Εξόρυξη Δεδομένων και Αλγόριθμοι Μάθησης

CEID_NE562

Πανεπιστήμιο Πατρών
Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής

Διδάσκοντες:

Χρήστος Μακρής, Βασίλειος Μεγαλοοικονόμου

Φοιτήτρια:

Καλαματιανού Δήμητρα(Α.Μ.: 1054406)

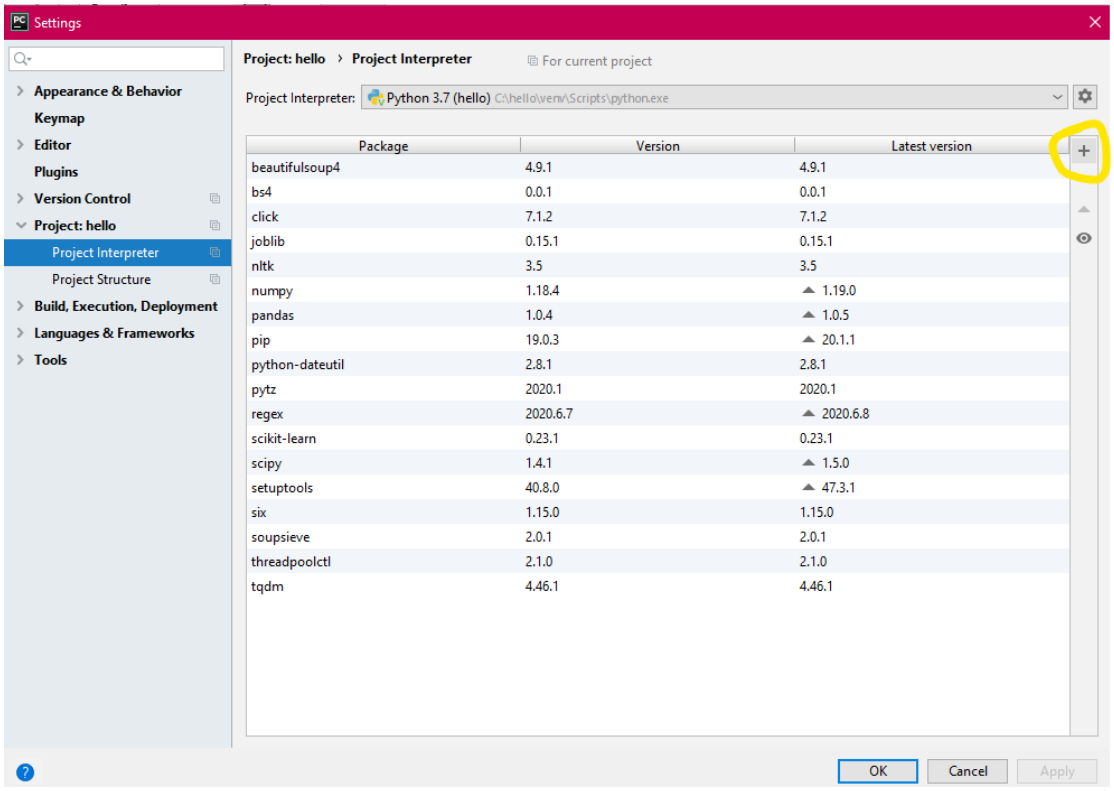
8ο εξάμηνο
2019-2020

Εγκατάσταση περιβάλλοντος υλοποίησης

Ο υπολογιστής χρειάζεται να έχει εγκατεστημένη κάποια έκδοση της **Python** για να μπορεί να τρέξει το πρόγραμμα γραμμένο σε αυτήν την γλώσσα. Κάποιος μπορεί να κατεβάσει όποια έκδοση της Python ακολουθώντας αυτόν τον σύνδεσμο: https://www.python.org/. Η υλοποίηση του project έγινε σε υπολογιστή με εγκατεστημένη την έκδοση Python 3.7.3.

Το **IDE** που χρησιμοποιήθηκε για την υλοποίηση του project είναι το PyCharm και συγκεκριμένα η free open-source (Community) edition. Η εγκατάσταση του είναι απλή, ακολουθώντας την σελίδα https://www.jetbrains.com/pycharm/ και επιλέγοντας Download εμφανίζεται η σελίδα που δίνει την δυνατότητα επιλογής operating system(WINDOWS, MAC, LINUX) καθώς και έκδοση του IDE(Professional ή Community). Πατώντας το κουμπί Download το πρόγραμμα κατεβαίνει στον υπολογιστή. Ανοίγοντας το κατεβασμένο αρχείο εμφανίζεται το setup του PyCharm, το οποίο προσφέρει επιλογές εγκατάστασης στον χρήστη.

