Homework 2

36-465/665, Spring 2021

Due at 6 pm on Thursday, 18 February 2021

1. *The "sandwich variance" for linear regression* In this problem, and this problem only, suppose that our data consists of IID pairs (X_i, Y_i), and that both X_i and Y_i are centered, so $\mathbb{E}[X_i] = \mathbb{E}[Y_i] = 0$. We want to estimate a linear regression of Y on X by least squares, so we would ideally like to find the b^* which minimizes $r(b) = \mathbb{E}\left[(Y - bX)^2\right]$. We do *not* assume that the true relationship between Y and X is linear.
   a. (5) It is known (e.g., from Lecture 2) that the optimal $b^* = \mathrm{Cov}[X, Y]/\mathrm{Var}[X]$. Use this to show that $\mathbb{E}[Y - b^*X] = 0$ and that $\mathrm{Cov}[Y - b^*X, X] = 0$.
   b. (5) Show that $r(b) = \mathrm{Var}[Y] + b^2\mathrm{Var}[X] - 2b\mathrm{Cov}[Y, X]$.
   c. (5) Show that the second derivative of $r(b)$ (with respect to b) is $r''(b) = 2\mathrm{Var}[X]$.
   d. (5) With finite data, we approximate $r(b)$ by $\hat{r}(b) = n^{-1}\sum_{i=1}^{n}(Y_i - bX_i)^2$. Define the residual for the i^{th} observation as $D_i(b) \equiv Y_i - bX_i$. Show that the first derivative of $\hat{r}(b)$ is $\hat{r}'(b) = \frac{-2}{n}\sum_{i=1}^{n} D_i(b)X_i$.
   e. (10) Explain why it's reasonable, under our assumptions, to estimate $\mathrm{Var}[\hat{r}'(b^*)]$ by

$$\hat{J}_n \equiv \frac{4}{n^2}\sum_{i=1}^{n} D_i^2(\hat{b})X_i^2$$

   "Reasonable" here means you don't need to give a formal proof, but you should give reasons to explain why $\hat{J}_n$ is connected to $\mathrm{Var}[\hat{r}'(b^*)]$. *Hints*: (i) What're the expectations of the summands in the definition of $\hat{J}_n$? (ii) Use sub-problem (a).
   f. (8) Find an expression for the standard error of $\hat{b}$, the minizer of $\hat{r}(b)$. Your answer should involve both $\hat{J}_n$ and the sample variance of X (and possibly other things). *Hints*: (1) Lectures 3 and 4; (2) remember that the standard error of an estimator is *defined* as the square root of its sampling variance; "the standard error of the mean", $\sigma/\sqrt{n}$, is the standard error of one particular estimator (which?), but every estimator has its own standard error.
   g. (5) Now assume that $Y = b^*X + \epsilon$ where ϵ is IID with mean 0 and variance σ^2. (That is, the usual linear-model assumptions hold.) Show that your expression for the standard error from the last sub-problem will converge on $\sigma/\sqrt{n\mathrm{Var}[X]}$ for large n.

Note: in this problem we do not assume that the linear regression model is right, or, if the relationship between Y and X *is* linear, assume that the noise around the regression line has constant variance. What we've just done, in the next-to-last sub-problem, is the calculation of a "robust standard error" (because it's still valid if the usual assumptions are broken). In particular, this is a "heteroskedasticity-consistent" (HC) robust standard error (because it works even if the noise is "heteroskedastic", i.e., does not have constant variance).

2. *Propagation of error and uncertainty in predictions.* The following technique, called "propagation of error", "the delta method", or "propagation of uncertainty", is often useful in simplifying complicated calculations about the variances of functions.
   a. (5) Suppose that $M = f(T)$, where the random variable T has expectation $\mu\$$ and variance σ^2. Use a Taylor expansion of f to explain why $\mathrm{Var}[M] \approx \left(f'(\mu)\right)^2\sigma^2$ when σ^2 is small.
   b. (6) Now suppose that $M = f(T_1, T_2, \ldots T_d)$, where T_i has expectation μ_i and variance σ_i^2. Assume

the T_i are uncorrelated with each other. Assuming all the σ_i^2 are small, explain why

$$\mathrm{Var}\left[M\right] \approx \sum_{i=1}^{d} \left(\frac{\partial f}{\partial t_i}(\mu_1, \ldots \mu_d)\right)^2 \sigma_i^2$$

c. (6) Suppose the situation is as in (b), but that $\mathrm{Cov}\left[T_i, T_j\right] = \rho_{ij}$, not necessarily equal to 0. Explain why

$$\mathrm{Var}\left[M\right] \approx \sum_{i=1}^{d} \left(\frac{\partial f}{\partial t_i}(\mu_1, \ldots \mu_d)\right)^2 \sigma_i^2 + 2\sum_{i=1}^{d-1}\sum_{j=i+1}^{d} \left(\frac{\partial f}{\partial x_i}(\mu_1, \ldots \mu_d)\right)\left(\frac{\partial f}{\partial t_j}(\mu_1, \ldots \mu_d)\right)\rho_{ij}$$

d. (3) Define $\mathbf{\Sigma}$ as the matrix with diagonal entries $\sigma_1^2, \ldots \sigma_d^2$, and off-diagonal entries ρ_{ij}. Is

$$\mathrm{Var}\left[M\right] \approx \nabla M(\mu_1, \ldots \mu_d) \cdot (\Sigma \nabla M(\mu_1, \ldots \mu_d)) \ ?$$

If so, explain why; if not, explain why not, and give a correct expression if possible.

e. (6) Now suppose that our model / strategy / prediction rule makes the prediction $s(x; \theta_1, \ldots \theta_d)$ on information x when the parameters are $\theta_1, \ldots \theta_d$. We have a variance-covariance matrix $\mathbf{c}$ for our estimated parameters $\hat{\theta}_1, \ldots \hat{\theta}_d$ (perhaps from the "usual asymptotics", or from something like problem 1, or perhaps from the Oracle). Explain, in words, how we could use $\mathbf{c}$, and the earlier parts of this problem, to get a variance for our prediction at $X = x$. What, if anything, would we need to calculate, beyond $\mathbf{c}$?

3. *ERM and the 0-1 loss.* In this problem, assume that we're doing classification with the 0-1 loss, and that our X variable is a two-dimensional set of coordinates on a plane. Our available rules/strategies/classifiers are all of the form "say 1 if $a \cdot x \geq b$ and say 0 otherwise", where a is a vector and b is a scalar. (These are called "linear classifiers".) Finally, abbreviate $\mathbb{P}\left(Y = 1 | X = x\right)$ as $p(x)$, and suppose that $p(x)$ is a smooth function of x.

a. (5) Explain why the true risk is a smooth function of a and b.

b. (5) Explain why the empirical risk is not a smooth function of a and b. In particular, explain why the empirical risk is a *discontinuous* function of the parameters. (You may find it helps to draw a picture.)

c. (5) Explain why applying "the usual asymptotics" (from lectures 3 and 4) to the 0-1 loss is dubious.

d. (5) Would your conclusion in (c) be altered if instead of linear classifiers, we used circles of varying radii and centers?

4. (1) Roughly how much time did you spend on this assignment?

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.