Homework 13

36-465/665, Spring 2021

Due by 6 pm on **Sunday, 9 May 2021**

Agenda: Random feature machines. Q1 and Q2 are theory; Q3 is an application, which can be done independently of the first two questions.

Questions 1 and 2 in this assignment will build towards showing the following result. We have a collection of functions $\psi(x; w)$ with a parameter or index w. For each w, $\max_x |\psi(x; w)| \leq 1$. We also have a distribution $\rho(w)$, and a constant $C > 0$. We draw $W_1, \ldots W_q$ independently and randomly from $\rho(w)$. We then get our data $(x_1, y_1), \ldots (x_n, y_n)$. These are IID, and the xs follow a pdf μ (which we don't know). After getting our data, we calculate our q-dimensional feature vectors, $\psi(x_i) = (\psi(x_i; W_1), \ldots \psi(x_i, W_q))$. Finally, we do a constrained empirical risk minimization over linear models in these features:

$$\hat{\beta} = \underset{\beta: \max_{i \in 1:q} |\beta_i| \leq C/q}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \beta \cdot \psi(x_i))$$

The claim is that if ℓ is Lipschitz-continuous, with Lipschitz constant L, then

$$\mathbb{P}\left(r(\hat{s}) - r(s^*) \geq O\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{q}}\right) LC\sqrt{\log(2/\delta)} \right) \leq \delta$$

We will work through this with the constants.

1. *The approximation part* In this question, we will think through what the best function we can hope to make from the basis functions ψ is, and how well it can be approximated using the *random* set of functions $\psi(x; W_1), \ldots \psi(x; W_q)$. To do so, we'll need to define two function classes:

$$\mathcal{F} \equiv \left\{ f(x) = \int b(w)\psi(x; w)dw \mid |b(w)| \leq C\rho(w) \right\}$$

That is, $\mathcal{F}_\rho$ consists of all the functions we can get by taking *infinite* combinations of the basic functions ψ, provided the weight $\psi(x; w)$ gets isn't too big compared to the probability $\rho(w)$ of that value of w. The other function class is

$$\hat{\mathcal{F}} \equiv \left\{ f(x) = \sum_{j=1}^{q} \beta_j \psi(x; W_j) \mid |\beta_j| \leq C/q \right\}$$

This is the class of *finite* combinations of the *random* functions $\psi(x; W_j)$, though with a similar bound on the coefficients. We are going to want to see how well the best function in $\mathcal{F}$ can be approximated using $\hat{\mathcal{F}}$. We also need to define a size or norm for functions; we'll use the $L_2(\mu)$ norm, defined by $\|f\| \equiv \sqrt{\int f^2(x)\mu(x)dx}$. The corresponding distance between two functions f and g is $\|f - g\| = \sqrt{\int (f(x) - g(x))^2 \mu(x)dx}$.

   a. (5) *Deviation inequality for functions* We start with a technical result about averaging random functions. Say $F_1, \ldots F_q$ are IID random functions of limited norm, $\mathbb{P}(\|F_1\| \leq m) = 1$. Write $\overline{F}$ for their average. Show that changing one of the F_i changes $\overline{F}$ by at most $2m/q$. Show that

$\mathbb{E}\left[\|\overline{F} - \mathbb{E}\left[\overline{F}\right]\|^2\right] \leq m^2/q$. Combine these with the bounded-difference inequality to show that

$$\mathbb{P}\left(\|\overline{F} - \mathbb{E}\left[\overline{F}\right]\| \geq \frac{m}{\sqrt{q}}\left(1 + \sqrt{2\log 1/\alpha}\right)\right) \leq \alpha$$

b. (8) *Approximating $\mathcal{F}$ by $\hat{\mathcal{F}}$* Pick any function $f^* \in \mathcal{F}$, which we can represent as $\int a(w)\psi(x;w)dw$ for some weight function $a(w)$. Define F_i as $\frac{a(W_i)}{\rho(W_i)}\psi(x;W_i)$. Show that $\mathbb{E}\left[F_i\right](x) = f^*(x)$. Define $\overline{F}$ as the average of $F_1, \ldots F_q$. Use the previous part to show that $\overline{F} \in \hat{\mathcal{F}}$, and that

