STATS 503, Homework 1

Please use a Jupyter notebook to write up your solutions. Submit your work through Canvas by uploading an html file that contains the html export of your notebook file. The html file should include all code (no cells hidden). Use markdown formatting to make clear which part of your notebook corresponds to each problem.

1. Consider the population of students taking STATS 503 this semester at the University of Michigan.

   (a) [**3 POINTS**] Name three variables **related to academics** that you could collect or measure about each student in this population. Of these three variables, one must be ordinal, one must be categorical, and one must be continuous.

   (b) [**3 POINTS**] Suppose you have collected a dataset containing these variables for the population of students taking STATS 503 this semester. Now consider using this dataset to make inferences about a different population. Name another population about which we could plausibly make inferences. Name another population that would be more difficult to make inferences about.

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide the number of samples (n) and the number of features (p).

   (a) [**3 POINTS**] We collect a set of data on the top 500 firms in the United States. For each firm, we record the profit, the number of employees, the industry, and the CEO salary. We are interested in understanding which factors affect CEO salary.

   (b) [**3 POINTS**] We are considering releasing a new product and want to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product, we have recorded whether it was a success or failure, the price charged for the product, the marketing budget, the competition price, and ten other variables.

   (c) [**3 POINTS**] We are interested in predicting the percentage change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence, we collect weekly data for all of 2012. For each week, we record the percent change in the USD/Euro, the precent change in the US market, the percent change in the British market, and the percent change in the German market.

3. [**3 POINTS**] You have a dataset. You are considering a collection of different methods that you could apply to this dataset in order to estimate the regression function. The methods range from very inflexible (only capable of representing a small class of true regression functions) to very flexible (capable of fitting a very large class of true regression functions).

   (a) True or false: typically, the bias of your estimate will be lower with more flexible methods.

   (b) True or false: typically, the variance of your estimate will be lower with more flexible methods.

(c) True or false: typically, irreducible error will be lower with more flexible methods.

4. [**4 POINTS**] Consider gathering a dataset and using it to estimate the regression function. Assume the data-generating distribution is well behaved (e.g., assume that the true regression function is continuous) and typical of what we would find in real-world settings. Answer true or false.

   (a) As the size of your training dataset tends to infinity (for a fixed number of neighbors, K), the bias of the KNN regression estimate will tend to 0.

   (b) As the size of your training dataset tends to infinity (for a fixed number of neighbors, K), the variance of the KNN regression estimate will tend to 0.

   (c) As the size of your training dataset tends to infinity, the bias of a least squares linear regression estimate will tend to zero.

   (d) As the size of your training dataset tends to infinity, the variance of a least squares linear regression estimate will tend to 0.

5. Download the `college_train` data set from canvas and load it as a data frame.

   (a) [**1 POINTS**] Access the `shape` property of your data frame to compute the number of samples we have and the number of variables measured about each sample.

   (b) [**1 POINTS**] Compute the mean and standard deviation of the `Books` feature.

   (c) [**2 POINTS**] For how many samples is the `Terminal` feature at least 90?

   (d) [**2 POINTS**] Create a new dataframe that only includes samples where `Private` is equal to `Yes`.

      i. What is the mean value of `Books` in the new dataframe?

      ii. What is the standard deviation of `Books` in the new dataframe?

   (e) [**1 POINTS**] Use a for loop to print the following text.

```
Hello world 1.
Hello world 2.
Hello world 4.
Hello world 8.
Hello world 16.
```

   (f) [**1 POINTS**] Create a markdown Jupyter cell with the following contents.

This is a section heading

In this section I have a bigger formula,

$$\sqrt{\frac{\alpha}{\sqrt{\beta^2 + \cos 3}}}$$

and an in-line formula, $\sqrt{3}$. I made **this text** bold.

Total points on this assignment: 30