CS-E4830 Kernel methods in machine learning, exam 12.04.2019 / 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 (5 points) Give short (a few sentences) definitions or appropriate description of the following concepts.

(a) Kernel functions

(b) Empirical and expected error

(c) Bias-variance tradeoff

(d) Union Bound

(e) Canonical Correlation Analysis

Q.2 (3 points) Explain the computational advantages of using a polynomial kernel of degree two as compared to using bigram features. Under what conditions using the features directly might be more beneficial?

Q.3 (4 points) Assume we have the kernels $k_m(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}_m(\mathbf{x}_i), \boldsymbol{\phi}_m(\mathbf{x}_j) \rangle, m = 1, \ldots, P$ at our disposal, where $\boldsymbol{\phi}_m(\mathbf{x}) = (\phi_{1m}(\mathbf{x}), \ldots, \phi_{Nm}(\mathbf{x}))^T \in \mathbb{R}^D$ is the feature vector underlying the kernel k_m.

For each kernel below, write down the equation for the underlying feature vector $\tilde{\boldsymbol{\phi}}_s(\mathbf{x})$, as a function of the feature vectors $\boldsymbol{\phi}_m, m = 1, \ldots, P$, so that $\tilde{k}_s(\mathbf{x}_i, \mathbf{x}_j) = \langle \tilde{\boldsymbol{\phi}}_s(\mathbf{x}_i), \tilde{\boldsymbol{\phi}}_s(\mathbf{x}_j) \rangle$ is satisfied for each $s \in \{a, b, c, d\}$.

(a) $\tilde{k}_a(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{P} k_m(\mathbf{x}_i, \mathbf{x}_j)$

(b) $\tilde{k}_b(\mathbf{x}_i, \mathbf{x}_j) = (k_1(\mathbf{x}_i, \mathbf{x}_j) + 1)^2$

Q.4 (3 points) Show that the kernel matrix is symmetric and positive definite.

Q.5 (3 points) Is the feature map $\phi(.)$ for a given kernel $k(.,.)$ unique? If yes, prove it. Otherwise, give a counter-example. Is the feature space unique?

Q.6 (4 points) State Representer theorem and discuss its implication for computing the prediction function values at training points $f(x_i)$ and regularizer $||f||^2_{\mathcal{H}}$ for solving ERM problems such as Kernel SVM and Kernel logistic regression.

Q.7 (4 points) Recall the formulation for Kernel Logistic Regression

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^N} \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-y_i [K\boldsymbol{\alpha}]_i)) + \frac{\lambda}{2} \boldsymbol{\alpha}^T K \boldsymbol{\alpha}$$

Show that the objective function is convex in $\boldsymbol{\alpha}$.

Q.8 (5 points) The primal optimization problem for SVM formulation with squared Hinge loss $\mathcal{L}(f(\mathbf{x}), y) = \max(0, 1 - yf(\mathbf{x}))^2$ as the loss function is given by

$$\begin{aligned} \min_{\mathbf{w},\boldsymbol{\xi}} \quad & \frac{1}{2}||\mathbf{w}||^2 + \frac{C}{2}\sum_{i=1}^{N} \xi_i^2 \\ \text{s.t.} \quad & y_i(\langle \mathbf{w}, \phi(\mathbf{x}_i)\rangle) \geq 1 - \xi_i, \quad i = 1, \ldots, N \end{aligned}$$

Derive the dual of the above problem.

Q.9 (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.10 (3 points) Explain the co-ordinate descent algorithm for solving optimization problems. Discuss how it is different from gradient descent.

Q.11 (3 points) State Bochner theorem and explain how it can be used for addressing machine learning problems with large number of training samples in the context of kernel methods.