

CS 410/510
Machine Learning
Winter, 2006

Homework 5: Bayesian Learning

Due Tuesday, February 21.

For this homework you will implement a naive Bayes classifier, and compare its performance on classifying spam and non-spam with that of the decision trees you used in Homework 2. You will be using the UCI spam database, as you did for Homework 2.

Here are the steps you need to perform:

I. Create binary attributes. The UCI spam data uses 57 continuous-valued attributes. You need to transform these into binary-valued attributes by finding a threshold c for each attribute that maximizes information gain. Write a program to do this using the algorithm we discussed in class (and described in the textbook in Section 3.7.2):

For each attribute a_i :

1. Sort the examples numerically with respect to a_i , lowest to highest
2. Find adjacent examples that differ in target classification.
3. Choose candidate threshold c_i as the midpoint of the corresponding interval.
4. Compute the information gain for each such candidate threshold c_i . Choose the one that gives highest information gain. (Break ties randomly.)

Report these attributes and corresponding c_i values in your writeup.

II. Train naive Bayes classifier. Now you have a set of 57 binary attributes of the form $a_i > c_i$. Use these binary attributes to train a naive Bayes classifier, using the training data UCI-spam.data that was given for Homework 2. For probabilities, use the m -estimate of probability, described in class and in the textbook (Section 6.9.1.1). Use $p = 1/2$, since each attribute has two possible values. Use $m = 2$.

III. Test naive Bayes classifier and compare its results with decision tree. Now run your naive Bayes classifier on the examples in UCI-spam.test. Report the accuracy on this test set. Compare it with the accuracy you obtained on this test set in Homework 2 with your (pruned) decision tree that was trained on UCI-spam.data. Also, for each hypothesis (your naive Bayes classifier and your decision tree), report the recall and the precision.

IV. Test significance of your results. (a) Given your results, what is the approximate probability that the hypothesis encoded by your naive Bayes classifier has a true higher accuracy than the hypothesis encoded by your decision tree? (See Section 5.5.1 in the textbook.) (b) Perform a paired t -test to compare the naive Bayes classifier and C.4.5. We will go over how to do this in class on Tuesday.

V. Summarize. Write a paragraph or two summarizing your results.

Here is what you need to turn in: **Hard copy:** Hand in a hard-copy of your (computer formatted, not handwritten) writeup as described above. **Electronic:** E-mail me your (well-commented) code for (a) determining the thresholds c for the continuous attributes and (b) implementing the naive Bayes classifier and instructions for running it.

Your homework will be graded based on (1) the clarity and completeness of your writeup and (2) the correctness and clarity of your code.