COM S 474/574: Introduction to Machine Learning
Homework #3

1. Please put required code files and report into a compressed file "HW#_FirstName_LastName.zip"
2. Unlimited number of submissions are allowed on Canvas and the latest one will be graded.
3. No later submission is accepted.
4. Please read and follow submission instructions. No exception will be made to accommodate incorrectly submitted files/reports.
5. All students are required to typeset their reports using latex. Overleaf (`https://www.overleaf.com/learn/latex/Tutorials`) can be a good start.

---

1. (30 points) You are provided with a training set of examples (see Figure 1). Which feature will you pick first to split the data as per the ID3 decision tree learning algorithm? Show all your work: compute the information gain for all the four attributes and pick the best one.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Figure 1: Table with training examples. Each row corresponds to a single training example. There are four features, namely, outlook, temperature, humidity, and wind. "PlayTennis" is the class label.

2. (20 points) **Principal Components Analysis**

   Given three data points: $(-1,-1),(0,0),(1,1)$.

   (a) (10 points) Show the first Principal Component (actual vector) without using Eigendecomposition. Justify your answer.

   (b) (10 points) If use the 1^{st} principle component to transform the data into 1-d space. What are the new data?

3. (50 points) **Principal Component Analysis:**

   In this homework, you will apply the principal component analysis to a collection of handwritten digit images from the USPS dataset. The USPS dataset is in the "data" folder: USPS.mat. The

starting code is in the "code" folder. The whole data has already been loaded into the matrix A. The matrix A has shape 3000×256 and contains all the images. Each row in A corresponds to a handwritten digit image (between 0 and 9) with size 16×16. You are expected to implement your solution based on the given codes. The only file you need to modify is the "solution.py" file. You can test your solution by running the "main.py" file.

(a) (20 points) In PCA, we obtain a projection matrix or reduce matrix $\boldsymbol{U} \in \mathbb{R}^{d \times p}$. Based on $\boldsymbol{U}$, we project the original centered data $\bar{\boldsymbol{X}} \in \mathbb{R}^{d \times n}$ into reduced data $\boldsymbol{Z} \in \mathbb{R}^{p \times n}$. Complete the ***_do_pca()*** method. You only need to center the data instead of applying mean normalization. Your code will be tested on $p = 10$, 50, 100, 200, total four different numbers of the principal components.

(b) (10 points) Based on the projection matrix $\boldsymbol{U}$ and reduce data $\boldsymbol{Z}$, we can reconstruct the original data $\boldsymbol{X'}$ by $\boldsymbol{UZ}$ and adding back the original means. Here you need to Complete the ***reconstruction()*** method to reconstruct the reduced data.

(c) (10 points) Based on the reconstructed data $\bar{\boldsymbol{X}}'$, we can compute measure the reconstruction error by $||\boldsymbol{X} - \boldsymbol{X'}||_F^2$. Complete the ***reconstruct_error()*** function to measuring the reconstruction error.

(d) (10 points) Run "main.py" to see the reconstruction results and summarize your observations from the results into a short report. When you run the "main.py" file, a subset (the first two) of the reconstructed images based on p = 10, 50, 100, 200 principal components will be automatically saved on the "code" folder. Please attach these images into your report also.

Note: You are NOT supposed to use existing PCA libraries; instead, you should write your own PCA. Please read the "Readme.txt" file carefully before you start this assignment.