EE 215 - Coding Assignment I

The goal of this assignment is playing around with a dataset and using correlation coefficient to understand the relation between the features. You will work with the Adult dataset from UCI Machine Learning Repository. The dataset link is https://archive.ics.uci.edu/ml/datasets/Adult. This dataset contains several features of individuals. The features of interest, for us, are described as follows:

- income: >50K, <=50K.
- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- sex: Female, Male.
- capital-gain: continuous.
- hours-per-week: continuous.

Approach

Our target feature is *income* and we will study the correlation between *income* and other features (age, workclass, ...). Given two feature vectors (e.g. income and age of 50 people), similar to random variables, we can calculate correlation, covariance, and correlation coefficient by treating them as samples from a joint distribution. For instance, suppose three people (corresponding to rows) have the following income, age, workclass (corresponding to columns) values:

$$\begin{bmatrix} > 50K,\ 30,\ Private \\ \leq 50K,\ 20,\ Federal-gov \\ > 50K,\ 40,\ State-gov \end{bmatrix}$$

To calculate correlation between these, we need numbers. Convert income to 0/1 by assigning 1 to >50K and 0 else. Now, we can calculate the correlation coefficient between income and age by using the vectors $X = [1\ 0\ 1]$ and $Y = [30\ 20\ 40]$. Viewing ith person as a sample from (X_i, Y_i) distribution, for N people, we can write

$$m_X = \frac{\sum_{i=1}^{N} X_i}{N},\ \sigma_X^2 = \frac{\sum_{i=1}^{N}(X_i - m_X)^2}{N}, \tag{1}$$
$$\rho(X,Y) = \frac{\sum_{i=1}^{N}(X_i - m_X)(Y_i - m_Y)}{N\sigma_X\sigma_Y}.$$

Setting $N = 3$ and using our income and age vectors, this yields $\rho \approx 0.87$. This positive value indeed makes sense because only person making <50K is the youngest one. Correlation between workclass and income may look less clear as there are many workclasses. However, we can create 0/1 vectors from individual workclasses. For instance, to study Private, we end up with $Z = [1\ 0\ 0]$ vector since only the first person has Private attribute. In this case, we find a correlation coefficient of $\rho(X, Z) = 0.5$.

Assignment

Based on the methodology described above, your tasks are as follows.

1. Plot the histogram of continuous features (age, capital-gain, hours-per-week).
2. Calculate the correlation coefficient between income and each of the continuous features.
3. Calculate the correlation coefficient between income and each value of each discrete feature. Discrete features are workclass, education, marital-status, relationship, and sex. For instance, for workclass, you need to go over the values "Private, Self-emp-not-inc, Self-emp-inc, ..." and look at the coefficient between income and 0/1 vector obtained by that workclass value as discussed above.
4. For each discrete feature, create a bar plot where x-axis is the feature values (e.g. "Private, Self-emp-not-inc, Self-emp-inc, ...") and y-axis is the correlation coefficients obtained in the previous question.

You are expected to return your code as well.

Remark: It is OK to find codes that read/format the dataset online or from your classmates. You are expected to do the rest yourself including calculation of correlation coefficient as described in (1).