

CS5470 Mid-Semester Exam

21-Sep-2017 2:20pm-5:20pm

Note: Please use **precise**, and **concise, mathematical** arguments. Please **justify**¹ each non-trivial statement appropriately. Please write **legibly**.

1. Let $\mathcal{X}$, and $\{-1, 1\}$ be the input, and output spaces respectively. Let $\mathcal{G}$ (model) be a given set of functions, each mapping $\mathcal{X}$ to $\{-1, 1\}$. Let the loss function, l , be the zero-one loss. Derive a high-probability upper-bound on the true-risk with empirical risk minimizer in terms of the true-risk with g^* and maximum-discrepancy of the loss-class, $\mathcal{G} \circ l$. Here, $g^* \in \mathcal{G}$ denotes the function that incurs least (true) risk among the functions in $\mathcal{G}$.

[15Marks]

2. In the context of the above set-up, derive a high-probability upper-bound on the maximum-discrepancy of the loss-class, $\mathcal{G} \circ l$, in terms of the conditional Rademacher complexity of the model, $\mathcal{G}$.

[15Marks]

3. Let $\mathcal{X}$, and $\{-1, 1\}$ be the input, and output spaces respectively. Let loss be the hinge-loss and the model be the bounded linear one: $\mathcal{G}_W = \{g \mid \exists w \in \mathbb{R}^n \ni g(x) = w^\top x \forall x \in \mathcal{X} \subset \mathbb{R}^n, \|w\| \leq W\}$. Then, in the lectures, we showed the following bound that holds uniformly for all $g \in \mathcal{G}$ with probability atleast $1 - \delta$:

$$R[g] \leq \hat{R}_m[g] + \frac{W}{\sqrt{m}} \Phi(R, \delta) + \Psi(m, \delta), \quad \forall g \in \mathcal{G}, \quad (1)$$

where² $R = \max_{x \in \mathcal{X}} \|x\|$. Given the model, i.e., given W , the upper bound on the true-risk (as given in (1)) is minimum for the empirical risk

¹You may use the following inequalities in your answers without proving them: Markov inequality, Hoeffding lemma, Chernoff-Hoeffding inequality, McDiarmid's inequality, Jensens's inequality, Cauchy-Schwartz inequality. However you need to precisely justify why a particular inequality applies in the context. In case you employ the McDiarmid inequality you may skip proving that the (random) function satisfies bounded difference property. Apart from the exceptions listed in this footnote every other non-trivial statement needs to be justified appropriately.

²In fact, $\Phi(R, \delta) = 2R + \frac{3R}{\sqrt{2}} \sqrt{\log \frac{2}{\delta}}$ and $\Psi(m, \delta) = 3\sqrt{\frac{1}{2m} \log \frac{2}{\delta}}$.

minimizer. Motivated by this observation, we propose the following for model selection, i.e., finding the right hyper-parameter $\hat{W}_m$, among the choices³, say, $W \in (0, B]$:

$$\hat{W}_m \equiv \operatorname{argmin}_{W \in (0, B]} \hat{R}_m[\hat{g}_m(W)] + \frac{W}{\sqrt{m}} \Phi(R, \delta), \quad (2)$$

where $\hat{g}_m(W)$ is the empirical risk minimizer with hyper-parameter W , i.e., $\hat{g}_m(W) = \operatorname{argmin}_{g \in \mathcal{G}_W} \hat{R}_m[g]$. Once the right $\hat{W}_m$ is known, ERM is performed in this model to obtain $\hat{g}_m(\hat{W}_m)$. Show that such an algorithm is statistically consistent⁴, i.e., show that $\left\{R[\hat{g}_m(\hat{W}_m)]\right\} \xrightarrow{P} R[g^*]$, where g^* denotes the function that incurs least (true) risk among the functions in $\mathcal{G}_B$.

[10Marks]

4. Consider the following linear classifier: $\operatorname{sign}(\mathbf{w}^\top \mathbf{x} - b = 0)$ where $\mathbf{w} = [0.6667, -1]^\top, b = 0$. Consider a data point $X = [X_1, X_2]^\top$; however X is not known exactly (it is uncertain, say modeled by a random variable). What is known about X is that its expectation is $[2, 0.75]^\top$, $0.5 \leq X_1 \leq 2.5$, $0.5 \leq X_2 \leq 1.5$ and it can be assumed that X_1 is independent of X_2 . Using Chernoff-Hoeffding bound or otherwise, lower-bound the probability that X lies on the positive side of the line $\mathbf{w}^\top x - b = 0$.

[10Marks]

5. Let Ω be a non-empty set of finite cardinality. Consider a function, $k : 2^\Omega \times 2^\Omega \mapsto \mathbb{R}$ given by⁵: $k(A_1, A_2) = |A_1 \cap A_2|$. Show that k is a valid kernel using the following two strategies:
 - (a) Explicitly write-down a finite-dimensional feature map $\phi : 2^\Omega \mapsto \mathbb{R}^n, n \in \mathbb{N}$, such that k evaluates dot-products (between mapped points in that n -dimensional space).

[5Marks]

- (b) Explicitly write-down a feature map $\phi : 2^\Omega \mapsto \mathcal{H}$, such that k evaluates inner-products (between mapped points in $\mathcal{H}$). Here, $\mathcal{H}$ must be the space of all (real-valued) random variables.

[5Marks]

³You may want to call B as hyper-hyper-parameter. This is because this fixes the model-class, from which the model is chosen, from which in turn the prediction function is chosen.

⁴Note that it is only asked to prove consistency and hence you need NOT derive any explicit learning bounds. Please do not waste time on deriving bounds.

⁵Here, $|A|$ denotes cardinality of A .

6. Let $\mathcal{G}$ be a model containing functions mapping input-space, $\mathcal{X}$, to output-space, $\mathcal{Y} \subset \mathbb{R}$ (model for regression⁶). For any $\epsilon > 0$, the covering number $\mathcal{N}(\mathcal{G}, \epsilon)$ is defined as the minimal $k \in \mathbb{N}$, such that $\mathcal{G}$ can be “covered” with k balls of radius ϵ . That is, there exist $g_1, \dots, g_k \in \mathcal{G}$, such that for any $g \in \mathcal{G}$, there exists an $i \in \{1, \dots, k\}$ such that $\max_{x \in \mathcal{X}} |g(x) - g_i(x)| \leq \epsilon$. Let us assume the covering number of $\mathcal{G}$ (at any $\epsilon > 0$) is finite⁷. Further, let us assume that there exists $M > 0$ such that $|h(x) - y| \leq M, \forall (x, y) \in \mathcal{X} \times \mathcal{Y}$. Let the loss function be squared loss: $l(y, g(x)) \equiv (y - g(x))^2$. Derive a high-probability bound on the true risk of g in terms of empirical risk of g and the covering number. This bound must uniformly hold⁸ for all $g \in \mathcal{G}$.

[20Marks]

7. Prove or disprove the statement:

(a) $k : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ given by $k(x, y) = (xy - 1)^2$ is a valid kernel.

[5Marks]

(b) $k : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ given by $k(x, y) = e^{-|x-y|}$ is a valid kernel.

[10Marks]

8. During the lecture on 14-9-2017 (the one just before this exam), we showed that the Gaussian kernel is positive-definite. Please write down the name of the matrix whose invertible nature was used to prove the result.

[5Marks]

⁶ $\mathcal{Y}$ is an uncountable set of real-values.

⁷Interested students may read further on “compact sets”.

⁸If done correctly, this bound will asymptotically approach zero, proving statistical consistency in this case. Note that covering number plays role analogous to Growth function.