CS-E4830 Kernel methods in machine learning, exam 29.05.2019 / Examiner: Rohit Babbar

Instructions: You have 3 hours to complete exam. No additional material is allowed. There are 11 questions for the total maximum of 40 points

Questions

Q.1 (4 points) Give short (a few sentences) definitions or appropriate description of the following concepts.

(a) Reproducing property of a kernel

(b) Duality gap

(c) Canonical Correlation Analysis

(d) Bochner Theorem

Q.2 (3 points) Define a kernel function in terms of feature map $\phi(.)$ and feature space $\mathcal{H}$. Use the definition to show that a kernel function is positive definite.

Q.3 (4 points) Let n be a positive number, and a function $k(.,.)$ is given by $k(x, y) = \sum_{d=1}^{D} \cos^n(x_d^2 - y_d^2)$ for all $x, y \in \mathbb{R}^D$. Show that the function $k(.,.)$ is a valid kernel.

Q.4 (3 points) Argue that the difference of kernel functions $k_1(.,.)$ and $k_2(.,.)$ may not be a valid kernel function.

Q.5 (4 points) Recall that, in the context of binary classification, the Parzen window classifier assigns a test instance x based on the distance to the centroids in the following way :

$$h(x) = \begin{cases} +1 \text{ if } ||\phi(x) - c_-||^2 > ||\phi(x) - c_+||^2 \\ -1 \text{ otherwise.} \end{cases}$$

where c_- and c_+ represent the centroids in the feature space of the negative and positive classes respectively. Show by deriving appropriate expressions for α_i and b, that the above decision function can be written in the following form $h(x) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i k(x, x_i) + b\right)$ such that $k(x, x_i) = \langle \phi(x), \phi(x_i) \rangle$. Here sgn(.) represents the sign function, and n is the total number of training samples.

F wrt f, D

Q.6 (4 points) Explain the principle of Empirical Risk Minimization and Uniform Convergence. Show that uniform convergence is sufficient for the consistency of Empirical Risk Minimization.

Q.7 (3 points) Recall the formulation of Kernel Ridge regression

$$\arg\min_f \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda ||f||_{\mathcal{H}}^2$$

where f belongs to an RKHS $\mathcal{H}$ and $\lambda > 0$ is the regularization parameter. Use Representer theorem to derive a finite dimensional equivalent problem and the closed form solution.

$f(x) = \langle f, k(x, \cdot) \rangle$

$\sum_{i=1} \alpha_i k($

Q.8 (3 points) Recall the definition of a convex set. A set C is convex if

$$\forall x_1, x_2 \in C \text{ and } 0 \leq \theta \leq 1 \Rightarrow \theta x_1 + (1-\theta)x_2 \in C$$

Assuming a set C is convex (i.e., it satisfies the above definition). Then prove that, For points $x_1, x_2, \ldots, x_n \in C$ and $\theta_1, \theta_2, \ldots \theta_n \geq 0$ such that $\theta_1 + \theta_2 + \ldots + \theta_n = 1$, the following holds

$$\theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n \in C$$

Q.9 (5 points) The primal SVM problem is given by the following optimization problem

$$\min_{\alpha \in \mathbf{R}^n, \xi \in \mathbf{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \alpha^T K \alpha \right\}$$

$\xi_i \geq 1 - y_i[K\alpha]_i$

such that

$$\begin{cases} 1 - y_i[K\alpha]_i - \xi_i \leq 0 & \text{for } i = 1, \ldots, n \\ -\xi_i \leq 0 & \text{for } i = 1, \ldots, n \end{cases}$$

$\xi_i \geq 0$

Derive the dual problem, and explain the notion of support vectors by invoking complementarity conditions.

Q.10 (3 points) Write the formulation of Principal Component Analysis and show how it is related to eigen value problem involving co-variance matrix. Is the optimization problem convex. Explain your answer.

Q.11 (4 points) In the context of ℓ_1-regularized SVM with squared hinge loss with co-ordinate descent, the problem of finding the step length along the j-th co-ordinate at the current iterate $\mathbf{w}_j$ is given by :

$$d = \min_{z} \left[\lambda |\mathbf{w}_j + z| + \mathcal{L}'_j(0; \mathbf{w}) z + \frac{1}{2} \mathcal{L}''_j(0; \mathbf{w}) z^2 \right]$$

where $\mathcal{L}_j(.)$, $\mathcal{L}'_j(.)$ and $\mathcal{L}''_j(.)$ represent the loss function, its first and second derivatives respectively, and are assumed to be known. Derive the expression for the optimal step length d.