CINF/CSCI 5931 Big Data Analytics Fall 2017

Assignment 2 Supervised Learning Techniques

Post Date: **Oct 12th, 2017** **Due Date**: **Oct 26th, 2017 by 11:59pm**

Important Notes:

1. This is **AN INDIVIDUAL** assignment, any form of collaboration is **PROHIBITED**.
2. You will turn in your codes in Jupyter Notebooks.
3. Make sure you **strictly follow** the turn-in instructions as follows, any deviation from the instruction will incur penalties :
   - **a.** Create a jupyter notebook for each problem, name the notebook **YourLastName_P[problem number].ipynb.**
   - b. In your notebook, use comments and markdown to explain your code as much as you can.
   - c. Make sure you clear all output in your notebooks before submission.
   - d. Combine all your notebooks and any other output files (if applicable) into one zip file and name it **YourLastName_A2** for submission.
4. **NO late submission will be accepted!**

Problem 1 (40 points)

An organization would like to keep its good employees because they usually possess valuable experience. Therefore, it is useful for a company to predict the likelihood an employee will leave so mitigations can be done.

You are given a dataset (named A2_P1) from human resource, your task is to build a classification model to predict if an employee will leave or not. You are required to complete this task using two different classification techniques. Your code should be in Python and use MLlib.

You also need to evaluate your model and provide measures of your models' error. Which technique works better?

Problem 2 (60 points)

Which country has the happiest citizens in the world? The World Happiness Report may tell you something about that. In this problem, you are given world happiness survey result data sets for three years (2015, 2016, and 2017). The ranking in terms of happiness of the countries is provided. Your job is to use regression to create a model that can predict the ranking (note that ranking may be an indirect results) as close as possible to the actual results. To help you understand better of the meaning of the attributes, you can refer to http://worldhappiness.report/faq/. You could start your model building using one year's data and apply it on other two years.

In addition, use the basic statistics and other data manipulation methods to answer the following questions (with code):

1. From 2015 to 2017, which country's happiness ranking increased the most?
2. From 2015 to 2017, which country's happiness ranking decreased the most?
3. For each year, provide the ranking of the happiest continents.