$$\mathbb{P}\left(\|\overline{F} - f^*\| \geq \frac{C}{\sqrt{q}}\left(1 + \sqrt{2\log 1/\alpha}\right)\right) \leq \alpha$$

c. (5) Suppose that the loss function is Lipschitz-continuous in the action, with Lipschitz constant L. Fix any function $f^* \in \mathcal{F}$, and let $\overline{F}$ be as in the previous part. Show that

$$\mathbb{P}\left(r(\overline{F}) \geq r(f^*) + \frac{LC}{\sqrt{q}}\left(1 + \sqrt{2\log 1/\alpha}\right)\right) \leq \alpha$$

Hint: For any random variable, $\mathbb{E}\left[|Z|\right] \leq \sqrt{\mathbb{E}\left[Z^2\right]}$. (You don't have to prove this, but see EC1.)

d. (10) Now let s^* be the function in $\mathcal{F}$ which minimizes the risk. Is it true that with probability at least $1-\alpha$, there is an $F^* \in \hat{\mathcal{F}}$ such that

$$r(F^*) \leq r(s^*) + \frac{LC}{\sqrt{q}}\left(1 + \sqrt{2\log 1/\alpha}\right)$$

Explain.

2. *The estimation part*

a. (8) Show that $\hat{\mathcal{R}}_n(\hat{\mathcal{F}}) \leq C/\sqrt{n}$. *Hint*: It's possible to do this straight from the definitions of empirical Rademacher complexity and of $\hat{\mathcal{F}}$, but you can also use known results from earlier in this course about the complexity of linear function classes, and the assumption that $|\psi(x;w)| \leq 1$ for all x and w.

b. (5) Show that the Rademacher complexity of the loss class is at most $2LC/\sqrt{n}$. *Hint*: HW12Q2e.

c. (12) Derive the final generalization error bound, of the form

$$\mathbb{P}\left(r(\hat{\beta}) \geq r(s^*) + g(n, L, C, q, \alpha)\right) \leq 2\alpha$$

Use your answers to earlier questions to determine the appropriate combination of terms in g. *Hint*: The 2 in $\leq 2\alpha$ is not a typo, but an indication that you should use a union bound.

3. **Try it out** Revisit the `anomaly.csv` data set from HW8Q3. Divide it randomly and equally into training and testing sets.

a. (8) Create a library of 40 random cosine transformations of the t variable. These should all be functions of the form $\cos(wt + b)$, with b uniformly distributed between $-\pi$ and π, and w following a standard Gaussian. Create a plot showing the first four of the random cosine transformations from your library. *R hint*: Look at the .Rmd file for lecture 22.

b. (8) Add new columns to the data frame giving the value of the cosine transformations at each data point. (This should add 40 new columns.) Are these new features uncorrelated with each other? Are they uncorrelated with t? Should they be?

c. (10) Split the data randomly and evenly (as evenly as possible) into a training and a testing set. For each $q \in 1:40$, fit a linear model for the `anomaly` variable as a function of the first q random features in your library, using only points in the training set. Calculate the mean squared error on the testing set. Plot the out-of-sample MSE versus q. What is the optimal value of q?

d. (10) Create a plot showing the actual data points, and the predicted values of the model with the optimal value of q. Comment.

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.

Extra credit

1. (3) Prove that for any (scalar) random variable Z, $\mathbb{E}\left[|Z|\right] \leq \sqrt{\mathbb{E}\left[Z^2\right]}$. *Hint*: Jensen's inequality.
2. (5) Repeat the exercise of Q3 for the same split into training and testing sets, but a *different* library of 40 random features. Do you select the same order of model? Should you? How similar are the fitted curves? *Should* they be similar?