Take Home Quiz 1: Regression

Worth 1% of final mark.
Assigned: Sept. 13, 2006
Due: Sept. 20, 2006 at start of class
Greg Grudic
CSCI5622

Please briefly answer the following questions.

1. Assume that data is generate from the following function

$$y = f(\mathbf{x}) + \rho$$

where $f(\mathbf{x}) \in \Re$ is a real valued function defined on $\mathbf{x} \in \Re^d$, and ρ is a random variable with mean zero (i.e. $E[\rho] = 0$) and finite variance $V[\rho] = c$. If $c = 25$ and $\mathbf{x} = (6, 9999, 8.001)$, how many different values of y can this equation generate? Similarly, if $c = 0$ and $\mathbf{x} = (33, 1, 8.001)$, how many different values of y can be generated by this equation?

2. Assume my model is linear as defined by:

$$\hat{y} = \hat{f}(\mathbf{x}) = \hat{\beta}_0 + \sum_{j=1}^{d} \hat{\beta}_j x_j = \hat{\beta}_0 + \left(\hat{\beta}_1, ..., \hat{\beta}_d\right) \mathbf{x}^T$$

When can the following learning algorithm for finding the coefficients β *NOT* produce a model?

$$\hat{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

3. If my model is $\hat{y} = 2x_1 + 3x_2 - 5$, what is $\hat{y}$ when a test point is given with $(x_1, x_2) = (1, 1)$?

4. Is Ridge regression or Lasso regression better for feature selection?

5. Assume I have used kernel ridge regression and obtained the following model:

$$\hat{y} = 99 + .8K(2, x) + 0.97K(15, x) + 0.97K(0.11, x)$$

where

$$K(\mathbf{z}, \mathbf{x}) = \exp\left[-\frac{\|\mathbf{z} - \mathbf{x}\|^2}{33.3}\right]$$

Can you determine the dimension of the input space (i.e. what is d)? Can you identify at least some of the inputs (i.e. $\mathbf{x}$'s) in the training data? If so, how many and what are they?

6. Assume a set of training data $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)$, $\mathbf{x}_i \in \Re^d$, $y_i \in \Re$, $d = 1000000$ and $N = 10$. Your boss asks you to build a model from this data that has the following form:

$$\hat{y} = \hat{a}_0 + \sum_{k=1}^{d} \hat{a}_k x_k$$

What are two ridge regression learning algorithms you could use to estimate the model coefficients $\hat{a}_0, ..., \hat{a}_d$. Which algorithm would be computationally more efficient?

7. Assume a set of training data $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)$, $\mathbf{x}_i \in \Re^d$, $y_i \in \Re$, $d = 2$ and $N = 25$. Write a Pseudo-code algorithm for calculating a kernel matrix K:

$$K = \begin{pmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \dots & K(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ K(\mathbf{x}_N, \mathbf{x}_1) & \cdots & K(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix}$$

Assume that a subroutine for evaluating a kernel for vectors $\mathbf{a}, \mathbf{b} \in \Re^d$ exists and is called using $v = ker(\mathbf{a}, \mathbf{b})$ (i.e. where v is the kernel output given inputs $\mathbf{a}, \mathbf{b} \in \Re^d$).

8. Assume you have implemented a linear ridge regression learning algorithm that is called as follows:

   $\hat{\beta}$ = LinearRidgeAlg(λ, X, Y)

   where Y is a column vector of training data outputs $y_1, ..., y_N$, λ is the ridge parameter, $\hat{\beta}$ is a vector containing the linear model coefficients, and X is a matrix of training data inputs (each row is one training example) as follows

$$X = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{pmatrix}$$

   What will $\hat{\beta}$ represent if, instead of passing X to the algorithm, I pass a kernel matrix K as calculated in the previous question? i.e.

   $\hat{\beta}$ = LinearRidgeAlg(λ, K, Y)