Τα packages εγκαθίστανται στο PyCharm ακολουθώντας τα παρακάτω βήματα:
1. Άνοιγμα καρτέλας File μέσα από το IDE.
2. Επιλογή Settings μέσα από την καρτέλα.
3. Επιλογή Project Interpreter μέσα από το καινούργιο παράθυρο που ανοίγει.
4. Επιλογή του «+» όπως φαίνεται στην παρακάτω εικόνα.



5. Πληκτρολόγηση του ονόματος του επιθυμητού package.
6. Επιλογή της από την λίστα και επιλογή κουμπιού Install Package.
Εμφανίζεται σχετικό μήνυμα για την ολοκλήρωση της εγκατάστασης.

Για την ολοκλήρωση του project χρειάστηκε να εγκατασταθούν τα εξής:
pandas
numpy
scikit-learn
nltk

Σύντομη περιγραφή της διαδικασία υλοποίησης - Σχολιασμός των τελικών αποτελεσμάτων

Ερώτημα 1 – Α

Πριν ακόμη ξεκινήσω με το γράψιμο κώδικα έκανα μια έρευνα για τα ζητούμενα και για τις Υποδείξεις που υπήρχαν μετά τα ερωτήματα.

Στην συνέχεια ξεκίνησα να γράφω κώδικα για το (Α) υποερώτημα του *ερωτήματος 1*. Για να βελτιωθούν τα αποτελέσματα προσπάθησα να βελτιώσω τις παραμέτρους του vector machine (SVM) classifier, πιο συγκεκριμένα τις παραμέτρους C και gamma. Ως παράμετρος για το kernel έγινε χρήση του rbf γιατί είναι μια συνήθης συνάρτηση kernel για vector machine classification. Οι καλύτερες τιμές για C και gamma είναι συνήθως οι παρακάτω:

‘’The best performing gamma value is: 0.23’’
‘’The best performing C value is: 1.03’’

όπως προκύπτουν από τον κώδικα του ερωτήματος, με μια μικρή απόκλιση κάθε φορά.

Υπολογίζονται accuracy, macro avg, weighted avg και overall accuracy για να έχουμε μια γενική εικόνα για τις μετρικές f1-score, precision και recall.

Αποτελέσματα μετά την εκτέλεση του κώδικα:

- Για C=1.03, gamma=0.23

```
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         3
           4       1.00      0.00      0.00         9
           5       0.65      0.43      0.52       179
           6       0.43      0.79      0.55       151
           7       1.00      0.00      0.00        55
           8       1.00      0.00      0.00         3

    accuracy                           0.49       400
   macro avg       0.85      0.20      0.18       400
weighted avg       0.63      0.49      0.44       400

Overall Accuracy: 0.492

The best performing gamma value is:  0.23
The best performing C value is:  1.03
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         3
           4       1.00      0.00      0.00         9
           5       0.66      0.68      0.67       179
           6       0.54      0.70      0.61       151
           7       0.85      0.31      0.45        55
           8       1.00      0.00      0.00         3

    accuracy                           0.61       400
   macro avg       0.84      0.28      0.29       400
weighted avg       0.65      0.61      0.59       400

Overall Accuracy: 0.6075
```

Οι τιμές των C και gamma αλλάζουν γιατί κάθε φορά που τρέχουμε τον κώδικα το dataset χωρίζεται σε διαφορετικά training και test set (με αναλογία 75%-25%). Παρακάτω παραθέτω κάποιες πιθανές τιμές αυτών των παραμέτρων και τα αποτελέσματα των μετρικών που αυτές βγάζουν.

- Για C=1.03, gamma=0.17

```
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         2
           4       1.00      0.00      0.00        14
           5       0.72      0.44      0.54       185
           6       0.44      0.84      0.58       152
           7       1.00      0.00      0.00        45
           8       1.00      0.00      0.00         2

    accuracy                           0.52       400
   macro avg       0.86      0.21      0.19       400
weighted avg       0.66      0.52      0.47       400

Overall Accuracy: 0.52

The best performing gamma value is:  0.17
The best performing C value is:  1.03
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         2
           4       1.00      0.00      0.00        14
           5       0.65      0.74      0.69       185
           6       0.54      0.59      0.56       152
           7       0.67      0.31      0.42        45
           8       1.00      0.00      0.00         2

    accuracy                           0.60       400
   macro avg       0.81      0.27      0.28       400
weighted avg       0.62      0.60      0.58       400

Overall Accuracy: 0.6025
```

