Introduction to Machine Learning (WISB 365)
Final Exam

Sjoerd Dirksen

6 July 2022, 13:30-16:30

This exam consists of 5 questions, worth 45 points in total. Please write your name and student number on every sheet of your solutions.

Question 1 [12 points]

Let $\{(x_1, y_1), \dots, (x_m, y_m)\} \subset \mathbb{R} \times \mathbb{R}$ be a training data set. Consider the ℓ_1-regularized least squares (or Lasso) regression problem

$$\min_{w \in \mathbb{R}} \frac{1}{2m} \sum_{i=1}^{m} (y_i - w x_i)^2 + \lambda |w|, \tag{1}$$

where $\lambda > 0$. For simplicity we will assume that

$$\sum_{i=1}^{m} x_i^2 = m.$$

(a) Show that the solution of (1) is given by

$$w^* = \operatorname{sign}(\langle x, y \rangle) \max\left\{ \frac{|\langle x, y \rangle|}{m} - \lambda, 0 \right\},$$

where $x = (x_1, \dots, x_m)$ and $y = (y_1, \dots, y_m)$.

Hint: show first that the solution of the optimization problem

$$\min_{w \in \mathbb{R}} \frac{1}{2} w^2 - xw + \lambda |w|$$

is given by

$$w^* = \operatorname{sign}(x) \max\{|x| - \lambda, 0\}.$$

(b) When will the solution of (1) be sparser than the solution of the unregularized least squares regression problem ((1) with $\lambda = 0$)?

Question 2 [9 points]

Let $\psi : \mathcal{X} \to \mathbb{H}$ be a feature map into a Hilbert space and let $K = K_\psi$ be the kernel associated with ψ. Consider a training data set $\{(x_1, y_1), \dots, (x_m, y_m)\} \subset \mathcal{X} \times \{-1, 1\}$. Let

$$n_y = |\{1 \leq i \leq m \ : \ y_i = y\}$$

be the number of training data with label $y \in \{-1, 1\}$ and let

$$c_y = \frac{1}{n_y} \sum_{1 \leq i \leq m \, : \, y_i = y} \psi(x_i)$$

be the corresponding class average. Consider the binary classifier $h : \mathcal{X} \to \{-1, 1\}$ that assigns a label to a datum x according to which class average is closest, i.e.,

$$h(x) = \begin{cases} 1 & \text{if } \|\psi(x) - c_1\|_2 \leq \|\psi(x) - c_{-1}\|_2 \\ -1 & \text{else.} \end{cases}$$

(a) Show that $h(x) = \text{sign}(\langle w, \psi(x)\rangle + b)$ for all $x \in \mathcal{X}$, where $w = c_1 - c_{-1}$ and $b = \frac{1}{2}(\|c_{-1}\|_2^2 - \|c_1\|_2^2)$.

(b) Show that $h(x)$ can be computed using only the kernel K, without accessing entries of $\psi(x)$ or w.

(c) Sketch a situation in which h cannot be expected to perform (much) better in classifying new data than simply randomly guessing the label.

Question 3 [8 points]

Prove or disprove that the following kernels are PD kernels.

(a) Let $\mathcal{X} = \{x \in \mathbb{R}^d \ : \ \|x\|_2 < 1\}$ and consider the kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$K(x, y) = \frac{1}{1 - \langle x, y\rangle}.$$

(b) Let $\mathcal{X} = \mathbb{R}^d$ and consider the kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$K(x, y) = \|x - y\|_2^2.$$

Question 4 [10 points]

Consider the quadratic optimization problem

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} w^T C w + c^T w \qquad \text{s.t.} \qquad Aw = b, \tag{2}$$

where $C \in \mathbb{R}^{d \times d}$ is positive semidefinite, $c \in \mathbb{R}^d$, $A \in \mathbb{R}^{m \times d}$, and $b \in \mathbb{R}^m$. Prove that w^* is a solution of (2) if and only if there is some $\lambda^* \in \mathbb{R}^m$ such that $(w^*, \lambda^*) \in \mathbb{R}^d \times \mathbb{R}^m$ is a solution of the system of equations

$$\begin{pmatrix} C & -A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} w^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix}.$$

Does (2) have a unique solution?

Question 5 [6 points]

In this question we will show that the first principal component identified by PCA can be interpreted as the direction in which the variance of the data is maximal. Consider $x_1, \ldots, x_m \in \mathbb{R}^d$ and suppose that the data is centered, i.e., $\frac{1}{m}\sum_{i=1}^m x_i = 0$. Let X be the random vector that takes value x_i with probability $\frac{1}{m}$, for $i = 1, \ldots, m$. Let $W = [w_1|\ldots|w_k] \in \mathbb{R}^{d \times k}$ be the output of PCA, where $k \geq 1$. Show that w_1 solves

$$\max_{w \in \mathbb{R}^d} \text{Var}(\langle X, w\rangle) \qquad \text{s. t.} \qquad \|w\|_2 = 1.$$