# Homework 3

36-465/665, Spring 2021

Due at 6 pm on Thursday, 25 February 2021

1. *Optimism and the "covariance penalty"* If we use data $(X_1, Y_1), \ldots (X_n, Y_n)$ to learn a predictive model $\hat{\mu}$, the "optimism" of our method is defined as how much worse that model would do on new data with the same values of $X$ but *independent* $Y$s. That is, for each $i$, $Y_i'$ has the same distribution as $Y_i$ (conditional on $X_i$), but is independent of $Y_i$, and the optimism (for regression) is

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \hat{\mu}(X_i))^2\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}(X_i))^2\right] \tag{1}$$

   In this problem and the next, we'll see how to build a simple, unbiased estimator of the optimism.

   a. (5) Show that the optimism (as defined above) is equal to

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[(Y_i' - \hat{\mu}(X_i))^2\right] - \mathbb{E}\left[(Y_i - \hat{\mu}(X_i))^2\right]\right) \tag{2}$$

   b. (5) Show that $\mathbb{E}\left[Y_i' - \hat{\mu}(X_i)\right] = \mathbb{E}\left[Y_i - \hat{\mu}(X_i)\right]$.

   c. (5) Show that the optimism is equal to

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Var}\left[Y_i' - \hat{\mu}(X_i)\right] - \mathrm{Var}\left[Y_i - \hat{\mu}(X_i)\right]\right) \tag{3}$$

   d. (5) Show that

$$\mathrm{Var}\left[Y_i - \hat{\mu}(X_i)\right] = \mathrm{Var}\left[Y_i\right] + \mathrm{Var}\left[\hat{\mu}(X_i)\right] - 2\mathrm{Cov}\left[Y_i, \hat{\mu}(X_i)\right] \tag{4}$$

   e. (5) Show that

$$\mathrm{Var}\left[Y_i' - \hat{\mu}(X_i)\right] = \mathrm{Var}\left[Y_i\right] + \mathrm{Var}\left[\hat{\mu}(X_i)\right] \tag{5}$$

   f. (5) Show that the optimism equals

$$\frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}\left[Y_i, \hat{\mu}(X_i)\right] \tag{6}$$

   g. (5) Explain why

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}(X_i))^2 + \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}\left[Y_i, \hat{\mu}(X_i)\right] \tag{7}$$

   is an unbiased estimate of

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \hat{\mu}(X_i))^2\right] \tag{8}$$

2. *Filling in a step* (5) In lecture 5, slide 20, it's asserted that "the deviating-below-expectation bound matches exactly" the bound on the probability of deviating above expectation by $\epsilon$. Show that this is true.

3. *Converting deviation bounds into high-probability bounds and sample-size bounds*

   a. (4) In lecture 5, we express the Chebyshev inequality for the sample mean as $\mathbb{P}\left(|\overline{X}_n - \mathbb{E}[X]| \geq \epsilon\right) \leq \mathrm{Var}[X]/n\epsilon^2$. That is, we start with the size of the deviation $\epsilon$, and give a bound on the probability of a deviation of that size or larger. Use this to give a formula, say $g(n,\alpha)$, where I can state my desired confidence level $\alpha \in (0,1)$, and you can guarantee that $\mathbb{P}\left(|\overline{X}_n - \mathbb{E}[X]| \leq g(n,\alpha)\right) \geq 1-\alpha$.

   b. (4) Continuing with the previous problem, find the minimum sample size $N(\epsilon,\alpha)$ which guarantees that if $n \geq N(\epsilon,\alpha)$, then $\mathbb{P}\left(|\overline{X}_n - \mathbb{E}[X]| \geq \epsilon\right) \leq \alpha$.

   c. (4) Repeat (a), but start from the Chernoff bound on sub-Gaussian random variables (given in Lecture 5).

   d. (4) Repeat (b), but again start from the Chernoff bound on sub-Gaussian random variables.

4. *The random projection trick* (a.k.a. the "Johnson-Lindenstrauss lemma" or "Johnson-Lindenstrauss theorem") Suppose we have $n$ different $p$-dimensional vectors $x_1, \ldots x_n$. A function which maps these vectors to new points, in $\mathbb{R}^p$ or any other space, is called an **isometry** if it preserves distances[1], so $\|F(x_i) - F(x_j)\| = \|x_i - x_j\|$ for all $i, j$. A function is a **$\epsilon$-approximate isometry** if[2]

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \|F(x_i) - F(x_j)\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2 \tag{9}$$

   Specifically, we will consider *linear* functions which map the vectors into $\mathbb{R}^q$, for $q < p$, and show that random linear projections are approximate isometries with high probability.

   a. (5) Create a random $p$-dimensional vector $V$, where the entries are all independent standard Gaussians, so $V_i \sim \mathcal{N}(0,1)$. Show that, for any non-random $p$-dimensional vector $x$ (other than 0), $V \cdot x/\|x\| \sim \mathcal{N}(0,1)$.

   b. (5) Create random vectors $V_1, \ldots V_q$ as in (a), and stack them as the rows in a $q \times p$ matrix $\mathbf{V}$. Using (a), show that

$$Z \equiv \frac{\|\mathbf{V}x\|^2}{\|x\|^2} \sim \chi^2_q \tag{10}$$

   c. (4) The moment generating function of a $\chi^2_q$ distribution is $M(t) = (1-2t)^{-q/2}$ for $t < 1/2$, and undefined for $t > 1/2$. Use this to show that

$$\mathbb{P}\left(|Z - q| \geq \epsilon\right) \leq 2\exp\left(-\epsilon^2/8q\right) \tag{11}$$

   (If you get stuck here, go ahead to do the rest of this question using this result, and come back to this part when you have time.)

   d. (5) Define $F(x) = q^{-1/2}\mathbf{V}x$. Use (c) to show that

$$\mathbb{P}\left(\frac{\|F(x)\|^2}{\|x\|^2} \notin [1-\epsilon, 1+\epsilon]\right) \leq 2\exp\left(-q\epsilon^2/8\right) \tag{12}$$

   e. (5) Use (d) to show that

$$\mathbb{P}\left(\text{for some } i,j, \frac{\|F(x_i - x_j)\|^2}{\|x_i - x_j\|^2} \notin [1-\epsilon, 1+\epsilon]\right) \leq n(n-1)\exp\left(-q\epsilon^2/8\right) \tag{13}$$

   *Hints*: for any two events $A$ and $B$, $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ ("the union bound"); how many pairs $i, j$ can we form from $n$ points?

[1] Throughout this problem, and this course, $\|\cdot\|$ indicates the ordinsry Euclidean length (or "norm") of a vector, so $\|v\| = \sqrt{\sum_{i=1}^p v_i^2}$ in Cartesian coordinates.

[2] You might think it more natural to state this in terms of the distances than of the squared distances, but this definition is easier to work with than the other one would be, as we'll see.

f. (5) Use (e) to show that with probability at least $1 - \alpha$, the $F$ we have built is an $\epsilon$-isometry, provided $q \geq \frac{16}{\epsilon^2} \log n/\alpha$. *Hint*: try upper-bounded the formula in (e) by something easier to invert.

g. (4) Suppose we have $10^9$ data points, each represented as ten-thousand-dimensional vectors. (For instance each could be a $100 \times 100$ pixel image.) How many random vectors would we need to use, to have 99% confidence that the resulting random projection was a 0.01-approximate isometry? What would change if each data point was had a million dimensions instead? (The answers will help explain why this technique is very widely used in industry for dimension reduction.)

5. (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.