- Για C=1.20, gamma=0.17

```
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         2
           4       1.00      0.00      0.00        14
           5       0.72      0.42      0.53       177
           6       0.47      0.86      0.61       162
           7       1.00      0.00      0.00        40
           8       1.00      0.00      0.00         5

    accuracy                           0.53       400
   macro avg       0.86      0.21      0.19       400
weighted avg       0.66      0.53      0.48       400

Overall Accuracy: 0.532

The best performing gamma value is:  0.17
The best performing C value is:  1.20
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         2
           4       0.00      0.00      0.00        14
           5       0.62      0.75      0.68       177
           6       0.54      0.56      0.55       162
           7       0.68      0.33      0.44        40
           8       1.00      0.00      0.00         5

    accuracy                           0.59       400
   macro avg       0.64      0.27      0.28       400
weighted avg       0.58      0.59      0.57       400

Overall Accuracy: 0.5875
```

- Για C=0.98, gamma=0.23

```
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         3
           4       1.00      0.00      0.00        14
           5       0.60      0.53      0.56       150
           6       0.48      0.75      0.59       171
           7       1.00      0.02      0.04        54
           8       1.00      0.00      0.00         8

    accuracy                           0.52       400
   macro avg       0.85      0.22      0.20       400
weighted avg       0.63      0.52      0.47       400

Overall Accuracy: 0.522

The best performing gamma value is:  0.23
The best performing C value is:  0.98
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         3
           4       1.00      0.00      0.00        14
           5       0.54      0.78      0.64       150
           6       0.54      0.53      0.54       171
           7       0.67      0.22      0.33        54
           8       1.00      0.00      0.00         8

    accuracy                           0.55       400
   macro avg       0.79      0.26      0.25       400
weighted avg       0.59      0.55      0.52       400

Overall Accuracy: 0.55
```

Ερώτημα 1 – Β

Για το (Β) υποερώτημα του *ερωτήματος 1* ζητείται να γίνει αφαίρεση του 33% των τιμών του της στήλης ph του training dataset και στην συνέχεια να εφαρμόσουμε τις παρακάτω μεθόδους:

1. Αφαίρεση της στήλης.
2. Συμπλήρωση των τιμών με το μέσο όρο των στοιχείων της στήλης.
3. Συμπλήρωση των τιμών χρησιμοποιώντας Logistic Regression
4. Εφαρμογή K-means και συμπλήρωση των τιμών που λείπουν με τον αριθμητικό μέσο όρο της συστάδας στην οποία ανήκει το δείγμα.

Για κάθε μια από τις μεθόδους έχω δημιουργήσει ένα copy του training dataset με το 33% των τιμών του να λείπει, υπάρχουν δηλαδή τα x_train_protos_tropos, x_train_deuteros_tropos, x_train_tritos_tropos, x_train_tetartos_tropos.

Στα νέα μητρώα που προκύπτουν εκπαιδεύω ένα SVM με τις καλύτερες παραμέτρους που βρήκα στο υποερώτημα Α, οι οποίες υπολογίζονται και αποθηκεύονται στις μεταβλητές best_C, για το C, και best_gamma, για το gamma. Έτσι για κάθε μια από τις 4 διαφορετικές μεθόδους εκπαιδεύω το SVM με τον εξής τρόπο: « `rbfSVM = SVC(kernel='rbf', C=best_C, gamma=best_gamma)` »

Στα αποτελέσματα υπάρχει η εξής αντιστοιχία:

`1os tropos` → Αφαίρεση της στήλης
`2os tropos` → Μέσος όρος στοιχείων
`3os tropos` → Logistic Regression
`4os tropos` → K-means

Αποτελέσματα μετά την εκτέλεση του κώδικα:

- Για C=1.03, gamma=0.23

```
los tropos:
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         3
           4       1.00      0.00      0.00         9
           5       0.66      0.68      0.67       179
           6       0.54      0.70      0.61       151
           7       0.85      0.31      0.45        55
           8       1.00      0.00      0.00         3

    accuracy                           0.61       400
   macro avg       0.84      0.28      0.29       400
weighted avg       0.65      0.61      0.59       400

Overall Accuracy 1ou tropou: 0.6075

2os tropos:
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         3
           4       1.00      0.00      0.00         9
           5       0.55      0.61      0.58       179
           6       0.46      0.59      0.52       151
           7       0.67      0.11      0.19        55
           8       1.00      0.00      0.00         3

    accuracy                           0.51       400
   macro avg       0.78      0.22      0.21       400
weighted avg       0.55      0.51      0.48       400

Overall Accuracy 2ou tropou: 0.51
```

