Lab3: Data Preparation

CPSC429/529 Machine Learning

In this lab assignment, you are given a adult dataset (`adult.data`) (`https://archive.ics.uci.edu/ml/datasets/adult`) and do data preprocessing. Specifically, you will do the followings:

1. Use `StandardScaler` to standardize the column of `fnlwgt`. Update the column of `fnlwgt` in the dataframe `df`, and print out the first five rows as shown below.

	age	workclass	fnlwgt	edu	edu-num
0	39	State-gov	-1.063611	Bachelors	13
1	50	Self-emp-not-inc	-1.008707	Bachelors	13
2	38	Private	0.245079	HS-grad	9
3	53	Private	0.425801	11th	7
4	28	Private	1.408176	Bachelors	13

2. Use `KBinsDiscretizer` to discretize the column of `age` (**parameter inputs**: 5 bins, ordinal, uniform) . Update the column of `age` in the dataframe `df`, and print out the first five rows as shown below.

	age	workclass	fnlwgt	edu	edu-num
0	1.0	State-gov	-1.063611	Bachelors	13
1	2.0	Self-emp-not-inc	-1.008707	Bachelors	13
2	1.0	Private	0.245079	HS-grad	9
3	2.0	Private	0.425801	11th	7
4	0.0	Private	1.408176	Bachelors	13

3. Use `OneHotEncoder` to encode the column of `race`. Save the output into a new dataframe, and print out the first five rows as shown below.

	race_ Amer-Indian-Eskimo	race_ Asian-Pac-Islander	race_ Black	race_ Other	race_ White
0	0.0	0.0	0.0	0.0	1.0
1	0.0	0.0	0.0	0.0	1.0
2	0.0	0.0	0.0	0.0	1.0
3	0.0	0.0	1.0	0.0	0.0
4	0.0	0.0	1.0	0.0	0.0

The jupyter notebook skeleton of lab 2 (`Lab3_DataPreparation.ipynb`) is given to you, so you can complete the remaining parts (**Hints:** Refer lecture slides). Submit this jupyter notebook (`Lab3_DataPreparation.ipynb`) to D2L.