Assignment 2: Imputing Methylation Status

Thee Chanyaswad
Department of Electrical Engineering
Princeton University
tc7@princeton.edu

Sergiy Popovych
Department of Computer Science
Princeton University
popovych@princeton.edu

Abstract

Streaming text data is ubiquitous: Google News, Twitter, Facebook, and many other websites have large numbers of posts from users every second. Often, we would like to classify these posts into specific categories as they arrive in order to organize the many new posts for other users to filter efficiently. In this assignment, we address the problem of classifying newsgroup posts using a number of different feature sets and methods for classification. We evaluate these methods on the 20 Newgroups data set, which contains 20,000 posts to twenty different newsgroups, where the newsgroup labels each post into each of twenty distinct classes. We train the classifiers on bag-of-words representations of each post, with and without feature selection. We find that decision trees and random forest classifiers uniformly have the highest precision and recall, whereas the perceptron and hinge loss perform poorly with respect to the other classifiers.

1 Introduction

The need for classifying streaming text is everywhere. The Associated Press releases articles that need to be filed under specific headings in Google News. Twitter and Facebook analyze each post in real time. Emails arrive in our inbox and must be filtered into *Spam* or *Inbox*. We are interested in understanding which classification methods perform well on this task. Furthermore, we would like to understand the properties of a classifier that are useful in performing this task in order to consider developing new classifiers for this purpose.

In this work, we evaluate ten separate classification methods in the task of classifying newsgroup categories from posts. For this task, we have modeled the posts using a bag-of-words representation. We evaluated the performance of each of the classification methods with and without feature selection to quantify the benefits of feature selection on classifier performance and also on the computational speed of each of the classifiers.

2 Related Work

This data set is the canonical text multi class classification data set in the machine learning community, and has been used as a benchmark for text classification methods in many contexts. In particular, recent text classification methods have used the bag-of-words representation of this data set projected down to a set of latent *topics*, where the projection is informed by the class label [**?**, **?**]. Then, for testing, new posts also are considered in the context of the low-dimensional topic space instead of their bag-of-words representation. This suggests that feature selection may benefit the classification task.

3 Methods

3.1 Problem definition

The objective is to predict the methylation level of the new samples, of which only the methylation of the sites on the illumina 450K chip is known. The algorithm learns the relationship between the methylation levels of the illumina 450K sites from training samples and those from the new samples. Then, the trained model is used to predict the methylation level of the off-chip sites on the new samples from those on the training samples. Therefore, this is a supervised regression problem.

3.2 Data processing

The dataset is derived from [1]. We use 165 features in our dataset. The first 33 features are the methylation level of the 33 training samples at the same location of the target location to be predicted. The subsequent 33 features are the the average methylation levels of the 33 training samples over the +100 locations next to the location of the target. The next 33 features are the averages over the -100 locations before the location of the target. Similarly, the next two sets of 33 features are the averages over +101 to +200 locations next to the location of the target, and -101 to -200 locations before the location of the target, respectively. The teacher (target) data are available for the locations on the illumina 45K chip. Thus, these sites are used for training.

3.3 Imputation methods

1. Six Neural Networks with the following topologies and configurations:

Networks	#of neurons	Connection	Bias	Input Layer	Hidden Layer	Output Layer
NN1	165 x 100 x 1	Fully-connected	False	Linear	Sigmoid	Linear
NN2	165 x 100 x 1	Fully-connected	True	Linear	Sigmoid	Linear
NN3	165 x 80 x 1	Fully-connected	True	Linear	Sigmoid	Linear
NN4	165 x 100 x 1	Fully-connected	True	Linear	Tanh	Linear
NN5	165 x 100 x 1	Fully-connected	True	Linear	Sigmoid	Softmax
NN6	165 x 150 x 1	Fully-connected	True	Linear	Sigmoid	Linear

3.4 Evaluation

We evaluate each of our methods on the off-chip sites of the new samples. We obtain the ground-truth for these sites from [1]. We use the Root-Mean-Square-Error ($RMSE$) and the coefficient of determination (r^2) as our main measures of success. The two measures are defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2} \tag{1}$$

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - p_i)^2}{\sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n}y_i)^2} \tag{2}$$

where y_i is the ground-truth, and p_i is the prediction. Ideally, we want $RMSE$ to be as close to zero as possible, while r^2 to be as close to 1 as possible.

Method	RMSE	r2
NN1	0.0784	0.7851
NN2	0.0793	0.7800
NN3	0.0965	0.6741
NN4	0.1015	0.6401
NN5	0.2601	-1.3657
NN6	0.0791	0.7813

Table 1: The comparative performance of the methods.

4 Results

4.1 Evaluation results

4.2 Computational speed

The variability in the time for training and testing these linear classifiers was substantial (Table **??**). In particular, we found that the KNN classifier, which does not perform training, takes the largest amount of time because of the all-by-all comparison that occurs during test phase. AdaBoost takes the second longest, but here the time is spent on training the weak classifiers and the weights of the linear combination of those weak classifiers. The fastest classifiers include the NB classifier, the perceptron, and the hinge loss classifier, followed by the linear SVM and then the decision tree and random forest classifiers.

4.3 Feature selection

These results highlight the benefits for some of the methods of reducing the number of features before training the classifiers. In particular, we found that using feature selection improved the precision for SVMS, AB, and RF classifiers, and the recall for KNN, LR, NB, SVMS, and RF classifiers. The largest improvement was for the SVMS and RF classifiers. The effect on the RF classifier might be mitigated by increasing the number of trees in the random forest for larger numbers of features, although this would slow down the training time proportionally. Across all methods, feature selection substantially improved the average wall clock time, e.g., improving the time of AdaBoost by 87.5%.

5 Discussion and Conclusion

In this work, we compared ten different classifiers to predict the newsgroup for a particular newsgroup post using bag-of-words features. We found that, considering precision, recall, and time, the decision tree and random forest classifiers showed superior performance on this task. The effect of feature selection was mostly on the time, although the improvement in performance was substantial for the random forest classifier on this task.

There are a number of directions to go that would improve these results. First, we could expand our data set using available data from these and other related newsgroups. Second, we could consider more sophisticated features for the newsgroup posts than dictionary word counts; bi-grams, post length, or punctuation may be useful in this classification task. Third, we could use the most promising models in a more problem-tailored way. In particular, because the random forest classifier showed such promise in this task, we could consider applying it to this problem using multi class class labels instead of one-versus-rest class labels, and reducing the dimension of the feature space using supervised latent Dirichlet allocation based methods [**?**, **?**].

References

[1] Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LTY, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500(7463):477–481.