Τεχνικές Εξόρυξης Δεδομένων
Εαρινό Εξάμηνο 2016-2017

2η Άσκηση, Ημερομηνία παράδοσης: 05-06-2017
Ομαδική Εργασία (2 Ατόμων)

Σκοπός της εργασίας

Σκοπός της εργασίας είναι η κατανόηση και η εξερεύνηση (data exploration) των δεδομένων εισόδου καθώς, η αξιολόγηση των χαρακτηριστικών τους (feature selection). Η υλοποίηση της εργασίας θα γίνει στην γλώσσα προγραμματισμού Python (όπως και η 1η άσκηση) με την χρήση των εργαλείων/βιβλιοθηκών: *jupyter notebook*, *pandas*, *gensim* και *SciKit Learn*.

Περιγραφή

Η εργασία σχετίζεται με δεδομένα δανειοληπτικής ικανότητας τα οποία ανήκουν σε δυο κατηγορίες (Good/Bad). Ο στόχος σας είναι να χρησιμοποιήσετε τα διαθέσιμα features προκειμένου να δημιουργήσετε ένα classification μοντέλο το οποίο θα μπορεί να αποφασίζει εάν θα πρέπει να δοθεί δάνειο σε έναν πελάτη της τράπεζας που αιτείται δάνειο ή όχι.

Το διαθέσιμο dataset περιέχει ένα σύνολο στηλών που αφορούν τα διαφορετικά χαρακτηριστικά των δανειοληπτών, ενώ η τελευταία στήλη αναφέρεται στο εάν ο δανειολήπτης ήταν τελικά καλός ή κακός (1 = Good, 2 = Bad).

Μια εκτενής περιγραφή των δεδομένων παρέχεται από το διαθέσιμο document στο παρακάτω link.

Πιο συγκεκριμένα θα έχετε τα εξής αρχεία:

1. Train_set.csv (*800 instances*): Το αρχείο αυτό θα χρησιμοποιηθεί για να εκπαιδεύσετε τους αλγόριθμους σας. Επίσης αυτό το dataset θα το χρησιμοποιήσετε για την αξιολόγηση των features καθώς και την οπτικοποίηση τους.
2. test_set.csv (*200 instances*): Το αρχείο αυτό θα χρησιμοποιηθεί για να κάνετε προβλέψεις για νέα δεδομένα. Περιέχει όλα τα πεδία του αρχείου εκπαίδευσης εκτός από την τελευταία στήλη, το πεδίο αυτό θα κληθείτε να το εκτιμήσετε χρησιμοποιώντας αλγόριθμους κατηγοριοποίησης

Λήψη Dataset

Τα datasets είναι διαθέσιμα στο eclass.

Οπτικοποίηση των Δεδομένων

Στο συγκεκριμένο ερώτημα θα πρέπει να οπτικοποιήσετε τα διαφορετικά feautes του dataset. Πιο συγκεκριμένα θα πρέπει στην αναφορά σας για κάθε feature να προσθέσετε τα παρακάτω:

- Εάν το feature είναι **categorical**:
    - ένα *histogram* για κάθε είδος δανειολήπτες που έχουν χαρακτηριστεί **Good**
    - ένα *histogram* για κάθε είδος δανειολήπτες που έχουν χαρακτηριστεί **Bad**
- Εάν το feature είναι **numerical**:
    - ένα box plot για κάθε είδος δανειολήπτες που έχουν χαρακτηριστεί **Good**
    - ένα box plot για κάθε είδος δανειολήπτες που έχουν χαρακτηριστεί **Bad**

Τέλος θα πρέπει να περιγράψτε τι παρατηρείτε από τα plots που δημιουργήσατε. Ποιά features περιμένετε να είναι πιο χρήσιμα για την κατηγοριοποίηση των πελατών,

(hint) Θα ήταν χρήσιμο αν στο ίδιο plot για κάθε feature οπτικοποιήσετε και τους Good και τους Bad με διαφορετικό χρώμα.

Υλοποίηση Κατηγοριοποίησης (Classification)

Σε αυτό το ερώτημα θα πρέπει να δοκιμάσετε τις παρακάτω μεθόδους Classification:

- Support Vector Machines (SVM)
- Random Forests
- Naive Bayes

Επίσης θα πρέπει να αξιολογήσετε και να καταγράψετε την απόδοση κάθε μεθόδου χρησιμοποιώντας 10-fold Cross Validation χρησιμοποιώντας τη μετρική accuracy.

