CDS503: Machine Learning
Academic Session: Semester 1, 2018/2019

School of Computer Sciences, USM, Penang

LAB EXERCISE (LAB 2)

Lab 2 Exercise

Look at the census income data (*adult_train.csv*) that is uploaded in eLearn. Look at each attribute and see what type of data it has.

Question 1: Do any **pre-processing** to data as *necessary*. Then, answer the following questions:

- What are the **types** of the attributes?
  - **age**
  - **workclass**
  - **fnlwgt**
  - **education**
  - **education-num**
  - **marital-status**
  - **occupation**
  - **relationship**
  - **race**
  - **sex**
  - **capital-gain**
  - **capital-loss**
  - **hours-per-week**
  - **native-country**
-
- Is there any **empty or null** values? What approach you use to address them (remove, replace, etc.)? and why?
- Any **unused** or **irrelevant** columns/attributes? What do you do to them?
- What attribute(s) might be **useful**?

Question 2:

Experiment with KNN machine learning algorithm to *predict* whether income **exceeds** $50K/year based on census income data (adult_test.csv). Use *default* KNN configurations and try **at least** two different values of *k*. Try conduct also with *custom* KNN configurations with **at least** 5 fold cross-validation. Compare the two KNN and specify your findings. Do higher values of *k* lead to better performance? Do cross-validation effect KNN performance?

Post your solution on Lab 02 Submission on **elearn@usm**. Make sure you to include your name and lab# on the submission post.

Format: in .ipynb

The due date is **21 October 2018 23:59**