Homework 2, due Thursday, Oct. 12, 2006
COMP 766-001 – Machine Learning for Bioinformatics

In the file ”Homework_02_data.txt” you will find a data matrix with 198 rows and 2 columns, taken from the Wisconsin Breast Cancer Database (Prognostic data set) from the UCI Machine Learning Repository. Each row corresponds to a patient. The values in the first column are the mean radii of cells in the patients' tumors (real-valued), which we have been calling "cell size" in class. The second column reports whether or not the patient had a recurrence of the cancer, with a 0 indicating no recurrence and a 1 indicating a recurrence.[1]

[1] (10 pts) Discretizing mean cell radius: Suppose we want to test for an association between mean cell radius and recurrence. Because mean cell radius is real valued, we may choose to discretize it before testing for an association. The two most common ways of doing this are as follows. Suppose we want to discretize into K different values. The equal-sized intervals approach divides the observed range of mean cell radius into K intervals of equal width. So for example, if mean cell radius varied between 10 and 30 in the data set (it doesn't, quite), then for $K = 4$ we would divide into the intervals $[10, 15)$, $[15, 20)$, $[20, 25)$ and $[25, 30)$. Each patient's mean cell radius would fall into precisely one of these intervals, and this defines the patient's discretized mean cell radius. Another very common approach is to choose intervals so that the number of data points in each bin is approximately equal. I will call this the equal-count approach.

Determine the intervals for discretizing mean cell radius using both the equal-sized and equal-counts approaches, for $K = 2$, $K = 3$, $K = 5$ and $K = 10$. This will give you 8 different discretizations in all. For each of these discretizations, state what the intervals are, and show the contingency table (patient counts) with columns corresponding to discretized mean cell radius.

[2] (10 pts) Testing association using χ^2: For each of your 8 contingency tables, use the χ^2-test to determine if there is a significant association between the two variables. If you have a statistical test with the package built in, you may use it. If not, you can use one of many web-based χ^2 calculators, such as "http://schnoodles.com/cgi-bin/web_chi_form.cgi". For each contingency table, state: the number of degrees of freedom, the value of the test statistic, and the p-value for rejecting the null hypothesis of independence.

[3] (10 pts) Testing assocation using mutual information: Using the procedure described in class, estimate the mutual information between discretized cell radius and recurrence for each of the 8 discretizations. Then, determine whether the estimated mutual information in each case is statistically significant, by using a permutation test. Report the estimated mutual information and estimated p-value for rejecting the null hypothesis of independence for each case.

[4] (0 pts) Which discretization do you think is best, and why?

[5] (10 pts) More on discretization: Suppose we have N samples of a continuous random variable, and we want to discretize that variable by dividing its observed range into K intervals. One intuitively appealing way of doing this is to choose the intervals so that the estimated information content of the discretized random variable is maximized. What choice of intervals achieves this? Justify your answer—with a formal proof, if possible.

[1] Incidentally, it has come to my attention that I've been mislabeling the axes on the two-by-two count matrix I've been presenting in class; the right display of the data is obtained by switching "cell-size" and "recur", or else by swapping the upper-right and lower-left counts.