Saarland University

The Elements of Stastical Learning

Assignement 5

Due Date: 03.01.2018

Thibault Schowing *Mat. 2571837*
Sarah Mcleod *Mat. 2566398*

December 28, 2017

Problem 1

Principal Components Analysis The first principal component is the direction of maximum variance in the data. Show that this first principal component also minimizes the residual sum of squares, which is here the squared distance between the projected data point and the original data point.

The RSS is given by:

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$$

The least squares regression model is given by the equation 6.16 and 6.17 of ISRL:

$$Z_m = \sum_{j=1}^{p} \o_{jm} X_j \qquad \text{(6.16, ISLR)}$$

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i, \quad i = 1, ..., n \qquad \text{(6.17, ISLR)}$$

In the RSS equation, $f(x)$ correspond to the value projected on the first component line and is given by: $f(x) = \o(x_i - \bar{x})$ (6.19, ISLR). So we have RSS:

$$RSS = \sum_{i=1}^{n}(\theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \varepsilon_i - \o(x_i - \bar{x}))^2$$

The first component has the highest variance so $Var(f(x))$ is a maximum and so the RSS is minimized.

NOT SURE OF THE $\o(x_i - \bar{x})$ thing !!!!

Problem 2

Show that regression splines of degree d with K knots form a vector space of dimension d + K + 1 by providing a balance of the degrees of freedom in every region of the input data range and the lost degrees of freedom due to the smoothness constraints at the knots. Do not use bases of the spline vector space for your argument.

Problem 3

Consider the truncated power series representation for cubic splines with K interior knots. Let

$$f(x) = \sum_{j=0}^{3} \beta_j X^j + \sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3$$

Problem 4