Ensemble Method for Spam Classification

Hannu Hartikainen 67524V
Eric Malmi 80351A

Classifiers

We use a combination of three classifiers with the following requirements: (1) high accuracy, (2) fundamentally different methodology from the others, and (3) probability outputs. This way we want to rid ourselves of the errors inherent to just one of the classifiers. The parameters of the classifiers are chosen by 10-fold cross-validation.

SVM

Support vector machine (SVM) is a very popular classification method which is often used when comparing different classifiers. We use SVM with RBF kernel and optimize parameter gamma.

Bernoulli mixture

Modeling the data with mixtures of Bernoulli distributions is a natural choice as the data is bivariate. Using a Bernoulli mixture has the advantage of taking into account correlations between variables whereas a single Bernoulli distribution would neglect any such correlations resulting in a Naive Bayes classifier. Furthermore, a Bernoulli mixture performs density estimation thus offering a probabilistic interpretation of the results.

A separate mixture is learned for both the spam and the ham data. We assign a class label for any new data instance according to the Maximum Likelihood method.

Random forest

The third classifier is chosen to be a decision tree in order to be different from the other two. The different trees supported by Weka were tested, and Random forest was chosen as a simple but highly accurate method.

Combining predictions

The predictions given by the three classifiers are simply averaged together. We test this ensemble by dividing the 1000 known samples into a training set (50%) and a validation set (50%).

Results

SVM: The best 10-fold cross-validation accuracy (98.4%) was obtained with parameter gamma=0.3.
Bernoulli mixture: The best 10-fold cross-validation accuracy (98.0%) was obtained with 17 spam components and 12 ham components.
Random forest: 10-fold cross-validation accuracy was 97.1% with the default parameters.
Model average: 99.2% validation accuracy was obtained by taking the mean of the predictions.