Homework 9

36-465/665, Spring 2021

Due at 6 pm on Thursday, 8 April 2021

Agenda: Structural risk minimization and model averaging.

Notation / Shared Assumptions Across Questions

We have a set of q different model classes, $S_1, S_2, \ldots S_q$. (Unless a question states otherwise, assume q is constant as sample size n grows.) For each model class,

$$\begin{aligned} s^*_k &\equiv \underset{s \in S_k}{\operatorname{argmin}}\, r(s) && (1)\\ \hat{s}_k &\equiv \underset{s \in S_k}{\operatorname{argmin}}\, \hat{r}(s) && (2)\\ \Gamma_k(n) &\equiv \max_{s \in S_k} |r(s) - \hat{r}(s)| && (3)\\ \alpha &\geq \mathbb{P}\left(\Gamma_k(n) \geq g_k(n, \alpha)\right) && (4) \end{aligned}$$

You can assume that:

$$\begin{aligned} \lim_{n \to \infty} g_k(n, \alpha) &= 0 && (5)\\ g_k(n, \alpha) &\leq g_{k+1}(n, \alpha) && (6) \end{aligned}$$

and that $g_k(n, \alpha)$ grows as $\alpha \to 0$.

Further definitions:

$$\begin{aligned} k^{\dagger} &\equiv \underset{k \in 1:q}{\operatorname{argmin}}\, r(s^*_k) && (7)\\ k^* &\equiv \underset{k \in 1:q}{\operatorname{argmin}}\, r(\hat{s}_k) && (8) \end{aligned}$$

1. *Uniform convergence across model classes* We'll establish some useful results for later.

   a. (10) Show that

   $$\mathbb{P}\left(\max_{k \in 1:q} \Gamma_k(n) \geq g_q(n, \alpha/q)\right) \leq \alpha \qquad (9)$$

   b. (8) Show that

   $$\mathbb{P}\left(\max_{k \in 1:q} |r(\hat{s}_k) - \hat{r}(\hat{s}_k)| \geq g_q(n, \alpha/q)\right) \leq \alpha \qquad (10)$$

   c. (8) Show that

   $$\mathbb{P}\left(\max_{k \in 1:q} |r(s^*_k) - r(\hat{s}_k)| \geq 2g_q(n, \alpha/q)\right) \leq \alpha \qquad (11)$$

   d. (8) Show that

   $$\mathbb{P}\left(\max_{k \in 1:q} |r(s^*_k) - \hat{r}(\hat{s}_k)| \geq 3g_q(n, \alpha/q)\right) \leq \alpha \qquad (12)$$

e. (6) Explain why, for any $\epsilon > 0$,

$$\mathbb{P}\left(\max_{k \in 1:q} |r(s^*_k) - \hat{r}(\hat{s}_k)| \geq \epsilon\right) \leq \alpha \tag{13}$$

for all sufficiently large n.

2. *SRM and risk bounds* Suppose we fix an $\alpha \in (0, 1)$, and select a model class by structural risk minimization, so

$$\hat{k} = \operatorname*{argmin}_{k \in 1:q} \hat{r}(\hat{s}_k) + g_k(n, \alpha) \tag{14}$$

   a. (6) Is $\mathbb{P}\left(r(\hat{s}_{\hat{k}}) \geq \hat{r}(\hat{s}_{\hat{k}}) + 3g_q(n, \alpha/q)\right) \leq \alpha$? Why or why not?

   b. (8) Is $\mathbb{P}\left(r(\hat{s}_{\hat{k}}) \geq \hat{r}(\hat{s}_{\hat{k}}) + g_{\hat{k}}(n, \alpha)\right) \leq \alpha$? Why or why not?

3. *Model averaging, exponential weighting, and uniform convegence* In this question, we use an ensemble of models, one model s_k from each class S_k. With weight w_k on S_k,

$$\overline{s}(x) = \frac{\sum_{k=1}^{q} w_k s_k(x)}{\sum_{j=1}^{q} w_j} \tag{15}$$

   It's convenient to introduce the normalized weights

$$u_k = \frac{w_k}{\sum_{j=1}^{q} w_j} \tag{16}$$

   which sum to 1 (unlike the w_k).

   a. (8) Suppose we could (somehow) use the true risk to set weights, so $w_k = \exp(-n\beta r(s^*_k))$, and $s_k = s^*_k$. (All we assume about β is that it's > 0 and constant in n.) Explain why the normalized weight u_k of every model class *other than* $k^\dagger$ will go to 0 exponentially fast in n. Find the exponential rate, i.e., find an expression for $\lim n^{-1} \log u_k$.

   b. (6) Suppose that the risk of the ensemble is continuous in the normalized weights. Show that, under the assumption of (a), $r(\overline{s}) \to r(s^*_{k^\dagger})$.

   c. (8) Now we make the weights $w_k = \exp(-n\beta\hat{r}(\hat{s}_k))$. Show that the normalized weight of every model class other than $k^\dagger$ goes to zero as $n \to \infty$. *Hint*: Use Q1.

   d. (5) Does the conclusion of (b), that $r(\overline{s}) \to r(s^*_{k^\dagger})$, still hold under the assumptions of (c)?

   e. (8) Is there any point to doing model averaging under these assumptions, as opposed to get selecting *a* model class? Refer to the earlier parts of this question in your answer (as well, possibly, as results from lecture).

4. (1) How much time did you spend on this problem set?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.