COMP30027 Machine Learning

Project 2 Report

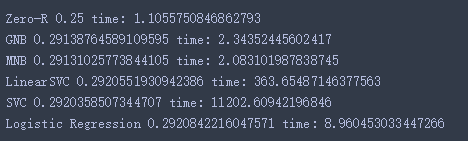
This machine learning task is to predictor the location of the author of a short text twitter.

**Week9 Tue** Attempt 1:

Use all the feature provided, use train top 10, top 50, top 100 dataset respectively.

Use 0-R, Gaussian Naïve Bayes, Multivariate Naïve Bayes, SVM, linear regression, K-NN with cosine similarity.

The result accuracy of the train top 10 with cross validation is shown below.





Summary:

At this point, I just simply tried all classifier I know, except decision tree, none of these models give a good prediction.

Thus, I need to do something to improve the model.

1. Feature selection, feature selection could make model faster and reduce the chance of overfitting
2. Ensemble method, combines classifier together to reduce the overall error rate.

Try ensemble first.

===========================================================

1. a basic description the task;
2. a short summary of some related literature;
3. a conceptual description of what you have done, including any learners that you have used, or features that you have engineered
4. an evaluation of your classifier(s) over the development tweets; You should also aim to have a more detailed discussion, which:
5. contextualises the behaviour of the method(s), in terms of the theoretical properties we have identified in the lectures;
6. attempts some error analysis of the method(s);
7. and, summarises the principal conclusions — which is to say, what a reasonably-informed reader will have learned from your efforts; And don’t forget:

1. A bibliography, which includes related work from your literature summary