1. **Designing Simple Pattern Analysis Systems**
   A Pattern Analysis System is a system responsible for automated recognition of patterns in data, in order to identify which of a set of categories (or classes) a new observation belongs to, or to estimate the value of a specific attribute. It contains several parts, which must be carefully designed.
   In this problem, you will get hands-on experience in designing a pattern analysis system for various different scenarios. In each one of the scenarios, you need to answer to some questions.

   (a) Predicting your final grade in this course

   Solution:
   1. Which types of prediction problems (classification, regression, etc.) does it belong to?
   Regression
   2. What sensors (if any) are needed?
   Nothing
   3. What is your training set?
   Since I have passed SML, list of SML and Pattern Recognition Grades for last semester's Pattern Recognition students who also have passed SML would be a good dataset.
   4. How do you gather your data?
   By asking from office or getting from Moodle.
   5. Which Feature do you select?
   Vector of SML and Pattern Recognition Grades (Grades for exam, home works, etc)
   6. Is there any pre-processing stage needed?
   We have to select students who passed both classes and remove the rest. This stage can be done in the preprocessing stage.
   7. Express the challenges and difficulties that may affect the outcome of your system.
   There might be some change in the evaluation of Statistical Pattern Recognition in this term comparing to the previous term. So in this case, the prediction would not work well even with the best model!
   8. How beneficial do you think it is to design such a system? Express the pros and cons of applying these systems instead of using a human observer.
   Pros: The system will be more precise, especially when have a big dataset.
   Cons: A human can identify the strength of a student by looking at the features roughly even if there is some noise in the student's grades. Using this knowledge, a human can predict a rough estimation. But the system can wrongly identify correlations especially when there is noise in the dataset and the dataset is not big enough.

   (b) Labeling a collection of books by their genres

   Solution: 1. Which types of prediction problems (classification, regression, etc.) does it belong to?
   Classification
   2. What sensors (if any) are needed?
   Nothing
   3. What is your training set?
   For each book, we have a set of (Name, author, number of pages, publisher) as features
   4. How do you gather your data?
   By getting information from a library
   5. Which Feature do you select?
   I use number of page and publisher as two basic features. I will also use some features related to the name of the book
   6. Is there any pre-processing stage needed?
   To get features related to book's name, we need to have some NLP preprocessing steps
   7. Express the challenges and difficulties that may affect the outcome of your system.
   Some books can be classified as more than one genre. As an example, suppose we have a fantasy story which has also some kind of mysteries. This book can be classified as "FANTASY". It can also be classified as "MYSTERY". In the dataset, we just have one label for each book. Suppose we have multiple books which have the "FANTASY" and the "MYS-

TERY" part. These books can be labeled differently. Using this dataset, the system can not learn the relationship between different features and labels. (i.e. these books have the toughly same features but different labels) As a result, this system can not predict well specifically on these kinds of books.
8. How beneficial do you think it is to design such a system? Express the pros and cons of applying these systems instead of using a human observer.
Pros: This system can help libraries to categorize tons of books really fast
Cons: Sometimes, this system can not predict well like a human.

(c) Predicting the winner of 2018-19 UEFA Champions League

Solution:
1. Which types of prediction problems (classification, regression, etc.) does it belong to?
Classification
2. What sensors (if any) are needed?
Nothing
3. What is your training set?
To predict the winner of the league, we have to use data to understand the strength and weakness of each team in each match to predict the winner of that game. To do so, we have to have two sets of features. We call the first set, Global features. This set is all about global features related to the strength and weakness of each team in general. (i.e number of goals in one match) The second set is pairwise features related to each pair of teams. It actually means how strong team A is when it faces team B. (i.e. number of goals in match between A and B) The result of each match can be predicted well according to the global features and the pairwise features.
4. How do you gather your data?
By using previous years records
5. Which Feature do you select?
I will use a set of features that contain both Global Features and Pairwise Features
6. Is there any pre-processing stage needed?
In the case of an incomplete dataset, we need to fill the n/a with reasonable value
7. Express the challenges and difficulties that may affect the outcome of your system.
Actually, this classification task has a great amount of noise! It is because every match has a lot of different factors (such as mental factors) that can not be reported in the dataset. So we should not expect to have a very great accuracy
8. How beneficial do you think it is to design such a system? Express the pros and cons of applying these systems instead of using a human observer.
Pros: The system can combine different features in the prediction
Cons: The expert can choose the most important features in the prediction. He can also consider factors which are not specified in the dataset such as emotional conditions