3os tropos:

	precision	recall	f1-score	support
3	1.00	0.00	0.00	3
4	1.00	0.00	0.00	9
5	0.70	0.78	0.74	179
6	0.58	0.65	0.61	151
7	0.65	0.36	0.47	55
8	1.00	0.00	0.00	3
accuracy			0.65	400
macro avg	0.82	0.30	0.30	400
weighted avg	0.66	0.65	0.63	400

Overall Accuracy 3ou tropou: 0.645

4os tropos:

	precision	recall	f1-score	support
3	1.00	0.00	0.00	3
4	1.00	0.00	0.00	9
5	0.65	0.69	0.67	179
6	0.53	0.68	0.60	151
7	0.84	0.29	0.43	55
8	1.00	0.00	0.00	3
accuracy			0.60	400
macro avg	0.84	0.28	0.28	400
weighted avg	0.65	0.60	0.59	400

Overall Accuracy 4ou tropou: 0.605

- Για C=1.03, gamma=0.17

```
los tropos:
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         2
           4       1.00      0.00      0.00        14
           5       0.64      0.74      0.69       185
           6       0.54      0.59      0.56       152
           7       0.67      0.31      0.42        45
           8       1.00      0.00      0.00         2

    accuracy                           0.60       400
   macro avg       0.81      0.27      0.28       400
weighted avg       0.62      0.60      0.58       400

Overall Accuracy 1ou tropou: 0.6

2os tropos:
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         2
           4       1.00      0.00      0.00        14
           5       0.57      0.75      0.65       185
           6       0.48      0.47      0.47       152
           7       0.62      0.11      0.19        45
           8       1.00      0.00      0.00         2

    accuracy                           0.54       400
   macro avg       0.78      0.22      0.22       400
weighted avg       0.56      0.54      0.50       400

Overall Accuracy 2ou tropou: 0.5375
```

3os tropos:

	precision	recall	f1-score	support
3	1.00	0.00	0.00	2
4	1.00	0.00	0.00	14
5	0.68	0.79	0.73	185
6	0.60	0.64	0.62	152
7	0.70	0.36	0.47	45
8	1.00	0.00	0.00	2
accuracy			0.65	400
macro avg	0.83	0.30	0.30	400
weighted avg	0.67	0.65	0.63	400

Overall Accuracy 3ou tropou: 0.65

4os tropos:

	precision	recall	f1-score	support
3	1.00	0.00	0.00	2
4	1.00	0.00	0.00	14
5	0.64	0.74	0.69	185
6	0.53	0.58	0.55	152
7	0.68	0.33	0.45	45
8	1.00	0.00	0.00	2
accuracy			0.60	400
macro avg	0.81	0.27	0.28	400
weighted avg	0.62	0.60	0.58	400

Overall Accuracy 4ou tropou: 0.5975

- Για C=1.20, gamma=0.17

```
los tropos:
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         2
           4       0.00      0.00      0.00        14
           5       0.61      0.75      0.67       177
           6       0.54      0.55      0.54       162
           7       0.68      0.33      0.44        40
           8       1.00      0.00      0.00         5

    accuracy                           0.58       400
   macro avg       0.64      0.27      0.28       400
weighted avg       0.58      0.58      0.56       400

Overall Accuracy 1ou tropou: 0.585

2os tropos:
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         2
           4       1.00      0.00      0.00        14
           5       0.55      0.74      0.63       177
           6       0.49      0.46      0.47       162
           7       0.56      0.12      0.20        40
           8       1.00      0.00      0.00         5

    accuracy                           0.53       400
   macro avg       0.77      0.22      0.22       400
weighted avg       0.55      0.53      0.49       400

Overall Accuracy 2ou tropou: 0.525
```

3os tropos:

	precision	recall	f1-score	support
3	1.00	0.00	0.00	2
4	1.00	0.00	0.00	14
5	0.69	0.80	0.74	177
6	0.60	0.62	0.61	162
7	0.47	0.35	0.40	40
8	1.00	0.00	0.00	5
accuracy			0.64	400
macro avg	0.79	0.30	0.29	400
weighted avg	0.65	0.64	0.62	400

