University of Washington

Stat / Biostat 527 Nonparametric Regression and Classification

Homework 1

Issued: Thursday, April 4, 2013 **Due:** Thursday, April 11, 2013

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. Divide the data into two parts by randomly choosing 140 samples as training set and the rest as test test. Perform ridge regression on the training set and predict fat content for the test set. 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.

(b) Show how the prediction performance (measured as Mean Square Error) in the test set changes with smoothing parameter and determine an optimal smoothing parameter.

Problem 1.3

With reference to Section 10.3.1 in "Bayesian and Frequentist Regression Methods", prove the optimal prediction for the following three loss functions.

(a) Show that minimization of expected quadratic loss, $E_{\mathbf{X},Y}\{[Y - f(\mathbf{X})]^{\mathbf{2}}\}$ leads to $\hat{f}(\mathbf{x}) = E[Y|\mathbf{x}]$.

(b) Show that minimization of expected absolute value loss, $E_{\mathbf{X},Y}\{|Y - f(\mathbf{X})|\}$ leads to $\hat{f}(\mathbf{x}) = \text{median}(Y|\mathbf{x})$.

(c) Consider the bilinear loss function

$$L(y, \mathbf{x}) = \begin{cases} a[y - f(\mathbf{x})] & \text{if } f(\mathbf{x}) < y \\ b[f(\mathbf{x}) - y] & \text{if } f(\mathbf{x}) > y \end{cases} \tag{1}$$

Deduce that this leads to the optimal $\hat{f}(x)$ being the $100 \times a/(a+b)$-th quantile of the distribution of Y given $\mathbf{X} = \mathbf{x}$.