Homework 5

36-465/665, Spring 2021

Due at 6 pm on Thursday, 11 March 2021

Agenda: Working with risk bounds; actually calculating a data-dependent risk bound for a practical classifier.

1. *Fully operational Rademacher risk bounds* Lecture 8 assumed that our loss function ℓ is bounded between 0 and m, and uses the bounded-difference (McDiarmid) inequality to show that

$$\mathbb{P}\left(r(\hat{s}) \geq \hat{r}(\hat{s}) + \epsilon\right) \leq \exp\left(-2n(\epsilon - \mathbb{E}\left[\Gamma_n\right])^2/m^2\right) \tag{1}$$

   a. (8) Lecture 8 went on to claim that we can pick any $\alpha \in (0,1)$, and that then

$$\mathbb{P}\left(r(\hat{s}) \geq \hat{r}(\hat{s}) + \mathbb{E}\left[\Gamma_n\right] + m\sqrt{\frac{\log 1/\alpha}{2n}}\right) \leq \alpha \tag{2}$$

   Show that this follows from equation 1.

   b. (6) Explain why Q1a and the fact that $\mathbb{E}\left[\Gamma_n\right] \leq 2\mathcal{R}$ implies that, for any $\alpha \in (0,1)$,

$$\mathbb{P}\left(r(\hat{s}) \geq \hat{r}(\hat{s}) + 2\mathcal{R} + m\sqrt{\frac{\log 1/\alpha}{2n}}\right) \leq \alpha \tag{3}$$

   c. (5) Lecture 8 showed that, for any $\epsilon > 0$,

$$\mathbb{P}\left(\mathcal{R}_n \geq \hat{\mathcal{R}}_n + \epsilon\right) \leq \exp\left(-2n\epsilon^2/m^2\right) \tag{4}$$

   Use this to show that for any $\beta \in (0,1)$,

$$\mathbb{P}\left(\mathcal{R}_n \geq \hat{\mathcal{R}}_n + m\sqrt{\frac{\log\left(1/\beta\right)}{2n}}\right) \leq \beta \tag{5}$$

   d. (10) Use Q1c and Q1b to show that, for any $\delta \in (0,1)$,

$$\mathbb{P}\left(r(\hat{s}) \geq \hat{r}(\hat{s}) + 2\hat{\mathcal{R}}_n + 3m\sqrt{\frac{\log 2/\delta}{2n}}\right) \leq \delta \tag{6}$$

   Hint: Consider $\alpha = \beta = \delta/2$ in the previous parts. (Why $\delta/2$?)

   e. (8) Now imagine that we get our data $Z_1, \ldots Z_n$, and we then generate a single random sequence $\sigma_1, \ldots \sigma_n$, by tossing n coins and setting $\sigma = +1$ for heads and $= -1$ for tails. We can then calculate

$$\hat{\hat{\mathcal{R}}}_n \equiv \max_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i f(Z_i)\right| \tag{7}$$

   Show that, for any $\epsilon > 0$,

$$\mathbb{P}\left(\hat{\mathcal{R}}_n \geq \hat{\hat{\mathcal{R}}}_n + \epsilon\right) \leq \exp\left(-n\frac{\epsilon^2}{2m^2}\right) \tag{8}$$

f. (8) Using the previous parts of this question, show that, for any $h \in (0,1)$,

$$\mathbb{P}\left(r(\hat{s}) \geq \hat{r}(\hat{s}) + 2\hat{\hat{\mathcal{R}}}_n + 7m\sqrt{\frac{\log 3/h}{2n}}\right) \leq h \tag{9}$$

(This is not the tightest bound we could get here, but optimizing the bound is un-informatively messy.)

2. *Computing exercise* The file `data-05.csv` on the class website contains a simple data set, with two predictor variables and a binary response which is either $+1$ or -1.

   a. (8) Load the data file and plot it, using color or shape to indicate which points are $+$ and which are $-$. Can you guess the rule used to assign the class labels?

   b. (8) Using the `tree` or `rpart` packages in R, fit a decision tree to predict the labels. Report the tree and the mis-classification rate on this data. (If you don't use R, these packages implement the CART algorithm of Breiman et al. (1984); there should be an equivalent package in your favorite language, but remember you weren't promised any computing support for other languages.)

   c. (10) Generate a sequence of Rademacher random variables, and use the same package to fit a new decision tree, which tries to predict those random variables. What is the mis-classification rate?

   d. (10) Explain how the mis-classification rate you got in Q2c can be converted into an estimate of the Rademacher complexity, what Q1 wrote as $\hat{\hat{\mathcal{R}}}$. What is that estimated Rademacher complexity?

   e. (8) Use the previous parts of this question, and Q1f, to give an upper bound on the mis-classification rate that will hold with 95% confidence.

3. (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.

References

Breiman, Leo, Jerome Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Trees*. Belmont, California: Wadsworth.