DSCI 552: Machine Learning for Data Science

Quiz 1

Instructors: Kristina Lerman (lerman@isi.edu)

burghard@usc.edu(burghard@usc.edu)

Deadline: **Monday, January 25, 2021 at 10 am PT**

You can access the quiz on Blackboard. The deadline to submit solutions is next Monday at 10 am (PST). As long as the quiz is open, you will be able to send multiple answers (only the last submitted answer will matter).

During this quiz, you will learn the first steps of a data modeling workflow. Suppose we are interested in modeling the pitch of human voice varies. The data (data-dsci552.csv) contains real measurements of voice *pitch* from a population of women (F) and men (*M*) in different *scenarios*. You will perform elementary statistical analysis of the data with the goal of quantifying variations in pitch.

Question 1 (2 points)

The first step is to clean the data. Are there non-sensical measurements? Are there any outliers? Describe your data cleaning steps.

Question 2 (2 points)

Exploratory analysis often includes plotting the histogram of your outcome variable (also known as response). What is the outcome variable in this problem? Create a histogram. Hint: Python library numpy can be useful to create a histogram.

Question 3 (2 points)

Calculate the mean, median, standard deviation for each gender subgroup, as well as for the entire population. Hint: use the Python library pandas's .loc[] function to separate data by gender.

Question 4 (2 points)

Create a boxplot of pitch variation in different scenarios. Don't forget to label the axes. You can paste the plot in your answer below. Hint: use the Python library seaborn's "boxplot" function.

Question 5 (2 points)

Is the variation in pitch due to gender more significant than variation due to context (different scenarios)? Explain your answer.

Question 6 (2 points)

Describe how you could apply what you have learned through modeling in a real-world application.