ARIN7102 Applied Data Mining and Text Analytics
Assignment 1

1. *Concepts*

(a) Logistic Regression.

Logistic Regression is a method to finish the binary classification task. For example, it can decide between two outcomes like "yes" or "no". This method use training dataset to fit a logistic curve and then give the predict result in test dataset.

The advantage of Logistic Regression is that it can work easily when the relationship between the variables is linear and the output is the possibility of belonging to a class.

The disadvantage is that it is not effective and can not obtain perfect result when facing highly complex or non-linear problems. And it will overfitting if there are too many input feature.

(b) Bagging and Boosting methods.

Bagging uses a model with better performance to make repeated judgments about a problem or task, and finally synthesizes these judgments and gives an answer.

Boosting uses multiple simple models for task processing, each with near-random judgment power. And each model constantly fixes only the problems encountered in the previous round, so as to increase the overall judgment ability.

The advantage is that they can improve accuracy compared to individual models. And it also a method to avoid overfitting. They work well with both linear and non-linear data.

The disadvantage is that it will use a lot of computer resource when training the models.

(c) Naive Bayes.

Naive Bayes is based on Bayes' Theorem, which calculates probabilities based on prior knowledge and dataset likelihood. It assumes that all features are independent. It calculated the likelihood of different outcomes and came up with the most likely choice.

The advantage is that it is simple and fast to train. And it works well with small datasets.

The disadvantage is that it assumes feature independence, which is often unrealistic.

(d) k-Nearest Neighbor (k-NN).

k-NN classifies a data point based on the majority label of its nearest neighbors. That is, it will calculate the distance between the target point and all other points, and choose the k points that are closest to the distance, by choosing the class that belongs to the most categories of these k points as the prediction category of the target point.

The advantage is that it can work well in small dataset and does not need for training.

The disadvantage is that it is computationally expensive for large datasets and sensitive to irrelevant or noisy features.

(e) Support Vector Machine (SVM).

SVM finds the best boundary that separates classes. In other words, it is possible to find a plane that divides the data set, and the sum of the distances from all points on this plane is the largest and can be maximized to classify different categories.

The advantage is that it is effective for high-dimensional and non-linear data and robust to overfitting in certain situations.

The disadvantage is that it does not perform well when there's a lot of noise in the data.

2. Feature Selection

(a) Count vectorizer.

The accuracy of the method using sklearn is: 0.8766666666666667
The accuracy of the method using your own method is: 0.8766666666666667
The difference between the two methods is: 0.0

(b) TF-IDF vectorizer.

The accuracy of the method using sklearn is: 0.8533333333333334
The accuracy of the method using your own method is: 0.8533333333333334
The difference between the two methods is: 0.0

3. Naïve Bayes

```
Performance on class <RELEVANT>, keeping stopwords
        Precision: 0.8082191780821918    Recall: 0.9414893617021277      F1: 0.8697788697788698

Performance on class <RELEVANT>, removing stopwords
        Precision: 0.7860262008733624    Recall: 0.9574468085106383      F1: 0.8633093525179856

Top features for class <IRRELEVANT>
        notes   14.016793595000978
        job     9.944346807264209
        news    7.576645186487014
        translate       7.153680921694979
        back    6.945258087613088
        cry     6.629564538176137
        senegal 6.629564538176137
        write   6.629564538176137
        love    6.603397773872288
        visa    5.682483889865262

Top features for class <RELEVANT>
        food    27.980721649484543
        hungry  26.43690721649484
        hunger  19.464536082474222
        tents   13.944742268041237
        aid     12.492164948453604
        dying   12.20164948453608
        water   10.822731958762889
        earthquake      10.168041237113398
        tent    9.296494845360822
        carrefour       8.974948453608242
```

4. k-Nearest Neighbors

(a) K-NN Classifier.

K=1:

Got 137 / 500 correct; accuracy is 27.40%
27.4

K=5:

Got 139 / 500 correct; accuracy is 27.80%
27.8

(b) K-NN cross validate + analysis.

Best k is 10
Got 141 / 500 correct; accuracy is 28.20%
28.2

5. Support Vector Machine

(a) This is a coding contest based on SVM and feature extraction and aims to encourage participants to apply their knowledge towards problem solving. The top 10% students i.e., based on student model's test accuracy performance will get (30 points) while the other, if their model is higher than the benchmark model and give a reasonable and detailed analysis, will receive a maximum point of (20 points). Also, do not forget that you also need to write down what methods/ parameters you have tried and the detailed analysis of results. If you use some extra libraries in your implementation, please let us know in your analysis so that we can also reproduce your model. You can write your analysis either in a pdf or in the notebook file.

After using the resnet34+SVM method, I finally got 88% accuracy. For specific experimental results and data analysis, please refer to the last module of the submitted Q5-svm notebook file.

(b) If we removed the non-support vectors, does the hyper-plane change or not? Please explain.

In a support vector machine (SVM), the hyperplane is determined only by the support vector, which is the data point closest to the hyperplane. Non-support vectors lie outside the boundary, and they have no effect on the computation of the hyperplane. Therefore, if you move except the support vector, neither the position nor the direction of the hyperplane will change. The optimization goal of SVM is to maximize the distance between the support vector and the hyperplane, and other data points (non-support vectors) do not participate in this process, so they do not affect the result.

(c) If we removed one of the support vectors, does the size of the optimal margin decrease, stay the same or increase? Removing different support vectors may cause different results. You need to discuss all these situations.

If a support vector is removed, the size of the optimal interval is affected, but the change depends on the importance of the removed support vector in the hyperplane definition.

If the removed support vector is the key to defining the hyperplane and the optimal interval, then the optimal interval is reduced. Because they directly determine the size and position of the interval. When these key points are removed, the model needs to recalculate the hyperplane, often resulting in a narrower interval.

Some support vectors are located on the interval boundary, but they have little effect on the final hyperplane position. If such a support vector is removed, the other support vectors can still define the same interval and hyperplane position. In this case, the size of the optimal interval does not change.

In some cases where the distribution of support vectors is small and uneven, removing a support vector may cause the remaining data points to rearrange the interval boundary, making the interval larger.