STAT 2270: Data Mining
University of Pittsburgh
Fall 2020

Homework 2

1. Suppose we have a dataset $\mathcal{D}_n = \{X_1, ..., X_n\}$ of size n from which we want to take a resample of size k uniformly at random with replacement (i.e. if $k = n$, then this is just a standard bootstrap sample).

   (a) What is the probability that any given observation in the original sample appears in the resample?

   (b) What is the expected number of unique observations from the original sample that appear in the resample? What is the expected proportion?

   (c) Now suppose $k = n$. What does this expected proportion converge to as $n \to \infty$?

   (d) There is a common notion of the "0.632 rule" for the bootstrap. Explain where this comes from given your answer in part (c).

2. Suppose we have a training dataset $\mathcal{D}_n$ with n observations of the form $Z_i = (X_i, Y_i)$ where the X_i is a vector of features and $Y_i \in \{0, 1\}$ is the binary response. Suppose that we draw B bootstrap samples $\mathcal{D}_1^*, ..., \mathcal{D}_B^*$. Let p_B denote the probability that any given observation in the original training sample appears in more than half of the bootstrap samples. Prove or disprove the following claim: the probability that *all* observations appear in at least half of the bootstrap samples is bounded above by p_B^n.

3. Show that k-nearest neighbors with $k \geq 1$ is a linear smoother and calculate its degrees of freedom.

4. Suppose we have a dataset $\mathcal{D}_n = \{(X_1, Y_1)..., (X_n, Y_n)\}$ but instead of using any information about the predictor variables, we define our regression estimator as simply the empirical mean over the observed response values. That is, for all x, we define

   $$\hat{f}(x) = \bar{y}$$

   Calculate the degrees of freedom for this estimator.

5. In class, we discussed how k-nearest neighbors is not quite the same as kernel regression with a uniform kernel. Give a simple example (one covariate) that demonstrates this.

6. Bootstrapping can be used to estimate the sampling distribution of *most* statistics, but not all:

   (a) Give an example of a statistic for which the bootstrap struggles to provide an accurate sampling distribution.

   (b) Demonstrate your example from part (a). Take 1000 samples of size 25 from a standard normal distribution, calculate your statistic on each one, and create a histogram.

(c) Now select one of your 1000 samples uniformly at random and take 1000 bootstrap samples. With each bootstrap sample, calculate the statistic and create a histogram.

(d) How does your bootstrap distribution differ from the "true" sampling distribution? Give a brief explanation for why the bootstrap fails in your case.

7. Consider the function

$$f(x) = (x-2)^4 - 4(x-2)^3 + 5(x-2).$$

(a) Plot this function (in the form of a line) in black for values of x from 0 to 6. Now consider the model $Y = f(X) + \epsilon$. Generate data by taking a random sample of size 50 from this model with $X_i \overset{iid}{\sim} Unif(0,6)$ and $\epsilon_i \overset{iid}{\sim} N(0,5^2)$ for $i = 1, ..., 50$ and add these points to your existing plot in a different color.

(b) Consider doing kernel regression on the data from part (a) using a Gaussian kernel with bandwidth h. For an appropriate (wide enough) range of bandwidth values, plot the 10-fold CV error of the model against h. Use this plot to select an appropriate value of h and call this h_G. Again plot the original function in black and now add this Gaussian kernel estimate to the plot (in the form of a line) in blue.

(c) Now suppose we switch to the Epanechnikov kernel. Using the value h_G you found in part (b), add the Epanechnikov kernel estimate to the plot as a dotted red line.

(d) Now repeat part (b) using the Epanechnikov kernel and call the optimal bandwidth you find h_E. On the same plot from part (c), add the Epanechnikov kernel estimate with bandwidth h_E as a solid red line.

(e) Compare the two different fits of the Epanechnikov kernel regression estimate. How much did the bandwidth change in part (d)? How does the best Epanechnikov kernel estimate compare with the best Gaussian kernel estimate?

