# EE239AS Special Topics in Signals and Systems

Project 2 Report:

**Classification Analysis**

**Group Members:**

Jiayu Guo(504513188), Yitian Hu(904516321),

Peidong Chen(204432674), Yang Yang(804522285)

Winter 2016

02/21/2016

## Problem (a)

In this project we work with “20 Newsgroups” dataset. We plot a histogram for the following 8 topics: comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, rec.autos, rec.motorcycles, rec.sport.baseball and rec.sport.hockey. We use the training part of the dataset. And the plot is shown below:

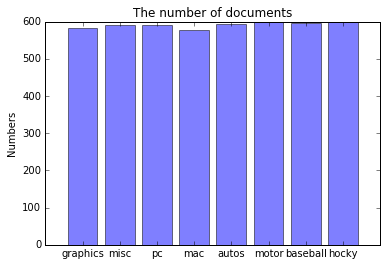


Fig.1 The plot of the number of documents per topic

From the plot above, we can see that they are evenly distributed.

And we compute that the number of documents in training part in the group "Computer Technology" is 2343, and that is the group "Recreational Activity" is 2389.

## Problem (b)

We use the "SnowballStemmer" library to exclude different stems of a word. And the number of terms of the train data we extracted is 55444.

## Problem (c)

For the class "comp.sys.ibm.pc.hardware", the 10 most significant terms that we find are:

dos, drive, bios, bus, com, card, disk, scsi, ide, controller

For the class "comp.sys.mac.hardware", the 10 most significant terms that we find are:

drive, apple, bit, mhz, monitor, memory, card, disk, scsi, mac

For the class "misc.forsale", the 10 most significant terms that we find are:

new, offer, sale, condition, good, sell, cover, ship, interest, price

For the class "soc.religion.christian", the 10 most significant terms that we find are:

people, say, church, jesus, faith, god, bible, believe, truth, good

## Problem (d)

We apply LSI to the TFxIDF matrix and pick k=50, and each document is mapped to a 50-dimensional vector. The reduced vector shape of the training set is (11314, 50)

## Problem (e)

We use SVM method to separate the documents into Computer Technology vs Recreational Activity groups.

Here is the plot of ROC curve

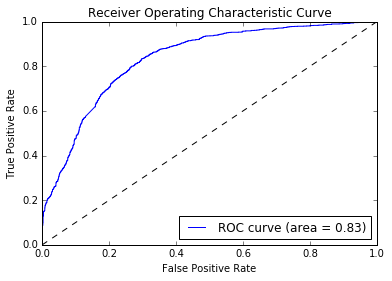


Fig. 2 ROC curve of problem (e)

The confusion matrix is

[[1241 319]

[ 448 1142]]

The accuracy score is 0.7565

The recall score is 0.71824

The precision score is 0.78166

## Problem (f)

In this problem, we use soft margin SVM approach. We find the best value of the parameter 𝛾 is 100.

Here is the plot of ROC curve

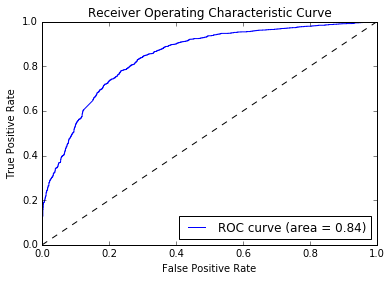


Fig. 3 ROC curve of problem (f)

The confusion matrix is

[[1278 282] [ 471 1119]]

The accuracy score is 0.76095

The recall score is 0.70378

The precision score is 0.7987

## Problem (g)

We use naïve Bayes algorith in this problem.

Here is the plot of ROC curve

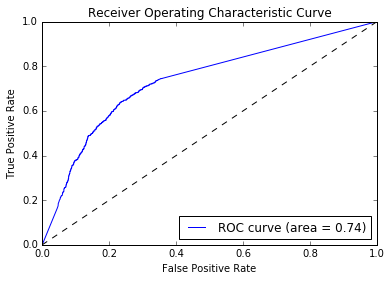


Fig. 4 ROC curve of problem (g)

The confusion matrix is

[[1257 303]

[ 686 904]]

The accuracy score is 0.6860

The recall score is 0.56855

The precision score is 0.74896

## Problem (h)

We use logistic regression in this problem.

Here is the plot of ROC curve

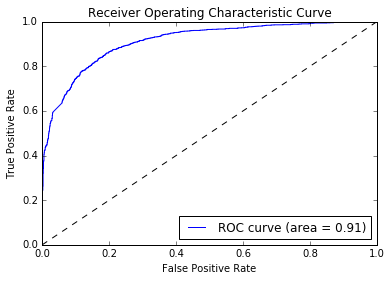


Fig. 5 ROC curve of problem (h)

The confusion matrix is

[[1335 225]

[ 315 1275]]

The accuracy score is 0.8286

The recall score is 0.8019

The precision score is 0.8500

## Problem (i)