EE 239AS
Special Topics in Signals and Systems
Project 2
Classification Analysis
Winter 2016

Liqiang YU, Kaiming WANG and Jun FENG
904592975, 504592374, 304588434

02-21-2016

Contents

1 Introduction **3**

2 Binary Classification **3**
2.1 Hard Margin SVM 3
2.2 Soft Margin SVM 4

3 Multiclass Classification **4**

1 Introduction

In this report, we implemented some classification models to classify the textual data from the "20 Newsgroups" dataset, including support vector machine (SVM), naive Bayes classifier and logistic regression classifier. Before the classification,there were some data preprocessing steps, like changing the number in each subset to make them balanced, transforming the textual data into TF-IDF matrix, implementing the singular value decomposition to reduce the dimension of TF-IDF matrix. The task included binary classification and multiclass classification. The results were measured with the metrics including the average of precision, recall and accuracy. Moreover, in order to characterize the trade-off between true positive rate (TPR) and false positive rate (FPR), the receiver operating characteristic (ROC) curve was plotted.

The report is organized as follows: in section 2 we discussed the classification results with support vector machine, naive Bayes classifier and logistics regression classifier. We compared the results from two SVM models : hard margin SVM and soft margin SVM, computed the precision, recall, accuracy and confusion matrix, plotted the ROC curve from three models. In section 3, we implemented two strategies for multiclass classification : "one VS one" and "one VS rest" and repeated the above procedures to test the results.

2 Binary Classification

In the binary classification problem, we chose eight classes and wanted to seperate them into two classes : Computer Technology and Recreational Activity. We assign the tag 0 to Computer Technology subclasses and tag 1 to Recreational Activity subclasses. The number of each subclass is almost the same so there is no need to balance them.

2.1 Hard Margin SVM

In the hard margin SVM, the objective function is

$$min\frac{1}{2}\|W\|_2^2$$

And the constraints is

$$y_i(W^T\overrightarrow{x_i} + b) \geq 1, i \in \{1, ..., n\}$$

In the program, we set C to 100000 to simulate the effect of hard margin. The average precision is 82.65%, the average recall is 83.14% and the accuracy is 82.44%.The confusion table is shown in Table 1. The ROC curve is shown in figure 1

Table 1: The confusion matrix of hard margin SVM

	Predicted Comp	Predicted Rect
Actual Comp	1275	285
Actual Rect	268	1322

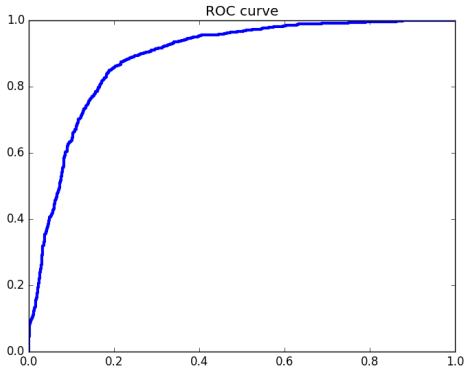

Figure 1: The ROC curve of hard margin SVM

2.2 Soft Margin SVM

The problem of the hard margin model is that it may overfit the data, therefore it's better to use the soft margin SVM. In the soft margin model, we add error parameter in the objective function and the constraints. Thus the objective function is changed to the following:

$$min\frac{1}{2}\|W\|_2^2 + \gamma\sum_{i=1}^{n}\xi_i$$

Accordingly, the constraints are changed to the following:

$$y_i(W^T\overrightarrow{x_i} + b) \geq 1 - \xi_i, \xi_i \geq 0, i \in \{1, ..., n\}$$

The γ is the hyperparameter here and different γ will affect the classification results. We implemented 5-fold cross validation to fit the model and choose the γ. The best γ we can get is 1000. The average precision is 96.61%, the average recall is 97.59% and the accuracy is 97.08%. The confusion matrix is shown in Table 2

3 Multiclass Classification

Table 2: The confusion matrix of soft margin SVM with $\gamma = 1000$

	Predicted Comp	Predicted Rect
Actual Comp	759	27
Actual Rect	19	771