

CS 410/510
Machine Learning
Spring, 2007

Homework 3: Bayesian Learning

Due Tuesday, April 24.

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 same spam database that you used for Homework 2.

Here are the steps you need to perform:

I. Discretize the data. You need to transform your continuous-valued attributes into a small number of discrete bins. For this homework, experiment with two different discretization approaches: (1) Use 10 equal-width bins for the data. (2) Use \sqrt{n} bins with \sqrt{n} instances in each bin, as described in class.

For the six different training sets you created for Homework 2 (part 3), do the following:

II. Train naive Bayes classifier. First calculate the prior probabilities of the two classes, $P(+)$ and $P(-)$. Then, for each attribute a_i and bin b_j , calculate $P(a_i \in b_j|+)$ and $P(a_i \in b_j|-)$, over your training set. For each of these probabilities (including the priors), use the m -estimate of probability, as described in class, with $p = 1/2$, since each attribute has two possible values, and $m = 2$.

III. Test the naive Bayes classifier and compare its results with those of your decision trees from Homework 2. Now run your naive Bayes classifier on the examples in your test set. Report the accuracy, precision, and recall of both the Bayes classifier and your decision-tree classifiers on this test set. (I.e., give a complete table of the results on the different experiments.)

IV. Summarize. Write a few sentences 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 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.