Homework 4: Underdetermined Problems and Model Validation

Due: 11/1/2021

1. *Exit polling.* Recall the exit polling example from class: Define μ to be the probability that a voter will vote for Biden in the 2020 US presidential election. Suppose that we poll n voters each at p different polling places. Let

$$z_{i,j} = \begin{cases} 1 & \text{if the } i\text{-th voter at location } j \text{ voted for Biden} \\ 0 & \text{otherwise} \end{cases}$$

Let $\nu_j = \frac{1}{n}\sum_{i=1}^{n} z_{i,j}$, the sample mean at polling location j.
The probability of obtaining k votes for Biden at a given location is given by the binomial distribution:

$$\mathbb{P}[k|n,\mu] = \binom{n}{k}\mu^k(1-\mu)^{n-k}$$

   (a) Assume the sample size $n = 10$ at each polling location. If all the voters have $\mu = 0.05$ compute the probability that at least one polling location will have $\nu_j = 0$ for the case of $p = 1$, $p = 1000$, and $p = 1,000,000$. Repeat for $\mu = 0.8$.

   (b) For the case $n = 6$ and $p = 2$ with $\mu = 0.5$ for both locations, plot the probability

$$\mathbb{P}[\max_j |\nu_j - \mu| > \epsilon]$$

   for $\epsilon \in [0, 1]$ (the max is over polling locations). On the same plot show the bound that would be obtained using the Hoeffding Inequality. Remember that for a single location, the Hoeffding bound is

$$\mathbb{P}[|\nu - \mu| > \epsilon] \leq 2e^{-2n\epsilon^2}$$

   Hint: if events A and B are independent, $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A]\mathbb{P}[B]$

2. *Rats!*
Your friend Alice works in a lab where she studies the genetic indicators of rat intelligence. To do this, she sequences their DNA, and tests how long it takes for the rats to

finish a maze. She has collected all of the rats' genetic data into a matrix X. Each row contains data for one rat and the columns are different alleles (a variant of a gene). The entry $X[i, j]$ is 1 if rat i has allele j but 0 otherwise. She's collected all of the rat maze finish times into the vector y.

One day she tells you that she has run a linear regression on her dataset to understand which alleles are important for rat intelligence. She has found a model w_A that fits the problem with very small residual sum of squares error.

(a) Alice is excited because she's found two alleles with very large coefficients which she thinks might be important for understanding rat intelligence. The coefficient for the first allele, $w_A[1]$, is very large and positive. The coefficient for the second allele, $w_A[2]$, is very negative. How do you interpret these coefficients? What would you predict about how long it would take a rat with each of these genes to finish the maze?

(b) Bob works in the same lab as Alice. The next day, he tells you that he's solved the same least squares problem with an error equally as small. But his coefficients, w_B, are very different from Alice's! His coefficient for the first allele, $w_B[1]$, is very negative while $w_B[2]$ is very positive. Why do you think Bob's coefficients are so different from Alice's?

(c) Candace, who also works in the same lab, tells you that she has found a rat with a genetic code that has never been seen before. All other lab rats had both allele 1 and allele 2. This new rat only has allele 1. Alice's model predicts that this rat will take 535 years to finish the maze. Bob's model predicts the rat will finish in -534 years. How will you answer Alice, Bob and Candace if they ask you to make a prediction?

(d) If Bob and Alice gave you their dataset X and y, what would you do to make a better prediction? *(There isn't necessarily a right answer here.)*

3. *Ridge regression and the SVD*

Recall the singular value decomposition from class. The SVD decomposes a matrix $X \in \mathbf{R}^{n \times d}$ into a product of simpler matrices

$$X = U \Sigma V^T,$$

where

- $U \in \mathbf{R}^{n \times r}$ and $V \in \mathbf{R}^{d \times r}$ have orthonormal columns

$$U^T U = I, \quad V^T V = I.$$

- $\Sigma \in \mathbf{R}^{r \times r}$ is diagonal with $\Sigma_{ii} = \sigma_i > 0$ for $i = 1, \ldots, r$.
- $r \leq \min(d, n)$ is the rank of X.

(This version is sometimes called the thin SVD.) In class we showed how to use the SVD to solve the least squares problem. In particular, we showed that $w^{\text{SVD}} = V\Sigma^{-1}U^T y$ satisfies the normal equations.

In class, we saw that when X is not full rank, least squares has no unique solution. We introduced ridge regression to deal with these underdetermined problems. In this problem we will draw a connection between w^{SVD} and w^{ridge}.

We will show that the w^{SVD} is the solution to least squares that minimizes the Euclidean norm $||w||$. That is, if we have a vector w s.t. $||y - Xw|| = ||y - Xw^{\text{SVD}}||$, then $||w|| \geq ||w^{\text{SVD}}||$.

(a) Let w be some solution to the least squares problem. It must satisfy the normal equations. Rewrite the normal equations ($X^T X w = X^T y$) using the matrices U, V, and Σ.

(b) Define a matrix V^c whose columns form an orthonormal basis for the orthogonal complement of V. That is, $(V^c)^T V = 0$ and $(V^c)^T V^c = I_{d-r}$.
We can decompose $w = Vw^{\|} + V^c w^{\perp}$ where $w^{\|} = V^T w$ and $w^{\perp} = (V^c)^T w$. Use the normal equations to derive a formula for $w^{\|}$ in terms of U, Σ, and V. In particular, this shows that any solution w to least squares must have the same $w^{\|}$.

(c) Decompose $w^{\text{SVD}} = Vw^{\text{SVD}\|} + V^c w^{\text{SVD}\perp}$ using the same decomposition. Derive a formula for $w^{\text{SVD}\|}$ and $w^{\text{SVD}\perp}$.

(d) Prove that $\|w\|^2 = \|V^T w\|^2$ if $V^T V = I_r$.

(e) Write out the Euclidean norm of w in terms of $w^{\|}$ and $w^{\perp}$. Conclude that for any solution w to least squares, $||w|| \geq ||w^{\text{SVD}}||$
Hint: Use the Pythagorean Theorem.

Through parts (a) – (d), we have shown that w^{SVD} is the solution to least squares with minimum norm. Recall that the ridge regression estimator,

$$w^{\text{ridge}}(\lambda) = \operatorname{argmin} ||y - Xw||^2 + \lambda ||w||^2,$$

minimizes the mean squared error and a weighted norm of w.

(f) Show that

$$\lim_{\lambda \to 0} w^{\text{ridge}}(\lambda) = w^{\text{SVD}}$$

Hint: Write w^{ridge} and w^{SVD} in terms of the σ_i's.

4. *Plotting bias and variance*: In this problem, we'll investigate the bias and the variance of two different estimators: a linear estimator and a cubic estimator. We'll see (once again) that fitting the data more precisely is not always a good idea.

Work through the notebook `https://github.com/ORIE4741/homework/blob/master/BiasVariance.ipynb`.