Overall Accuracy 3ou tropou: 0.64

4os tropos:

	precision	recall	f1-score	support
3	1.00	0.00	0.00	2
4	1.00	0.00	0.00	14
5	0.62	0.75	0.68	177
6	0.56	0.57	0.56	162
7	0.65	0.38	0.48	40
8	1.00	0.00	0.00	5
accuracy			0.60	400
macro avg	0.81	0.28	0.29	400
weighted avg	0.62	0.60	0.58	400

Overall Accuracy 4ou tropou: 0.5975

- Για C=0.98, gamma=0.23

```
los tropos:
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         3
           4       1.00      0.00      0.00        14
           5       0.54      0.78      0.64       150
           6       0.54      0.53      0.54       171
           7       0.67      0.22      0.33        54
           8       1.00      0.00      0.00         8

    accuracy                           0.55       400
   macro avg       0.79      0.26      0.25       400
weighted avg       0.59      0.55      0.52       400

Overall Accuracy 1ou tropou: 0.55

2os tropos:
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         3
           4       1.00      0.00      0.00        14
           5       0.46      0.78      0.58       150
           6       0.56      0.47      0.51       171
           7       0.80      0.07      0.14        54
           8       1.00      0.00      0.00         8

    accuracy                           0.50       400
   macro avg       0.80      0.22      0.20       400
weighted avg       0.58      0.50      0.45       400

Overall Accuracy 2ou tropou: 0.5025
```

```
3os tropos:
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         3
           4       1.00      0.00      0.00        14
           5       0.60      0.81      0.69       150
           6       0.60      0.61      0.60       171
           7       0.96      0.46      0.62        54
           8       1.00      0.00      0.00         8

    accuracy                           0.62       400
   macro avg       0.86      0.31      0.32       400
weighted avg       0.68      0.62      0.60       400

Overall Accuracy 3ou tropou: 0.625

4os tropos:
              precision    recall  f1-score   support

           3       1.00      0.00      0.00         3
           4       1.00      0.00      0.00        14
           5       0.54      0.78      0.64       150
           6       0.54      0.52      0.53       171
           7       0.61      0.20      0.31        54
           8       1.00      0.00      0.00         8

    accuracy                           0.54       400
   macro avg       0.78      0.25      0.25       400
weighted avg       0.58      0.54      0.51       400

Overall Accuracy 4ou tropou: 0.5425
```

Όπως φαίνεται από τα παραπάνω αποτελέσματα δεν μπορούμε να αποφασίσουμε καθολικά και με σιγουριά ότι κάποια μέθοδος είναι συγκριτικά καλύτερη των άλλων διότι έχουμε τέσσερα μόνο παραδείγματα και ένα συγκεκριμένο dataset. Φαίνεται, όμως, ότι η μέθοδος συμπλήρωσης τιμών χρησιμοποιώντας Logistic Regression δίνει καλές μετρικές και στα 4 παραδείγματα διαφορετικών παραμέτρων C και gamma. Οι μέθοδοι αφαίρεσης στήλης καθώς και εφαρμογής K-means μοιάζουν να δίνουν μετρικές με μικρή απόκλιση. Τέλος, η μέθοδος συμπλήρωσης τιμών με το μέσο όρο των στοιχείων της στήλης δίνει τις μικρότερες μετρικές σε κάθε παράδειγμα παραμέτρων C και gamma.

Ερώτημα 2

Αφού ολοκληρώθηκε το *Ερώτημα 1* συνέχισα με το *Ερώτημα 2*. Ασχολήθηκα, αρχικά, με την παρακάτω διαδικασία:

1. Δημιουργία ενός διανύσματος λέξεων με χρήση του nltk.word_tokenize.

2. Stemming με την βοήθεια της PorterStemmer() .

3. Stopwords removal με χρήση του stopwords.words("english").

4. Ανάθεση ως βάρος την τιμή tf-idf και μετατροπή σε vector με το TfidfVectorizer().

Στην συνέχεια χώρισα τα training-test dataset με αναλογία 75%-25%. Ως νευρωνικό δίκτυο επέλεξα το MLPClassifier, το οποίο εκπαιδεύτηκε και η απόδοσή του χρησιμοποιώντας τις μετρικές f1-score, precision και recall είναι αρκετά μεγάλη όπως φαίνεται παρακάτω:

	precision	recall	f1-score	support
0	0.83	0.84	0.84	3726
1	0.74	0.72	0.73	2274
accuracy			0.80	6000
macro avg	0.79	0.78	0.78	6000
weighted avg	0.80	0.80	0.80	6000

Overall Accuracy: 0.7982