CS 545 Machine Learning
Homework 4: Naive Bayes Classification and Logistic Regression
Haomin He

How did Naïve Bayes do compare with your SVM from Homework 3? Do you think the attributes here are independent, as assumed by Naïve Bayes? Does Naïve Bayes do well on this problem in spite of the independence assumption? Speculate on other reasons Naïve Bayes might do well or poorly on this problem.

First of all, shuffle the data and separate it into two sets, a training and test set; and each have about 40% spam, 60% not-spam. Next, calculate the mean and standard deviation for each feature. Compute the prior probability for each class in the training data. Then use the Gaussian Naïve Bayes algorithm to classify data in the test set, whether spam or not spam.

SVM has higher accuracy (0.9252) compare to Naïve Bayes (0.8058). I think the attributes here are not independent. If they are independent, Naïve Bayes would have a higher accuracy. I think Naïve Bayes does well (with accuracy of 80%) on this problem in spite of the independence assumption. Naïve Bayes might do poorly if the data is not continuously distributed.

The accuracy, precision, recall and confusion matrix for the test set:
Accuracy: 0.8058
Precision: 0.6845
Recall: 0.9404
Confusion matrix:

True positive: 853.0	False negative: 54.0
False positive: 393.0	True negative: 1002.0

Describe what library you used and what parameter values you used in running logistic regression.

I used sklearn.linear_model.LogisticRegression() to initialize logistic regression computation. Make the training data fit the linear model by using function sklearn.linear_model.LinearRegression.fit(Training_data, Target_values). Classify the test data using the Logistic Regression function sklearn.linear_model.LinearRegression.predict(Testing_data). It returns predicted values 1 or 0 / spam or not spam.

Give the accuracy, precision, and recall of your learned model on the test set, as well as a confusion matrix.
Accuracy: 0.9343
Precision: 0.9164
Recall: 0.9154

Confusion matrix:

True positive: 1327.0	False negative: 75.0
False positive: 76.0	True negative: 823.0

Write a few sentences comparing the results of logistic regression to those you obtained from Naïve Bayes and from your SVM from Homework 3.

Logistic regression has highest accuracy (0.9343) compare to Naïve Bayes (0.8058) and SVM (0.9252).
Logistic regression has highest precision (0.9164) compare to Naïve Bayes (0.6845) and SVM (0.9124).
Naïve Bayes has highest recall (0.9404) compare to logistic regression (0.9154) and SVM (0.8963).