Topics on High-Dimensional Data Analytics

Homework 1

Problem 1. In this problem, you will show that the truncated power basis functions, shown below, represent a basis for a cubic spline with one knot.

$$h_1(x) = 1, \quad h_2(x) = x, \quad h_3(x) = x^2, \quad h_4(x) = x^3, \quad h_5(x) = (x-\xi)_+^3.$$

In other words, you will show that a function of the form

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)_+^3$$

is indeed a cubic regression spline, regardless of the values of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. Follow the following steps.

a. Find a cubic polynomial

$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$$

such that $f(x) = f_1(x)$ for all $x \leq \xi$. Express a_1, b_1, c_1, d_1 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

b. Find a cubic polynomial

$$f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3$$

such that $f(x) = f_2(x)$ for all $x > \xi$. Express a_2, b_2, c_2, d_2 in terms of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. You have now establish that $f(x)$ is a piecewise polynomial.

c. Show that $f_1(\xi) = f_2(\xi)$. That is, $f(x)$ is continuous at ξ.

d. Show that $f_1'(\xi) = f_2'(\xi)$. That is, $f'(x)$ is continuous at ξ.

e. Show that $f_1''(\xi) = f_2''(\xi)$. That is, $f''(x)$ is continuous at ξ.

Therefore, $f(x)$ is indeed a cubic spline.

Problem 2. Consider the leave-one-out cross-validation scheme. Define the cross-validation error for the i^{th} data point as $CV_i := y_i - \hat{f}_{(i)}(x_i)$, where $\hat{f}_{(i)}(x_i)$ is the predicted value of y_i obtained with data points $i = 1, \cdots, i-1, i+1, \cdots, n$. Consider the Kernel regression, where

$$\hat{f}(x) = \frac{\sum_{i=1}^n K(x, x_i) y_i}{\sum_{i=1}^n K(x, x_i)}.$$

Prove that

$$CV_i = \frac{y_i - \hat{f}(x_i)}{1 - \dfrac{K(x_i, x_i)}{\sum_{j=1}^{n} K(x_i, x_j)}}.$$

Given this fact, how many models do we need to fit in order to compute the mean square cross-validation error?

Problem 3. In an ultrasonic welding process, the power signal of the machine during each operation cycle is measured by a sensor. Each measured signal consists of 51 data points. A sample of 90 signals was collected, refer to 'X1.txt'. In the file, row i represents a signal collected at time index i and column j represents observation j in a signal.

a. Use the following models to estimate the mean function of the data:

   - Cubic splines (use 8 knots)
   - B-splines (use 8 knots)
   - Smoothing splines (choose the optimal lambda)
   - Kernel regression with Gaussian kernel (choose the optimal lambda)

   Plot the estimated mean functions along with the sample average signal.

b. By computing the mean squared error, select the best model.

Problem 4. Can a machine detect a cardiac abnormality? In medicine, an electrocardiogram (ECG) is an exam that allows physicians to detect a heart disease. In practice, a doctor performs an ECG, and based on the shape of the signal obtained, he determines if the heart is behaving normally. Due to the high mortality rate of cardiac diseases, it is very important to detect correctly and promptly an abnormal ECG. This is why, in recent years, there has been an increasing interest in using computers to detect cardiac abnormalities. The aim of this problem is to achieve recognition of abnormal ECG results, by treating the ECG signals as functional data, extracting relevant features and then using a classification method to discriminate between normal and abnormal ECGs. We will use a public data set available at the UCR Times Series Classification Archive. The training data set can be found as 'ECG200TRAIN', and the testing data set as 'ECG200TEST'.

Use B-splines and FPCA to classify the ECG as normal or abnormal. You can use the classification method of your preference.