Επιλογή Features

Στο συγκεκριμένο ερώτημα θα πρέπει να αξιολογήσετε την ποιότητα των διαθέσιμων features σχετικά με την κατηγοριοποίηση των πελατών ως Good ή Bad, χρησιμοποιώντας τον καλύτερο Classifier που βρήκατε από το προηγούμενο ερώτημα.

Πιο συγκεκριμένα θα πρέπει να υπολογίσετε το Information Gain για κάθε feature. Στην συνέχεια να υπολογίσετε το accuracy του classifier, αφαιρώντας κάθε φορά και ένα feature.
Θα πρέπει να παρουσιάσετε:

- Σε ένα plot πως μεταβάλλεται το μέσο accuracy για 10-fold cross-validation καθώς αφαιρείτε features από τον Classifier.
- Σε έναν πίνακα το feature που επιλέξατε να αφαιρέσετε σε κάθε επανάληψη καθώς και το αντίστοιχο Information Gain.

(Hint) Σχετικά με τον υπολογισμό του Information Gain μπορείτε να χρησιμοποιήσετε τον ορισμό από το Wilkipedia από το παρακάτω link.
Τα numerical attributes μπορείτε να τα μετατρέψετε σε categorical στο συγκεκριμένο ερώτημα διακριτοποιώντας τα σε 5 bins.

Forum Επικοινωνίας
piazza

Θα πρέπει αναλυτικά να τεκμηριώσετε τα βήματα που ακολουθήσατε. Το report σας να μην ξεπερνάει τις 30 σελίδες.

Αρχεία Εξόδου

Ο κώδικας θα πρέπει για τα ερωτήματα που αφορούν το Classification θα πρέπει να δημιουργεί τα παρακάτω αρχεία

- EvaluationMetric_10fold.csv
- testSet_Predictions.csv

Το format των αρχείων EvaluationMetric_10fold.csv φαίνεται παρακάτω:

Statistic Measure	**Naive Bayes**	**Random Forest**	**SVM**
Accuracy			

Το format του αρχείου testSet_Predictions.csv, το οποίο θα περιέχει τις κατηγορίες των πελατών που δίνονται στο Test set φαίνεται παρακάτω:

Client_ID	Predicted_Label
1	Good
...	...
10	Bad
...	...

Για το αρχείο “testSet_Predictions.csv” θα πρέπει να χρησιμοποιηθεί αυστηρά η παραπάνω μορφοποίηση διαχωρίζοντας τα δυο πεδία με τον χαρακτήρα TAB (‘\t’) και επίσης θα πρέπει στην πρώτη γραμμή να υπάρχουν οι δυο επικεφαλίδες (Client_ID και Predicted_Label) και ακολούθως οι προβλέψεις του μοντέλου σας στις επόμενες γραμμές διευκρινίζοντας το Client_ID του πελάτη από το test set και το αντίστοιχο label.

Σχετικά με το παραδοτέο

Ο φάκελος που θα παραδώσετε θα έχει το όνομα Ass1_όνοματεπώνυμο1_ΑΜ1_ονοματεπώνυμο2_ΑΜ2. Ο φάκελος θα περιέχει:

1. ένα κείμενο με τον σχολιασμό στα πειράματα που κάνατε και στις μεθόδους που δοκιμάσατε σε μορφή PDF. Η αναφορά σας θα πρέπει να περιέχει και τους πίνακες με τα αποτελέσματα των αρχείων εξόδου.
2. τα ζητούμενα αρχεία εξόδου.
3. τα αρχεία κώδικα που γράψατε.

Το εκτενές κείμενο που θα παραδώσετε, θα περιέχει την περιγραφή των δοκιμών σας και οτιδήποτε σκεφτείτε για να δείξετε τι δοκιμές κάνατε, για ποιο λόγο έχουν τα συγκεκριμένα αποτελέσματα οι μέθοδοι που επιλέξατε, πως λειτουργούν αυτές οι μέθοδοι και σχολιασμό των αποτελεσμάτων σας. Όλες οι εργασίες θα αξιολογηθούν στη βάση της σωστής τεκμηρίωσης και στο βαθμό που υλοποιούν τα ζητούμενα της εργασίας, όχι με βάση την κατάταξη που επιτυγχάνουν στα αποτελέσματα της κατηγοριοποίησης δεδομένων.