AMS 598: Big Data Analysis (Fall 2024)

Assignment #3

Due Nov 5th, 2024

- You need to submit 1) a report in PDF and 2) your .ipynb file, both to Brightspace.
- Your PDF report should include (1) answers to the non-programming part, and (2) necessary analysis of the programming part.
- Please put all your files (PDF report and .ipynb file) into a compressed file named "Assi3-FirstName LastName.zip"
- All students are highly encouraged to typeset their reports using Word or LaTeX. In case you decide to hand-write, please make sure your answers are clearly readable in scanned PDF.
- Unlimited number of submissions are allowed and the latest one will be timed and graded.
- Only write your code between the following lines. Do not modify other parts.
  ### YOUR CODE HERE

  ### END YOUR CODE
- Please read and follow submission instructions. No exception will be made to accommodate incorrectly submitted files/reports.

---

1. [15 pts] **Cross-Validation**. Suppose we want to compute cross-validation error on 100 training examples. We need to compute error N1 times, and the cross-validation error is the average of the errors. To compute each error, we need to build a model with data of size N2, and test the model on the data of size N3.
   (a). If we use 5-fold cross-validation, what are the appropriate numbers for N1, N2, N3?
   (b). If we use 10-fold cross-validation, what are the appropriate numbers for N1, N2, N3?
   (a). If we use leave-one-out cross-validation, what are the appropriate numbers for N1, N2, N3?

2. [35 pts] (Coding Task) **Bootstrapping.** Bootstrapping is a resampling technique used in data science to estimate the distribution of a statistic by repeatedly sampling with replacement from the observed data. In this assignment, you are tasked with analyzing the effectiveness of a new drug using a small dataset consisting of only 8 measurements (given in notebook); each measurement is associated with an effectiveness score for a particular individual.
   Useful Resource: Youtube Video Bootstrapping Main Ideas!!!:
   https://www.youtube.com/watch?v=Xz0x-8-cgaQ&ab_channel=StatQuestwithJoshStarmer
   (a) [5 pts] Complete the ***sample()*** method, which generates a new bootstrap dataset (the dataset resulting from one bootstrapping round) from the original dataset.
   (b) [5 pts] Complete the ***generate_bootstrap_datasets()*** method, which generates a number of bootstrap datasets using the ***sample()*** method. For (a) and (b), you can refer to slides 29/44 of lecture "Lecture12_ Resampling".

(c) [5 pts] Complete the ***calculate_means()*** method, which calculates the **mean** for each bootstrap dataset.
(d) [10 pts] Complete the ***plot_histogram()*** method, which plots a histogram of means. You should set the number of bins to be a number from 20 to 50. You can refer to slides 29/44 of lecture "Lecture12_Resampling".
(e) [10 pts] Summarize your results and provide insights on whether or not this drug is effective, and explain why.

3. [50 pts] **Decision Tree**. Consider the following table of observations:

No.	Outlook	Temperature	Humidity	Windy	Play Golf?
1	sunny	hot	high	false	N
2	sunny	hot	high	true	N
3	overcast	hot	high	false	Y
4	rain	mild	high	false	Y
5	rain	cool	normal	false	Y
6	rain	cool	normal	true	N
7	overcast	cool	normal	true	Y
8	sunny	mild	high	false	N
9	sunny	cool	normal	false	Y
10	rain	mild	normal	false	Y
11	sunny	mild	normal	true	Y
12	overcast	mild	high	true	Y
13	overcast	hot	normal	false	Y
14	rain	mild	high	true	N

From the classified examples in the above table, construct two decision trees (by hand) for the classification "Play Golf." For the first tree, use Temperature as the root node. (This is a really bad choice.) Continue the construction of tree as discussed in class for the subsequent nodes using information gain. Remember that different attributes can be used in different branches on a given level of the tree. For the second tree, follow the Decision Tree Learning algorithm described in class. At each step, choose the attribute with the highest information gain. Work out the computations of information gain by hand and draw the decision tree.