MATH 829 - Spring 2016
Introduction to data mining and analysis
Homework #3

Instructions:

(1) You are allowed to work in groups of 1-4.
(2) Show (and briefly explain) all of your work to receive full credit.
(3) Submit your work before **Friday, March 25th, 2016**.

Theoretical part

Problem 1. Let $\xi_1 < \xi_2$ be two real numbers. Define

$$h_1(x) = 1, \quad h_3(x) = x^2, \quad h_5(x) = (x-\xi_1)_+^3$$
$$h_2(x) = x, \quad h_4(x) = x^3, \quad h_6(x) = (x-\xi_2)_+^3.$$

a) Show that every linear combination $\sum_{i=1}^6 \lambda_i h_i(x)$ with $\lambda_i \in \mathbb{R}$ admits two continuous derivatives on $\mathbb{R}$.

b) Suppose $f \in C^2(\mathbb{R})$ is equal to a cubic polynomial in $(-\infty, \xi_1]$, $[\xi_1, \xi_2]$, and $[\xi_2, \infty)$. Show that $f(x) = \sum_{i=1}^6 \lambda_i h_i(x)$ for some $\lambda_i \in \mathbb{R}$.

Problem 2. Let $N \geq 2$ and let $(x_i, y_i)_{i=1}^N \subset \mathbb{R}^2$ with $a < x_1 < \cdots < x_N < b$ for some $a, b \in \mathbb{R}$. Let $s(x)$ be a natural cubic spline that interpolates the sequence y_i, i.e., $s(x_i) = y_i$. Let $f \in C^2([a,b])$ be any function such that $f(x_i) = y_i$.

a) Define $h(x) := f(x) - s(x)$. Use integration by parts to show that

$$\int_a^b s''(x)h''(x)\ dx = -\sum_{i=1}^{N-1} s'''\left(\frac{x_j + x_{j+1}}{2}\right)(h(x_{j+1}) - h(x_j)) = 0.$$

b) Prove that

$$\int_a^b f''(x)^2\ dx \geq \int_a^b s''(x)^2\ dx$$

with equality if and only if $h \equiv 0$ on $[a,b]$.
Hint: Write $f''^2 = (f'' - s'' + s'')^2 = (h'' + s'')^2$.

c) Fix $\lambda > 0$. Prove that the minimiser of

$$\min_{f \in C^2([a,b])} \sum_{i=1}^N (y_i - f(x_i))^2 + \lambda \int_a^b f''(x)^2\ dx.$$

is a natural cubic spline with knots at $x_1, \dots, x_N$.
Hint: Given a function $f \in C^2([a,b])$, consider a natural cubic spline "competitor" $s(x)$ such that $s(x_i) = f(x_i)$.

Problem 3. For $x, x' \in \mathbb{R}^p$, let $K_1(x,x') := e^{-\gamma\|x-x'\|_2^2}$ and $K_2(x,x') := (1 + \langle x, x' \rangle)^d$ where d is a positive integer. Show that K_1 and K_2 are positive semidefinite kernels.

Problem 4. Consider the following three points in $\mathbb{R}^2$: $x_1 = (3,2)^T, x_2 = (3,0)^T, x_3 = (1,1)^T$, with labels $y_1 = 1, y_2 = 1, y_3 = -1$.

a) Draw the three points in the Cartesian plane. Intuitively, what is the line $\beta_0 + \beta_1 x + \beta_2 y = 0$ that maximizes the margin in the associated support vector machine classification problem?

b) Prove that your guess in a) is the unique solution of the problem

$$\min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^2} \frac{1}{2}\|\beta\|_2^2$$
$$\text{subject to } y_i(x_i^T \beta + \beta_0) \geq 1.$$

Python part

Problem 5. The `spam` dataset (available on Sakai) contains statistics (e.g. frequency of occurrence of words) for 4,601 emails. Each email is marked as spam/not spam. Compare the performance of logistic regression, LDA, QDA, and SVM to predict whether or not an email is spam.

Problem 6. The `phoneme` dataset (available on Sakai) contains the log-periodogram of 695 and 1,022 recordings of the phonemes "aa" and "ao" respectively.

a) Write a function to compute the matrix of splines $\mathbf{H}$ as discussed in class and during the lab.

b) Use a logistic regression model with smooth coefficients to predict the phonemes. Use knots uniformly distributed in $[1, 256]$. Plot the prediction error of your model as a function of the number of knots used.