University of Washington

Stat / Biostat 527 Nonparametric Regression and Classification

Homework 1

Issued: Thursday, April 3, 2014 **Due:** Thursday, April 10, 2014

Problem 1.1

For the light detection and ranging (LIDAR) data, fit polynomials in range of increasing degree and comment on the fit to the data. These data are in the `SemiPar` package, and are named `lidar` (command: `data(lidar)`).

(a) Fit linear, quadratic, degree 6 and degree 8 polynomial models and compare the fitted curves to the data points.

(b) Based on (a), what degree is required to obtain an adequate fit to the data? One method of assessing the latter is to examine residual plots.

(c) What degree is required to obtain a best prediction performance? Divide the data into two parts by randomly choosing 50% of data as training set and the rest as test set. Fit polynomials to the training data with degrees $d = 1, 2, 3, ..., 11, 12$. Plot the Mean Square Error in the training set and in the test set respectively. Comment on the relationship between degrees in the polynomials and the prediction error.

Problem 1.2

Within the `faraway` library there are data (`data(meatspec)`) on fat content (the response) in 215 samples of finely chopped meat, along with 100 covariates measuring the absorption at 100 wavelengths. Perform ridge regression on the entire dataset for varying values of the smoothing parameter. Specifically, set the smoothing parameter λ from 0 to 10^{-7} by a 10^{-9} difference.

(a) Plot how the parameter estimates change as a function of the smoothing parameter. Interpret any patterns in the plot. What happens to the parameter estimates when λ gets close to 0 or 10^{-7}? Please plot the estimated coefficients **unscaled**.

(b) Plot how the effective degrees of freedom change as a function of the smoothing parameter. Comment on any trends you see. Thinking about what penalized regression procedures do, explain how the two previous plots relate to each other. What is it about the ridge regression solution that allows us to analytically calculate the effective degrees of freedom?

Problem 1.3

(a) Show that minimization of expected absolute value loss, $E_{\mathbf{X},Y}\{|Y - f(\mathbf{X})|\}$ leads to $\hat{f}(\mathbf{x}) = \text{median}(Y|\mathbf{x})$. [**Hint**: you can assume $Y|\mathbf{x}$ is a continuous distribution with density $f_{y|x}(y|x)$].

(b) Consider a linear regression model with p parameters, fit by least squares to a set of training data $[(x_1, y_1),...,(x_N, y_N)]$ drawn at random from a population. Let $\hat{\beta}$ be the least squares estimate. Suppose we have some test data $[(\tilde{x}_1, \tilde{y}_1),...,(\tilde{x}_M, \tilde{y}_M)]$ drawn at random from the same population as the training data. If $Err_{train}(\beta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta^T x_i)^2$ and $Err_{test}(\beta) = \frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \beta^T \tilde{x}_i)^2$, prove that:

$\Rightarrow E_{\mathbf{X},\mathbf{Y}}[Err_{train}(\hat{\beta})] \leq E_{\mathbf{X},\mathbf{Y}}[Err_{test}(\hat{\beta})]$, where $E_{\mathbf{X},\mathbf{Y}}$ denotes averaging over all random components in this setup (i.e. $x_i/\tilde{x}_i$ and $y_i/\tilde{y}_i$).