Assignment 3, Machine Learning,

Team\_39

Task 1

Members:

Patrick

Simon

Jinwei Yuan 17306137

1. Data & Pre-Processing [10% of the assignment’s marks] Briefly describe the dataset used, including the format of the data, and how the data was processed. Explain how and why you (pre-)processed the data to make it suitable for your analysis.

The dataset we use is twitter-user gender classification. This dataset has more than 20,000 rows and has 26 columns. The target is “gender” and there are empty values and “unknown” values in the “gender” column, we ignore all these empty and “unknown” values and only focus the data/rows which has “male” or “female” in “gender” column. We also shuffle the data. After that the dataset is split into a train set (70%) and a test set (30%).

2. Algorithm & Feature Selection [30% of the assignment’s marks]

1. Text classification: we use ‘description’ as feature and ‘gender’ as target. Before applying three classifiers, **Multinomial Naïve Bayes Classifier(MNB)**, **Stochastic Gradient Descent Classifier(SGDC)**, **Multi-layer Perceptron Classifier(MLP)**, to the train data, we use CountVectorizer to tokenizing text and build a dictionary with the occurrence counts of each different word and the gender labels. This is because classification based on text or sentences makes no sense, transferring text in to occurrences of words over the dataset help us to change the problem from text classification into classification based on features with numbers. Then we use tf-idf to change the occurrences to frequencies. After this we choose parameters which influence the result more and do not affect the running time too much, we choose running grid search cross validation on MNB, SGDC and MLP to find the best values for selected parameters with the best scores for each classifiers.

From the figure **1** , we can see that for MNB, 'clf\_\_alpha': 0.5, 'clf\_\_fit\_prior': True, 'tfidf\_\_use\_idf': False, 'vect\_\_ngram\_range': (1, 2) has best result around 0.609. For SGDC from figure **2**, when the parameter values are 'clf\_\_alpha': 0.0001, 'tfidf\_\_use\_idf': True, 'vect\_\_ngram\_range': (1, 2), the score, 0.603, is the best. For MLP from figure **3**, when the parameter values are 'clf\_\_alpha': 1e-05, 'tfidf\_\_use\_idf': True, the score, 0.567, is the best.

Note: The best parameter of ‘tfidf\_\_use\_idf’ in MLP might be False sometimes, because when doing n-fold cross validation, random split for training and testing is used. But this is OK, as the alpha for

1. ……

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| accuracy | clf\_\_alpha | clf\_\_fit\_prior | tfidf\_\_use\_idf | vect\_\_ngram\_range |
| 0.594 | 0.1 | True | True | (1, 1) |
| 0.6 | 0.1 | True | True | (1, 2) |
| 0.604 | 0.1 | True | FALSE | (1, 1) |
| 0.606 | 0.1 | True | FALSE | (1, 2) |
| 0.585 | 0.1 | FALSE | True | (1, 1) |
| 0.59 | 0.1 | FALSE | True | (1, 2) |
| 0.596 | 0.1 | FALSE | FALSE | (1, 1) |
| 0.596 | 0.1 | FALSE | FALSE | (1, 2) |
| 0.601 | 0.25 | True | True | (1, 1) |
| 0.605 | 0.25 | True | True | (1, 2) |
| 0.608 | 0.25 | True | FALSE | (1, 1) |
| 0.608 | 0.25 | True | FALSE | (1, 2) |
| 0.594 | 0.25 | FALSE | True | (1, 1) |
| 0.596 | 0.25 | FALSE | True | (1, 2) |
| 0.598 | 0.25 | FALSE | FALSE | (1, 1) |
| 0.599 | 0.25 | FALSE | FALSE | (1, 2) |
| 0.606 | 0.5 | True | True | (1, 1) |
| 0.609 | 0.5 | True | True | (1, 2) |
| 0.608 | 0.5 | True | FALSE | (1, 1) |
| **0.609** | **0.5** | **True** | **FALSE** | **(1, 2)** |
| 0.599 | 0.5 | FALSE | True | (1, 1) |
| 0.599 | 0.5 | FALSE | True | (1, 2) |
| 0.6 | 0.5 | FALSE | FALSE | (1, 1) |
| 0.601 | 0.5 | FALSE | FALSE | (1, 2) |
| 0.608 | 0.75 | True | True | (1, 1) |
| 0.609 | 0.75 | True | True | (1, 2) |
| 0.606 | 0.75 | True | FALSE | (1, 1) |
| 0.608 | 0.75 | True | FALSE | (1, 2) |
| 0.6 | 0.75 | FALSE | True | (1, 1) |
| 0.601 | 0.75 | FALSE | True | (1, 2) |
| 0.599 | 0.75 | FALSE | FALSE | (1, 1) |
| 0.599 | 0.75 | FALSE | FALSE | (1, 2) |
| 0.608 | 1 | True | True | (1, 1) |
| 0.607 | 1 | True | True | (1, 2) |
| 0.605 | 1 | True | FALSE | (1, 1) |
| 0.606 | 1 | True | FALSE | (1, 2) |
| 0.6 | 1 | FALSE | True | (1, 1) |
| 0.602 | 1 | FALSE | True | (1, 2) |
| 0.597 | 1 | FALSE | FALSE | (1, 1) |
| 0.599 | 1 | FALSE | FALSE | (1, 2) |

**Figure 1**

|  |  |  |  |
| --- | --- | --- | --- |
| GridSearchCV using sgdc\_clf&sgdc\_parameters | | | |
| accuracy | clf\_\_alpha | tfidf\_\_use\_idf | vect\_\_ngram\_range |
| 0.587 | 0.001 | True | (1, 1) |
| 0.586 | 0.001 | True | (1, 2) |
| 0.569 | 0.001 | FALSE | (1, 1) |
| 0.571 | 0.001 | FALSE | (1, 2) |
| 0.602 | 0.0001 | True | (1, 1) |
| **0.603** | **0.0001** | **True** | **(1, 2)** |
| 0.597 | 0.0001 | FALSE | (1, 1) |
| 0.6 | 0.0001 | FALSE | (1, 2) |

**Figure 2**

|  |  |  |
| --- | --- | --- |
| GridSearchCV using mlp\_clf&mlp\_parameters | | |
| accuracy | clf\_\_alpha | tfidf\_\_use\_idf |
| 0.527 | 0.0001 | True |
| 0.517 | 0.0001 | FALSE |
| **0.567** | **0.00001** | **True** |
| 0.561 | 1.00E-05 | FALSE |

Figure 3

3. Evaluation [30% of the assignment’s marks]

1. Text classification: From the grid search cross validation from 2.a., we get three different classifiers equipped with their best parameters. We do 3-fold cross validation on these three classifiers with our test data to get the accuracy score. From the figure **4**, we can see the classifier based on Multinomial Naïve Bayes is the best, so for text gender classification, Multinomial Naïve Bayes is the best among the three.

|  |  |  |
| --- | --- | --- |
| accuracy on test set | | |
| **MNB** | SGDC | MLP |
| **0.56467569** | 0.52504107 | 0.5061069 |

**Figure 4**