CS-E4830 Kernel methods in machine learning, exam date 08.09.2021 / Examiner: Rohit Babbar

Instructions: You have 4 hours to complete exam. Scan (or take a picture) and send your answer sheets by 18:00pm today (08.09.2021). This is an open book exam but no additional material apart from the lecture videos/slides can be used. Consulting others to write your answers is not allowed. There are 10 questions for a total maximum of 50 points.

Questions

Q.1 (8 points) Give short (a few sentences) definitions or appropriate description of the following concepts.

(a) Kernel functions

(b) Complementary slackness

(c) Union Bound

(d) Multiple Kernel Learning

Q.2 (4 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 (6 points in total) 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 $\boldsymbol{\phi}_m(\mathbf{x}) = (\phi_{1m}(\mathbf{x}), \ldots, \phi_{Dm}(\mathbf{x}))^T \in \mathbb{R}^D$ 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 $s \in \{a, b\}$.

(a) (2 points) $\tilde{k}_a(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{P} k_m(\mathbf{x}_i, \mathbf{x}_j)$

(b) (4 points) $\tilde{k}_b(\mathbf{x}_i, \mathbf{x}_j) = (k_1(\mathbf{x}_i, \mathbf{x}_j) + 1)^2$

Q.4 (4 points) Check if $K(x, x') = \max(x, x')$ such that $x, x' \in \mathbb{R}^+$ is a valid kernel or not. If yes, prove it; give a counter-example otherwise.

Q.5 (5 points) State Representer theorem and discuss its implications 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.6 (6 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.7 (7 points) The primal optimization problem for linear SVM formulation with squared Hinge loss $(\max(0, 1 - y\mathbf{w}^T\mathbf{x}))^2$ as the loss function is given by

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

Using the method of Lagrange multipliers, derive the dual of the above problem.

Q.8 (5 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.9 (5 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.