Assignment 1, Foundations: 50 points

Due September 6, 2024, 11:59 PM CST

Instructions

Your homework submission must cite any references used (including articles, books, code, websites, and personal communications). All solutions must be written in your own words, and you must program the algorithms yourself. Submit your solutions as a PDF to Canvas: `https://canvas.tamu.edu/`.

Your programs must be written in Python. The relevant code to the problem should be in the PDF you turn in or submitted as a separate file (*e.g.*, Python file (.py), Jupter/Google Colab Notebook (.ipynb)). If a problem involves programming, then the code should be shown as part of the solution to that problem. If you solve any problems by hand just digitize that page and submit it (make sure the problem is clearly labeled and legible).

If you have any questions, please reach out to Dr. Peeples.

1 Linear Algebra- 10 points

For each of the following problems, state whether or not the the operation is defined (i.e., valid and can be computed) and, if it is defined, what is the size of the resulting answer. For all of the following problems let $\mathbf{X}$ be a $M \times N$ matrix, $\mathbf{Y}$ be a $N \times N$ matrix, $\mathbf{a}$ be a M × 1 vector, $\mathbf{b}$ be a $N \times 1$ vector and s be a scalar.

1. $\mathbf{XY}$
2. $\mathbf{YX}$
3. $\mathbf{YX^T}$
4. $\mathbf{aX}$
5. $\mathbf{a^TX}$
6. $\mathbf{aX^T}$
7. $\mathbf{a^Tb}$
8. $\mathbf{b^Tb}$

9. $\mathbf{b}\mathbf{b}^{\mathbf{T}}$

10. $s\mathbf{X}+\mathbf{Y}$

2 Probability- 5 points

The following frequency table is generated from an experiment of rolling a fair dice. You can do this problem by hand or create a python script.

	1	2	3	4	5	6
Counts	200	100	300	50	150	200

1. Compute and plot the empirical probability mass function.

2. Compute and plot the empirical cumulative distribution.

3 Statistics- 10 points

Given two datasets, $\mathbf{A} : [4, 16, 4, 5, 1]^T$ and $\mathbf{B} : [8, 3, 4, 8, 7]^T$, compute the following statistical (**sample**) measures by hand for each dataset:

1. Mean

2. Median

3. Range

4. Variance

5. For which dataset ($\mathbf{A}$ or $\mathbf{B}$), would a new value of 18 be more likely to be an outlier and why?

4 Data Attributes- 10 points

4.1 Numerical

Given the following data matrix, $\mathbf{D} = \begin{pmatrix} 0.1 & 1.6 & 2.7 & 3.2 & 4.1 & 5.9 \\ 0.3 & -0.4 & -1 & -1.6 & -2.2 & -2.8 \end{pmatrix}^T$, transform and plot the new data using (can either compute and plot by hand or creating python script):

1. Min-max normalization

2. Standard score normalization (*i.e.*, standardization)

4.2 Categorical

Please access the Iris Flower Dataset via Sklearn and select the following two attributes: petal length and petal width.

1. Discretize the petal length into four groups (very short, short, long, very long) and petal width into three groups (short, medium, and long) using equal-width intervals.

2. Create contingency table for the two attributes.

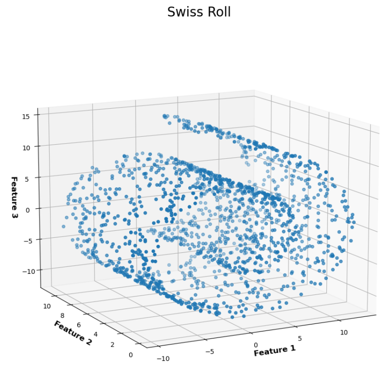
3. Are these two features independent? Why or why not? (hint: use Chi-squared test with $\alpha = 0.01$)

5 Dimensionality Reduction- 15 points

For this question, please download all three datasets provided along with this assignment: *swissroll.txt*, *spheres.txt*, and *ellipsoid.txt*. Import these files into your programming software. All datasets have 1500 samples and 3 features/dimensions. Now, consider **D** to be a dataset (*i.e.*, size 1500 × 3 or $D \in \mathbb{R}^{1500\times3}$). For example, in Python, you can use a script to generate a plot of the dataset **D** :

```
import matplotlib.pyplot as plt
import numpy as np

D = np.loadtxt(“swissroll.txt”)
fig = plt.figure()
ax = plt.axes(projection=‘3d’)
ax.scatter3D(D[:,0], D[:,1], D[:,2])
ax.set_xlabel(‘Feature 1’, fontsize=12, =dict(weight=‘bold’))
ax.set_ylabel(‘Feature 2’, fontsize=12, fontdict=dict(weight=‘bold’))
ax.set_zlabel(‘Feature 3’, fontsize=12,fontdict=dict(weight=‘bold’))
ax.set_title(‘Swiss Roll’, fontsize = 20)
```



For each dataset:

1. Find the covariance matrix.
2. Find the eigenvectors and eigenvalues of the covariance matrix.
3. Find (and plot) the projection of the data points into the 2-D and 1-D principal components (hint: use PCA and **do not** normalize data before PCA). After projecting the data into 2-D and 1-D, provide a short discussion (2-3 sentences) of the results for each dataset that answers the following question: Does the projection preserve the "important" or "most informative" structure for the original data? Why or why not? (hint: analyze quantitative and qualitative observations)