Tyler Sorg
Machine Learning
Project 4: Naïve Bayes Classification

Description of experiment:

I split the data in training and testing subsets, with 40% spam in both. I created a probabilistic model by calculating the means and standard deviations of each feature for each class. I approximated the features as normally distributed, and then calculated the probabilities of a new example being one class or the other by multiplying all the posterior probabilties for each feature given each class. Taking the argmax was the classification heuristic.

Results:

Accuracy	0.829130
Precision	0.716456
Recall	0.937086

Confusion Matrix	**Classified as Spam**	**Classified as Not Spam**
Actually Spam	849	57
Actually Not Spam	336	1058

The classification accuracy was 82.9%. Due to underflow errors, I replaced posterior probabilities that were too small with arbitrary small numbers. When the replacement was 10^-1, the accuracy was 59.6%. The reported accuracy, precision, recall, and confusion matrix are due to replacing a too-small posterior probabilities with 10^-200.

The precision always lagged behind the recall. With comparatively large posterior probability replacements, precision fell to 49%. It seems that accuracy and precision are both proportional to changes in the standard deviation in the calculation of a normal distribution. Larger standard deviations (smaller probabilities) seems to increase accuracy and precision. Why that is, I'm not sure.

To elaborate, when the standard deviation is large, the probability approaches 0. This is because the forumula for N(x_i; mean_ji, std_ji) approaches 1/(sqrt(2*pi*std)*e^(0) as std approaches infinity. So, the probability approaches 1/(sqrt(2*pi*std) when std is large. Taking the log of small numbers can cause underflow. When this happened, I replaced that probability with an arbitrarily small number. The smaller it was, the more accurate the classifications became.

Discussion:

Do you think the attributes are independent, as assumed by Naïve Bayes?

No. Some spam words and symbols are probably closely related with others, such as free, money, order, mail, receive, and $. When corresponding words like this are used, spammyness of the message seems more likely.

Also this is a question that leads me to believe the assumption isn't true.

Does Naïve Bayes do well on this problem in spite of the independence assumption.

In spite of the independence assumption, the classification accuracy was high. Assuming independence can still lead to good approximations.

Speculate on other reasons why Naïve Bayes might do well or poorly on this problem.

With 57 features to consider, the curse of dimensionality is avoided if they are all considered separately. Having 2300 examples to train with might not be enough without making such assumptions.

I also read in a journal article that dependencies between features can either be distributed evenly or cancel each other out, making the naive assumptions relatively safe and effective.