EE 215 - Coding Assignment II

This exercise still focuses on the Adult dataset and you should be able to use some of your previous code. The goal of this assignment is playing with random vectors and introduction to linear prediction. 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.
- sex: Female, Male.
- capital-gain: continuous.
- hours-per-week: continuous.

As a first step, convert the income and sex features into 0/1 variables by assigning 0/1 to income≤50K/ income>50K and to sex=Male/sex=Female. With this, we can create a feature vector $\boldsymbol{x} \in \mathbb{R}^5$ where its five entries are income, age, sex, capital-gain, hours-per-week respectively.

We can generate a random vector $\boldsymbol{X} \in \mathbb{R}^5$ by picking one person at random and assigning $\boldsymbol{X}$ to be his/her feature vector. Let $\boldsymbol{x}_i$ be the feature vector of the ith person for $1 \le i \le N$ where N is the dataset size. Then, the (empirical) covariance, correlation, and mean of $\boldsymbol{X}$ is given by

$$\text{cov}(\boldsymbol{X}) = \text{corr}(\boldsymbol{X}) - \mu\mu^T \quad , \quad \text{corr}(\boldsymbol{X}) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T \quad , \qquad \mu = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i. \tag{1}$$

Our goal is predicting one of the features in terms of the others in the least square sense. Denote the 4 dimensional vector obtained by the other entries by $\bar{\boldsymbol{X}}$. Linear prediction is done by picking a vector $\boldsymbol{a}$ and defining the function $\hat{\boldsymbol{X}}_i = \boldsymbol{a}^T \bar{\boldsymbol{X}}$. Our goal is finding the vector $\boldsymbol{a}$ such that prediction error $\mathbb{E}[(\boldsymbol{X}_i - \hat{\boldsymbol{X}}_i)^2] = \mathbb{E}[(\boldsymbol{X}_i - \boldsymbol{a}^T \bar{\boldsymbol{X}})^2]$ is minimized. Differentiating over $\boldsymbol{a}$, the optimal value $\boldsymbol{a}_\star$ is given by

$$\mathbb{E}[\bar{\boldsymbol{X}} \boldsymbol{X}_i] = \mathbb{E}[\bar{\boldsymbol{X}} \bar{\boldsymbol{X}}^T] \boldsymbol{a}_\star \implies \boldsymbol{a}_\star = \mathbb{E}[\bar{\boldsymbol{X}} \bar{\boldsymbol{X}}^T]^{-1} \mathbb{E}[\bar{\boldsymbol{X}} \boldsymbol{X}_i]. \tag{2}$$

Note that the matrices $\mathbb{E}[\bar{\boldsymbol{X}} \bar{\boldsymbol{X}}^T]$ and $\mathbb{E}[\bar{\boldsymbol{X}} \boldsymbol{X}_i]$ are submatrices of the correlation matrix $\text{corr}(\boldsymbol{X})$. Hence, you can find $\boldsymbol{a}_\star$ from the correlation matrix. Your assignment is described next.

Assignment

1. Find and print the covariance, correlation, correlation coefficient, and mean of the vector $\boldsymbol{X}$.
2. For each i, calculate the $\hat{\boldsymbol{X}}_i$ variable. Plot histogram of $\boldsymbol{X}_i, \hat{\boldsymbol{X}}_i$ on the same figure. Histograms should use the same range on the x-axis and same number of bins. For example python code try running[1]. Are the histograms similar?
3. Let us see how well we did. A good measure is printing the relative error $\mathbb{E}[(\boldsymbol{X}_i - \hat{\boldsymbol{X}}_i)^2] / \mathbb{E}[\boldsymbol{X}_i^2]$. This should better be less than 1!
4. So far, we worked with the $\boldsymbol{X}$ vector. Repeat steps 2 and 3 with the de-biased vector $\boldsymbol{X}_{\text{new}} = \boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]$. Feel free to comment on any difference on the relative error (when using $\boldsymbol{X}_{\text{new}}$ rather than $\boldsymbol{X}$).

[1] plt.hist(np.random.randn(1000),bins=50,range=(-3,3),histtype='step')
plt.hist(np.random.randn(1000)/2,bins=50,range=(-3,3),histtype='step')

5. Steps 2,3,4 should be done for all $1 \leq i \leq 5$. Report the feature with the most predictability, i.e. smallest relative error (for both $\boldsymbol{X}$ and $\boldsymbol{X}_{\text{new}}$ if different).

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 e.g. writing the script for (1) and (2).