8. Consider a standard linear model setup and suppose we have 50 (independent) covariates, but that only 5 of these (say the first five) have non-zero coefficients, with $\beta_i = i$ for $i = 1, ..., 5$. Finally, suppose we observe covariate values uniformly at random from the unit rectangle with standard iid Gaussian noise (i.e. $\epsilon \overset{iid}{\sim} N(0,1)$).

(a) Take a sample of size 25 and use 5-fold cross validation to fit a lasso model and record the resulting error. Call this $Err_{Lasso,1}(25)$.

(b) Repeat part (a) 1000 times to obtain $Err_{Lasso,1}(25), ..., Err_{Lasso,1000}(25)$. In addition, each time you fit a lasso model, determine which covariates are given non-zero coefficient estimates. Take these covariates and fit a standard linear model on only these. Record this error to obtain $Err_{OLS,1}(25), ..., Err_{OLS,1000}(25)$.

(c) In general, define

$$\overline{Err}_{Lasso}(n) = \frac{1}{n}\sum_{i=1}^{n} Err_{Lasso,i}(n)$$

and equivalently for the OLS errors. Calculate $\overline{Err}_{Lasso}(25)$ and $\overline{Err}_{OLS}(25)$ and take the difference.

(d) Explore if/how the difference in errors changes with n. For example, you might consider something like $n = 25, 50, 100, 250, 500, 1000$.

(e) You might expect things to behave differently depending on the problem set-up. What if there were even more parameters and many of them had non-zero coefficients? What if the covariates weren't independent but (at least some) exhibited some correlation? Make some modifications to the original problem to explore these aspects and see what you can conclude.

(e) When (if ever) would it be good to do OLS after Lasso? Intuitively, why might it make sense to do that? This is closely related to the idea of the *relaxed* lasso [Meinshausen (2007), 'Relaxed lasso', Computational Statistics & Data Analysis]. Have a look at this paper and give a short description of what is being suggested. Code up this estimator and apply it to some of the settings above to see when it tends to outperform the Lasso and OLS methods alone.

9. Assume the same original setup as in problem 7: a standard linear model with 50 independent covariates with non-zero coefficients for the first 5 variables only.

(a) Draw a sample of size 25 with iid Gaussian noise assuming covariate values are sampled uniformly at random from the unit rectangle. Fit a lasso model using 5-fold CV to choose λ and fix this value of the tuning parameter.

(b) Suppose we want to estimate the degrees of freedom for this model (i.e. we didn't know the theoretical result we talked about in class). We could bootstrap the rows of our data so that each time we get dataset of the form $\tilde{\mathcal{D}} = \{(\tilde{x}_1, \tilde{y}_1), ..., (\tilde{x}_{25}, \tilde{y}_{25})\}$ where we save the values $\tilde{y}_i$ as well as the fitted values $\hat{y}_i$ for $i = 1, ..., 25$. Repeat this process with B bootstrap samples and at the end, calculate the empirical covariance between each of the $\tilde{y}_i$ and the $\hat{y}_i$ which we can use as a proxy for $\text{Cov}(\hat{y}_i, y)$:

$$\widehat{\text{Cov}}(\hat{y}_i, y_i) = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{y}_i^{(b)} - \hat{\mathbb{E}}(\hat{y})\right) \left(\tilde{y}_i^{(b)} - \hat{\mathbb{E}}(\tilde{y})\right).$$

We could then estimate the degrees of freedom by

$$\hat{\text{df}}(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \widehat{\text{Cov}}(\hat{y}_i, y_i).$$

Using your data and λ from part (a), try this with B = 1000.

(c) The estimator from (b) often doesn't work that well. Why not? Hint: It may not be what you first think – consider what's fixed vs. random.

(d) A better way to estimate the degrees of freedom is with a *residual* bootstrap. This is carried out in exactly the same fashion as part (b) except that now each bootstrap dataset will be of the form $\tilde{\mathcal{D}} = \{(x_1, \tilde{y}_1), ..., (x_{25}, \tilde{y}_{25})\}$ where

$$\tilde{y}_i = \hat{y}_i + \epsilon^*$$

where ϵ^* selected uniformly at random from the 25 residuals calculated on the original data. Do this procedure to estimate the degrees of freedom with $B = 1000$. Why would you expect this to be a better estimate?