COMP24112 Lab 1 Report

March 12, 2023

1 Explaination

In this experiment, I use k-NN algorithm with the cosine distance. I compared the Euclidean distance and the Cosine distance. And the Cosine distance has a better performance. The reason Cosine has a better performance is that it measure the direction of two vectors's similarity. It only consider the angle between them. However, Euclidean distance measure the distance and it consider the magnitude or length. When the data is high-dimensional, it is very sensitive. So in text classification the Cosine is better than Euclidean.

2 Discussion

The training accuracy test how well the model fits the training data and the testing accuracy measures how well the model generaliazes to a set of new data. Normaly, the expected training accuracy will be higher than the testing accuracy, it's because the model is trianed to fit the training data. Due to the testing accuracy test how the model fit to new data, so if the model overfit the training data, it may perform worse than training data.

3 Analyze

In the experiment I use k-NN algorithm with cosine distance, I analyze the affact of different k value to the accuracy of the model. I seperate the data set to training data and testing data and the resule shows that when k has a small value, the model has a high variance and low bias and the result is overfitted. When k has a large number, the model has a low variance and high bias so that in this case the result is underfitted. In this dataset I suggest that a reasonable value for k is between 10-20. K value in this interval shows the balance between bias and variance. However, it is important to say that the best value for k may be changed due to the dataset and the problem changed.

4 Analyze 5 articles

It is important to say that classification can be subjective and the result may be different by different people. These articles I'd like to classify them into 'sport'. My classifier can make an appropriate class prediction for them because in the first four articles the accuracy of my classifier has around 90%.

5 The performance of the classifier

In my classifier, the accuracy of each class is 96% in average. If there is no training data or very limited training data, it will result in a degradation of the classifier's performance on that category. Because of the lack of training data, the classifier will not be able to learn the linguistic patterns and features specific to sports news, resulting in the inability to accurately classify sports news into the correct category.

6 Zero shot and few shot leaning

In this lab, the model is trained to be a few shot leaning model. It is because the aim of the few shot learning is to train the model to recognise a unknown class. In this lab, we have a big number of training data. So this can be considered as a few shot learing model.

7 Estimate the interval

In my lab, I recorded all the data in a dataframe from k=1 to k=50. And I can read a specific line in the dataframe and use the data in it. We take the error rate when K=1 and use the formula below to get the interval.

$$\text{error1} - 1.645 \times \sqrt{\frac{\text{error1} \times (1 - \text{error1})}{480}}$$

$$\text{error1} + 1.645 \times \sqrt{\frac{\text{error1} \times (1 - \text{error1})}{480}}$$

8 The probability of the higher true error

In this lab when k value is 45 the error rate is about test error is about 0.06, and this value is obviously higher than the value when k has the vlaue of 1. We estimate the probability that it also has higher true error using the formula below. We can get the zp value and use the function Get_p_value to get the value of the probability.

$d = |error1 - error45|$

$$\sigma = \sqrt{\frac{error1 * (1 - error1)}{480} + \frac{error45 * (1 - error45)}{480}}$$
$$z_p = \frac{d}{\sigma}$$

9 Hyperparameter

In this section, I used k-fold method. In this method, we have a trianing set and a testing set. In the training set we seperate it to a traing set and a validation set to test its validation. First we train the model using the validation set to test whether it is valid. After that we test the whole model with the testing set.