DS 100: Principles and Techniques of Data Science **Date: April 19, 2019**

Discussion #10

Name:

Tree-Based Methods and Cross-Validation

1. Now that we're more familiar with tree based methods, we can discuss how to properly implement them to solve regression and classification problems. One of the most difficult aspects of using these methods is selecting optimal hyperparameters (e.g. the depth of the tree, or the number of trees in a Random Forest). Luckily, cross-validation methods can be used for this task.

   (a) Recall that cross-validation (CV) can be used to assess the risk (i.e. the performance) of a model using only the learning set. Describe in your own words how this is accomplished using K-fold CV. Can you illustrate the procedure of 5-fold CV?

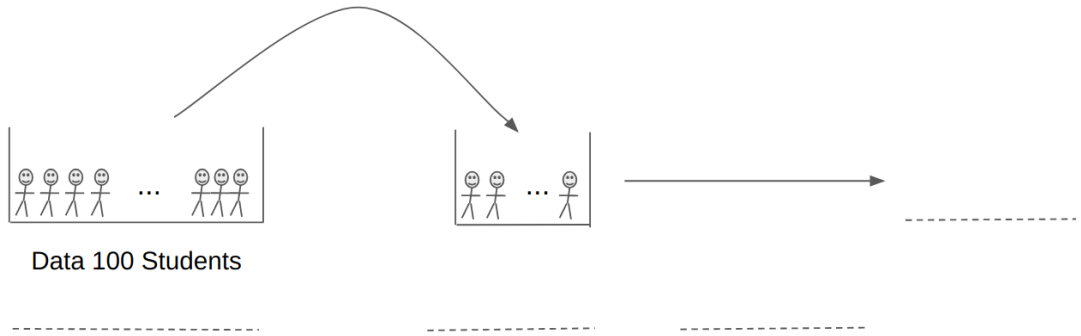
   (b) Now suppose we wish to fit a tree in a binary classification context. We can use CV methods to select the optimal value for the maximum tree depth parameter. Write the steps to accomplish this using K-fold cross-validation and a vector d of possible maximum tree depths, $d = (1, 2, 3, ..., N)$.

   (c) Suppose now you wish to fit a Random Forest model instead of a classification tree. How would you modify your pseudo code to identify the appropriate maximum depth and the number of trees to include in your model?

The Bootstrap

2. Suppose we wish to estimate the proportion of cigarette smokers among the students enrolled in Data 100.

   (a) Using a box model, we wish to represent the process of estimating the proportion of cigarette smokers. Fill the dotted lines in the image below using the following words: population, sample, compute estimate, and $\hat{p}$. Given that there are about 900 students enrolled in Data 100, should the random sample be drawn with or without replacement?



(b) Now that we have our estimate, we would like to carry out some statistical inference by creating a confidence interval. Assuming that our sample is a good representation of the population, we can use the bootstrap method to do this. Using a box model, illustrate the process of generating bootstrap estimates for the proportion of smokers in Data 100.

(c) Describe how you would compute the 95% confidence interval of proportion of smokers using bootstrap estimates.

(d) Suppose estimate that the proportion of smokers in Data 100 is 3.5%, and the 95% confidence interval found using the bootstrap is $(2.3\%, 4.7\%)$. How do we interpret this interval?

3. We can use the bootstrap to carry out inference on the slope of a simple linear regression. Recall that a simple linear regression model is defined as follows

$$E[Y|X] = \theta_0 + \theta_1 X$$

where (X, Y) are continuous random variables. Y is the response, X is the feature. Using the data to estimate the intercept and the slope, we arrive at the following equation:

$$f_{\hat{\theta}}(x) = \hat{\theta}_0 + \hat{\theta}_1 x_1$$

(a) Using a box model, describe the process of computing the 95% confidence interval of $\hat{\theta}_1$.

(b) Now that we have some intuition for how the bootstrap works in a simple linear regression, let's think about how we might implement this for a multivariate linear regression. Suppose we wish to fit a model of the following form:

$$f_{\hat{\theta}}(x) = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \cdots + \hat{\theta}_p x_p$$

and we would like to generate confidence intervals around our estimates of θ. Outline in pseudo-code a non-parametric bootstrap based approach to estimate the 95% confidence interval around each θ_i. Assume there are n data points.