CS 1571: Homework 4

Instructed by *Prof. Diane Litman*

Grader: *Ahmed Magooda*

Due on Dec. 6, 2017

Zac Yu (LDAP: zhy46@)

December 6, 2017

1 Report for Naive Bayes: Spam Detection

1.1 Abstract

We implemented a Naive Bayes classifier for detecting e-mail spam, and tested the classifier on a publicly available spam dataset using 5-fold cross-validation.

1.1.1 Output

This report is based on the output file, best viewed in a Markdown renderer or as a rendered Markdown in HTML.

1.2 Training Approach

1.2.1 Dataset

We used the SPAM E-mail Database from the University of California, Irvine's public Machine Learning databases[1] that consists of 4601 samples of 58 preprocessed attributes (57 features and 1 classification).

1.2.2 Data Spiting

We split the dataset to 5 groups for a 5-fold cross-validation - for iteration k, group k is used as the test set while the rest four are used for training. The five groups are divided based the samples' original position (row number modulo 5) in the dataset so that group 1 consists of samples with row number congruent to 0, group 2 consists of samples with row number congruent to 1, etc.

Iteration	Pos in Train	Neg in Train	Pos in Dev	Neg in Dev
1	1450	2230	363	558
2	1450	2231	363	557
3	1450	2231	363	557
4	1451	2230	362	558
5	1451	2230	362	558

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.DOCUMENTATION

1.2.3 Computing Feature Probabilities

For each iteration, we computed for four probabilities for each feature F_i, namely $P(F_i \leq mu_i \mid spam), P(F_i > mu_i \mid spam), P(F_i \leq mu_i \mid \neg spam)$, and $P(F_i > mu_i \mid \neg spam)$, where mu_i is the overall mean value for the feature. The probabilities for the first feature for all five iterations are shown in the attached table below. Note that instead of using a traditional smoothing algorithm, we naively replaced zero probabilities with 0.0014.

Iteration	$P(F_1 \leq mu_1 \mid spam)$	$P(F_1 > mu_1 \mid spam)$	$P(F_1 \leq mu_1 \mid \neg spam)$	$P(F_1 > mu_1 \mid \neg spam)$
1	0.708966	0.291034	0.883408	0.116592
2	0.704828	0.295172	0.886150	0.113850
3	0.702759	0.297241	0.893770	0.106230
4	0.717436	0.282564	0.886099	0.113901
5	0.715369	0.284631	0.889238	0.110762

1.3 Validation

1.3.1 Computing Predictions

For each sample in the test set, we predicts its probability of being a spam email message with a Naive Bayes classifier we built with from the training set, of the feature probabilities for that particular iteration. If the resulting probability is greater than 0.5, it will be classified as a spam message.

1.3.2 Error Rates

The predictions are checked against the classification in ground truth. We observed that our classifier exhibit the following error rates.

Fold	**False Pos**	**False Neg**	**Overall**
1	0.055556	0.187328	0.107492
2	0.052065	0.195592	0.108696
3	0.053860	0.140496	0.088043
4	0.057348	0.132597	0.086957
5	0.057348	0.162983	0.098913
Avg	0.055235	0.163799	0.098020

1.4 Analysis

1.4.1 Training v.s. Error Rates

The samples in the data set were spit to almost even groups so the affect of the ratio of negative/positive samples on the training set to the error rates across iterations is unclear. Specifically, iteration 2 and 3 have the same number of positive and negative samples in the training set, yet the false negative rate on the test set is still varies relatively largely. However, within iterations, since the ratio of negative samples is large than that of positive samples, the false positive rates are generally smaller than the false negative rates.

1.4.2 Method v.s. Error Rates

As expected, the overall error rates are significantly lower that what they would be by blindly choosing the majority class. For instance, for iteration 1 with group 1 being the test set, the majority group is negative. If all samples are classified are negative, we will have an overall error rate of $0.394137 > 0.107492$. More extremely, the false negative rate will be $1 \gg 0.187329$ yet the false positive rate will be $0 \ll 0.0555556$. The false positive rate, however, is misleading considering that the recall/sensativity is also 0.