Mid-Semester Examination

CS725: Foundations of Machine Learning

07-Sep-2015

Note: Your answers MUST employ precise mathematical statements rather than vague arguments in English.

1. All of us like ice cream cones right? If so, we should also like kernels! Because the ice-creams we buy in shops are cones in this 3-d world, whereas the set of all kernels over an input space $\mathcal{X}$ is indeed a (convex) cone in the world of functions (from $\mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$;). Note that proving this statement is the same as proving the following statement: if k_1, k_2 are kernels over $\mathcal{X}$, then so is $k \equiv \rho_1 k_1 + \rho_2 k_2$, $\forall\ \rho_1 \geq 0, \rho_2 \geq 0$. Now, your job is to prove the latter statement.

   [3Marks]

2. In the context of the previous problem, let $\phi_i : \mathcal{X} \mapsto \mathcal{H}_i$ be a feature map[1] for the kernel k_i. Now, construct a valid feature map $\phi : \mathcal{X} \mapsto \mathcal{H}$ for the kernel k by explicitly involving ϕ_1, ϕ_2 and $\mathcal{H}_1, \mathcal{H}_2$. In particular, clearly write down $\mathcal{H}$ in terms of $\mathcal{H}_1, \mathcal{H}_2$ and then ϕ in terms of ϕ_1, ϕ_2. *Hint: Think about the relation between ϕ and ϕ_1, ϕ_2 in the special case $\mathcal{X} = \mathbb{R}^2$, and $k_1(x, z) = x_1 z_1, k_2(x, z) = x_2 z_2$, where $x = [x_1\ x_2]^\top$ and $z = [z_1\ z_2]^\top$, and $\rho_1 = \rho_2 = 1$.*

   [3Marks]

3. Consider the problem of estimating the regions of high density of a distribution F using iid samples from it.

   (a) Clearly write down the empirical risk minimization formulation corresponding to this problem as presented in the lectures using the Linear model. More specifically, write down the version with 'C' as the hyper-parameter rather[2] than 'W'. Briefly describe why/when this formulation is supposed to perform the estimation task at hand.

   [3Marks]

[1] i.e., $k_i(x, z) = \langle \phi_i(x), \phi_i(z) \rangle_{\mathcal{H}_i}$

[2] i.e., your objective should look like $\frac{1}{2}\|w^2\| + \frac{C}{m}\sum_{i=1}^{m} \cdots\cdots$

(b) Suppose the input space is $\mathcal{X} = \mathbb{R}^2$ and this linear model is specified through the kernel $k(x, z) = \left(x^\top z\right)^2$. Now write down a feature map $\phi : \mathcal{X} \mapsto \mathbb{R}^3$ corresponding to this kernel[3].

[1Mark]

(c) Consider the training dataset

$$\mathcal{D} = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} -10 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -10 \end{bmatrix} \right\}.$$

Using this training set and explicitly involving the feature map ϕ, solve the minimization problem and express the optimal solution w^* in terms of C. Please simplify as much as possible[4]. *Hint: Note that the minimization problem reduces to solving simple unconstrained 1-dimensional convex quadratic objectives.*

[5Marks]

(d) Illustrate/draw the region of high density (in the input space $\mathcal{X}$) for $C = 1$.

[1Mark]

(e) By looking at the expression for the optimal solution or otherwise, argue that this kernel is "bad" (i.e., does not handle the estimation problem well even if a good model selection algorithm is employed for tuning the hyper-parameter C).

[1Mark]

(f) Repeat the above exercise using the kernel $k'(x, z) = \left(x^\top z\right)^2 + 2$. Argue that this kernel is "good". In particular, write down a C value for which the optimal solution is "intuitive".

[3Marks]

[3] Note that this is NOT the canonical feature map associated with the RKHS of the kernel.

[4] It is indeed possible to obtain an analytical expression for the optimal solution in this case.