STA 414/2104, Spring 2007 — Assignment #4

Due at start of class on April 13. Note that this assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own.

In this assignment you will apply Gaussian process regression models with various covariance functions, as well as ordinary least-squares linear regression, to data from two sources — measurements of gene expression in yeast, and the output of a simulator of glaciation in North America. The aim is to draw conclusions about the relative performance of these methods, on the data sets tried, in a statistically valid way.

I have provided R functions for Gaussian process models, which were part of a solution to an assignment last year. You can use these functions without change for this assignment.

The Gaussian process functions take a covariance function as a parameter. You should try three covariance functions. The first corresponds to a simple Bayesian linear regression model:

$$Cov(t_i, t_j) \;=\; 10^2 + \rho^2 \sum_{k=1}^{D} x_{ik} x_{jk} + \sigma^2 \delta_{ij}$$

The second uses an exponential covariance term in which all the inputs interact:

$$Cov(t_i, t_j) \;=\; 10^2 + \theta^2 \exp(-\eta^2 \sum_{k=1}^{D} (x_{ik} - x_{jk})^2) + \sigma^2 \delta_{ij}$$

The third uses a sum of exponential terms for each input separately, giving an additive model:

$$Cov(t_i, t_j) \;=\; 10^2 + \sum_{k=1}^{D} \theta^2 \exp(-\eta^2 (x_{ik} - x_{jk})^2) + \sigma^2 \delta_{ij}$$

Here, x_i is the vector of inputs for case i, and x_{ik} is input k for case i. The parameters σ, ρ, θ, and η are to be estimated by maximum likelihood, using the function provide (which imposes a minimum value of 0.01). You need to supply initial values for the parameters, for which small values of ρ, θ, and η would be appropriate. Letting σ initially equal the sample standard deviation of the training targets would be reasonable.

The gene expression data is associated with a paper by Gasch, *et al* (Molecular Biology of the Cell, December 2000), available via the course web page, who investigated how the expression levels of genes in yeast differ according to the environment in which the yeast cells are grown. For example, in one environment, the yeast are grown at a high temperature; in another, they are grown with one nutrient in short supply. I eliminated some environments and some genes to avoid missing data, after which I was left with 98 environments and 4562 genes. I then randomly selected four training sets of 150 genes each, as well as a test set of 3000 genes. I chose two of the environments to predict the response in, and 8 environments to use in predicting this response. (Data on the other environments is not used for this assignment.) Gasch, *et al* were not interested in this task of predicting the response of a gene in an environment it has not been tested in, but this task is related to the vaguer task of determining the functions of genes.

This data is available from the course web page. The files `a4a-x-train`, `a4a-y1-train`, and `a4a-y2-train` contain the inputs and targets for 600 training cases. You are to split this data into four training sets of 150 cases each, by taking the first 150, the next 150, etc. Files `a4a-x-test`, `a4a-y1-test`, and `a4a-y2-test` contain the inputs and targets for the 3000 test cases.

The data from the simulator of glaciation comes from a project I am working on with Lev Tarasov and Richard Peltier in the physics department. The aim of the project is to reconstruct the history of

glaciation in North America over the last 250,000 years, using a complex simulation program. Many of the parameters of this simulator are unknown, however. These parameters control the dynamics of ice movement, aspects of the assumed climate over this time period, etc. These parameters need to be estimated on the basis of how well the simulated history of glaciation using given parameter values matches the available data on past glaciation. One set of data is the rate at which the land at various locations is currently rebounding upwards, as a result of the land no longer being weighed down by ice.

Machine learning is involved in this project because the simulator takes quite a long time to run — several hours on a vector supercomputer. Because of this, it's not feasible to explore possible parameters values by just running the simulator for each set of parameter values that are considered. Instead, using past runs of the simulator we try to predict what the simulator would do if run with given parameters. The predictions aren't going to be perfect, of course, but the hope is that they will be good enough to choose some good sets of parameters, which can then be checked using a relatively small number of actual simulator runs. The task we will look at is predicting the simulators output regarding the upward rebound from glaciation at one location, from the values of the 23 simulator parameters used for that simulation run.

This data is available from the course web page. Files `a4b-x-train` and `a4b-y-train` contain the inputs and targets for 800 training cases. You are to split this data into four training sets of 200 cases each, by taking the first 200, the next 200, etc. Files `a4b-x-test` and `a4b-y-test` contain the inputs and targets for 1000 test cases.

There are therefore three problems on which to test the Gaussian process methods (with three different covariance functions) and the ordinary least-squares regression method — two regarding gene expression and one regarding the glaciation simulator. For each of these three problems, you have four non-overlapping training sets, and a reasonably large test set. This provides the basis for statistically valid comparisons of the methods. For reference, you should also look at the trivial method of simply predicting that the response in all test cases is equal to the sample mean of the response in the training cases. (So there are five methods in all.) You should judge the quality of the predictions made by a method by the average squared error on test cases.

If our purpose is to learn something about which machine learning methods are better for tasks of various sorts, we aren't interested in performance on any particular training set, or any particular test case, but rather the expected performance for a randomly sampled training set and a randomly sampled test case. We estimate this expected performance by taking an average over the training sets and test cases that we've tried, but before declaring that method A is better than method B, we need to decide whether the observed difference in performance is statistically significant (ie, whether it is likely to reflect a real difference in expected performance, not just chance variation due to the particular training and test cases we use.)

You should judge the statistical significant of the difference in performance between two methods using a paired t test, applied to the average squared error on test cases for different training sets. This will be appropriate provided the size of the test set is large enough that uncertainty due to the random selection of test cases is negligable compared to the uncertainty due to the random choice of training set. Of course, the Gaussian assumption behind the t test also needs to be well-enough satisfied.

You should hand in a listing of your R functions and scripts, with suitable but not excessive comments, the results of your comparisons, including both the average squared error on test cases and the p-values from your tests of significance, and a discussion of the results. In your discussion, you should try to explain the results as best you can, perhaps by looking at the estimated values for the hyperparameters, or by looking at scatterplots of variables.