρασαν τελικά;
3. Ποια είναι η ευστοχία (accuracy) ενός μοντέλου το οποίο προβλέπει πάντα ότι ο χρήστης δε θα αγοράσει (ανεξαρτήτως των χαρακτηριστικών του);

Επιπλέον προαιρετικά, μπορείτε να προσθέσετε διαγράμματα που να δείχνουν την κατανομή μεταβλητών καθώς και τη σχέση τους με τη μεταβλητή στόχο.

Ερώτημα 2: `prepare_data`

Σ' αυτό το ερώτημα θα υλοποιήσουμε μια συνάρτηση η οποία φορτώνει και προετοιμάζει το σύνολο δεδομένων. Συγκεκριμένα η συνάρτηση αυτή είναι η `prepare_data`:

```
X_train, X_test, y_train, y_test = prepare_data(df, train_size=None,
                                                shuffle=True, random_state=None)
```

όπου

- `df` είναι Pandas DataFrame το οποίο περιέχει τα δεδομένα, όπως διαβάζονται από τη συνάρτηση `read_csv`
- `train_size`, `shuffle`, `random_state` όπως χρησιμοποιούνται από την `train_test_split` του scikit-learn

Η συνάρτηση πρέπει να κάνει τα εξής:

- Θα αφαιρεί τα χαρακτηριστικά `Month`, `Browser`, `OperatingSystems` (αποφασίζουμε να τα αγνοήσουμε εδώ για απλούστευση)
- Θα μετατρέπει τις boolean τιμές σε αριθμητικές
- Θα εφαρμόζει One-hot encoding στις μεταβλητές `Region`, `TrafficType`, `VisitorType` (πχ μέσω της `get_dummies` της Pandas)
- Θα χωρίζει τη μεταβλητή `Revenue` που είναι ο στόχος από τις υπόλοιπες
- Θα χωρίζει το σύνολο δεδομένων σε σύνολο εκπαίδευσης και δοκιμής, χρησιμοποιώντας την `train_test_split` του scikit-learn, χρησιμοποιώντας τις παραμέτρους που δόθηκαν στην είσοδο
- Θα επιστρέφει το σύνολο εκπαίδευσης και δοκιμής, όπως στο παραπάνω αρχέτυπο

Αποθηκεύστε τη συνάρτηση σε ένα ένα αρχείο `linear_classification.py`. Όλα τα επόμενα ερωτήματα θα πρέπει να υλοποιηθούν στο ίδιο αρχείο.

Ερώτημα 3: Προετοιμασία δεδομένων

Στο ίδιο αρχείο, διαβάστε το σύνολο δεδομένων και καλέστε την `prepare_data` για 70%-30% χωρισμό σε σύνολο εκπαίδευσης και δοκιμής και σπόρο 42 (μέσω της μεταβλητής `random_seed`) ώστε να μπορείτε να επαναλάβετε τα πειράματα.

Επιπλέον, εφαρμόστε γραμμική κανονικοποίηση στα δεδομένα, πχ μέσω του `MinMaxScaler()`. Προσοχή, υπολογίζουμε τις παραμέτρους για το μετασχηματισμό τιμών μόνο στο σύνολο εκπαίδευσης, και έπειτα εφαρμόζουμε το μετασχηματισμό, τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής. Δε χρειάζεται να μετασχηματίσουμε τη μεταβλητή - στόχο.

Από εδώ και πέρα θα χρησιμοποιούμε μόνο τα κανονικοποιημένα δεδομένα.

Ερώτημα 4: Υλοποίηση μοντέλου

Στο ίδιο αρχείο, δημιουργήστε ένα αντικείμενο της κλάσης `LogisticRegression` το οποίο δε θα έχει παράμετρο `penalty`, ενώ φροντίστε να αυξήσετε τον μέγιστο αριθμό επαναλήψεων ώστε να εξασφαλίσετε ότι ο αλγόριθμό σας θα συγκλίνει. Προσέξτε και τα αντίστοιχα προειδοποιητικά μηνύματα.

Έχοντας εκπαιδεύσει το μοντέλο, εφαρμόστε το τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμής, ώστε να καταγράψετε τις προβλέψεις του, $\hat{y}_{\text{train}}$ και $\hat{y}_{\text{test}}$, αντίστοιχα.

Ερώτημα 5: Αξιολόγηση μοντέλου

Στη συνέχεια υπολογίστε και εκτυπώστε:

- Την ευστοχία του μοντέλου στο σύνολο εκπαίδευσης
- Την ευστοχία του μοντέλου στο σύνολο δοκιμής
- Τον πίνακα σύγχυσης (μπορείτε να χρησιμοποιήσετε τη συνάρτηση `confusion_matrix`, αν θέλετε)

Απαντήστε στα ακόλουθα ερωτήματα:

- Πως ερμηνεύετε τον πίνακα σύγχυσης;
- Τι τροποποιήσεις ή επιπλέον πειράματα θα υλοποιούσατε ώστε να βελτιώσετε το μοντέλο σας;

Καλή επιτυχία

Αναφορές

[1] Cemal Okan Sakar, Suleyman Olcay Polat, Mete Katircioglu, and Yomi Kastro. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and lstm recurrent neural networks. *Neural Computing and Applications*, 31:6893 – 6908, 2018.