COMP 652 - Homework 3

Assigned: Oct 8, 2008
Due: Oct 20, 2008
Late: Not accepted late

1. Exploring smoothing for discrete distributions. (10 pts.)

Consider the time-to-recurrence data given in the file "TTR.txt". This has the time to recurrence, in months, among 46 patients in the Wisconsin data set who had recurrences. We want to estimate a discrete distribution for this data. However, as the longest recurrence happened after 78 months, and there are only 46 data points, most months do not appear in the data set at all. Using the empirical frequency of a month to estimate the change of a recurrence after that amount of time is not a good idea.

Professor X and Professor Y are having a disagreement about how best to smooth this distribution. Professor X says that we should group the months together. With a group size of S, for example, months 1 to S from one group, months $S+1$ to $2S$ for another group, and so one. Professor X says that we should make a discrete distribution over groups of months. So for example, the maximum likelihood estimate of a recurrence between months 1 and S, is simply the empirical frequency of such recurrences in the data set.

Professor Y says the idea of grouping months is good, but should be implemented slightly differently. He says that we should construct a discrete distribution over months. But, the probability of any month in the range 1 to S should be given by the average of the maximum likelihood estimates for each month. For example, if $S = 3$ and months 1, 2, and 3 have maximum likelihood estimates of 0.25, 0.0 and 0.05, then we should actually assign each of those months the average probability of 0.1. Likewise for the group of months $S + 1$ to $2S$, $2S + 1$ to $3S$ and so on.

A (2.5 pts.) Write code that implements Professor X's idea. For a given set of positive integer data (such as in "TTR.txt") and for any specified choice of S, your code should calculate the discrete distribution described by Professor X.

B (2.5 pts.) Similarly, write code that implements Professor Y's idea.

C (5 pts.) To use either professor's idea, one needs to choose S. In part C, you will implement a version of cross-validation to choose a good S. Specifically, divide the TTR data set into two halves of equal size. One half will be the training set, and one the validation set. For every possible choice of S (from 1 to 78, as the largest recurrence time is 78), use the training set and your code from parts A and B to build the distributions suggested by Professors X and Y. Then, compute the probability of the validation set (i.e., the product of the probabilities of each validation datum) under each of the two distributions. Create plots that show the probability of the validation set (or the log probability, if you prefer) across different choices of S. Which choice of S is best for Professor X's style of smoothed distribution? Which choice of S is best for Professor Y's style of smoothed distribution? Comment. Which Professor has the better idea?

2. Gaussian discriminant analysis. (10 pts.)

Recall that in linear discriminant analysis (LDA) for binary classification we assume that $P(\mathbf{x}|y=0)$ is Gaussian, and so is $P(\mathbf{x}|y=1)$. Each Gaussian has its own mean, μ_0 and μ_1, but they share the same covariance matrix Σ. In the lecture, we saw what the maximum likelihood estimate for Σ is. However, sometimes we do not have enough data to fit a full covariance matrix.

A. (2 pts.) Suppose we assume that the covariance matrix has the special form $\Sigma = \sigma^2 I$, where I is the identity matrix and σ is unknown. Derive a formula for the maximum likelihood estimate of σ. (You can use, without proving, that the maximum likelihood estimates of μ_0 and μ_1 are simply the sample means among all $\mathbf{x_i}$ where $y_i = 0$ or $y_i = 1$ respectively.) More specifically, you should write down the likelihood of the input vectors $\mathbf{x}_i$ given the classes y_i, μ_0, μ_1 and σ, and maximize this likelihood with respect to σ:

$$\arg\max_{\sigma} l(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m | y_1, y_2, \ldots, y_m, \mu_0, \mu_1, \sigma)$$

B. (2 pts.) Again, assuming that $\Sigma = \sigma^2$ derive a formula for the decision boundary—the set of $\mathbf{x}$ for which $P(y=0|\mathbf{x}) = P(y=1|\mathbf{x}) = \frac{1}{2}$. How does the decision boundary depend on σ?

C. (2 pts.) Using the LDA approach with the assumption $\Sigma = \sigma^2 I$, fit the Wisconsin data for predicting whether or not the cancer recurs based on all available input features. The data is in files "wpbc_x.txt" and "wpbc_yrecur.txt". Report μ_0, μ_1, σ, and how many of the training examples are classified correctly.

D. (4 pts.) Now, assume a slightly more general model in which

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Derive the maximum likelihood estimates for σ_1, σ_2, …, σ_n. Fit such a model to the Wisconsin data. Report μ_0, μ_1, σ_1, …, σ_n, and how many of the training examples are classified correctly.