Topics on High-Dimensional Data Analytics

Homework 1

Problem 1. The behavior of polynomials fit to data tends to be erratic near the boundaries. The polynomials fit beyond the boundary knots behave even more wildly than the corresponding global polynomials in that region. Assuming the function is linear near the boundaries (where we have less information anyway) is often considered reasonable. A *natural quadratic spline* adds additional constraints, namely that the function is linear beyond the boundary knots. Let

$$\boldsymbol{y}(\boldsymbol{x}) = \begin{cases} \sum_{j=1}^{3} \beta_j x^{j-1} + \sum_{k=1}^{K} \theta_k (x-\xi_k)_+^2 & for\ \ x \in [\xi_1, \xi_K] \\ \sum_{j=1}^{2} \beta_j x^{j-1} + \sum_{k=1}^{K} \theta_k (x-\xi_k)_+^2 & \text{for } x \text{ outside interval } (\xi_1, \xi_K) \end{cases}$$

Determine a set of bases for $\boldsymbol{y}(\boldsymbol{x})$ such that $\frac{d^2}{dx} y(x) = 0$ for x outside interval (ξ_1, ξ_K).

Problem 2. Let $B_{i,j}(x)$ be the i^{th} B-spline basis function of a uniform quadratic B-spline with five knots. The B-spline curve is defined as

$$f(x) = \sum_{i=0}^{2} B_{i,2}(x)\, P_i$$

$$B_{i,0}(x) = \begin{cases} 1 & if\ \ \tau_i \le x < \tau_{i+1} \\ 0 & otherwise \end{cases}$$

$$B_{i,j}(x) = \frac{x - \tau_i}{\tau_{i+j} - \tau_i} B_{i,j-1}(x) + \frac{\tau_{i+j+1} - x}{\tau_{i+j+1} - \tau_{i+1}} B_{i+1,j-1}(x)$$

Drive an expression for $B_{0,2}(x)$, $B_{1,2}(x)$ and $B_{2,2}(x)$.

Problem 3. Data provided in this section shows the electrical energy produced by coal in the United States from 1950 to 2018.

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

(a) Cubic Splines (vary the number knots say, from 6 to 15)
(b) B-splines (vary the number knots say, from 6 to 15)
(c) Smoothing Splines (choose the optimal lambda)
(d) Kernel regression with Gaussian kernel (choose the optimal lambda)

Use the leave-one-out cross-validation scheme to compute the mean squared error (MSE) and select the best model. Plot the functions along with the fitted curve.

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.