(d) Grouping students in a dorm by their personalities

Solution:
1. Which types of prediction problems (classification, regression, etc.) does it belong to?
Clustering. Each group forms a cluster. We want our clusters to have some desirable attributes.
2. What sensors (if any) are needed?

3. What is your training set?
Our training set is set of all students' data. This dataset does not have any label for students.
4. How do you gather your data?
We need to ask students to fill a survey to gather their preferences and personalities
5. Which Feature do you select?

Those features that significantly separate one cluster's data from the others
6. Is there any pre-processing stage needed?
No
7. Express the challenges and difficulties that may affect the outcome of your system.
This problem is clustering with constraints. (i.e. having limit on the number of student in each dorm room) Therefore it is harder than normal clustering.
8. How beneficial do you think it is to design such a system? Express the pros and cons of applying these systems instead of using a human observer.
Pros: This system can match students who are not familiar with each other (but they have the same personalities)
Cons: Some people are friends to each other and they are good choices to be grouped in the same group. But the data that we collect will not necessarily give a high score to this group

(e) Predicting the price of meat in the coming month

Solution:
1. Which types of prediction problems (classification, regression, etc.) does it belong to?
Regression. (Time series)
2. What sensors (if any) are needed?
Nothing
3. What is your training set?
Price of meat in the previous months, the price of other stuff such as Dolar and Oil and chicken
4. How do you gather your data?
From related association for meat, etc.
5. Which Feature do you select?
Features that can express all important factors for changing the price
6. Is there any pre-processing stage needed?
There might be some missing data that we should fill with proper values or remove them from the dataset.
7. Express the challenges and difficulties that may affect the outcome of your system.
The price of meat can be changed in a directorial way. (Someone set a price for meat and force all sellers to sell the meat at the specific price). So the system can not learn the correlation between different prices properly.
8. How beneficial do you think it is to design such a system? Express the pros and cons of applying these systems instead of using a human observer.
Pros: System can manipulate different factors for the prediction
Cons: Some part of knowledge about the market is not reflected in the data set and an expert can recognize them better.

2. **Getting More Familiar with the Art of Feature Extraction**
The success of a pattern recognition system is heavily dependent on the Feature Extraction stage, where the goal is to extract distinctive properties of input patterns that best help in differentiating between the categories of the input data.
In this problem, you are going to get more familiar with the importance of feature extraction stage. Here, the focus is mainly on classification problems.
First, assume a simple Facial Recognition problem. Please state what features might be used to best distinguish among the following sets.

(a) African and Non-African
Face Color, Hair Color

(b) Happy and Neutral
Mouth direction, Mouth Size, Teeth

(c) Young and Adult
The edges in the face (Children have smooth face), beard

(d) Male and Female
beard, hair length, hair style, face wrinkles,

(e) Facial Recognition
Hair color, eyebrow thickness, skin color, face proportion and etc. For more information visit check out this paper.

In the second scenario, consider an Optical Character Recognition (OCR) problem, in which your task is to extract meaningful features which are invariant to the following changes, when the goal is to detect a letter 'A' in a desired text.

These features are invariant to affine transformation:

- Number of loops (O has 1 loop and 8 has two loops)
- Number of Crossing Points (H has 2 crossing points)
- Number of End points (H has 4 end points)

These features are also useful but they are not completely invariant to affine transformation (They are sensitive to rotation, shear, and etc):

- Number of Horizontal Lines (H has 1 horizontal line)
- Number of Vertical Lines (H has 2 vertical line)
- Number of Diagonal Lines (A has two diagonal line, one with positive slope and one with negative slope)

3. **Feature Selection: Evaluating Features to Select 'Good' Ones**
The answer to this question can be found in different files:
For Code + Report you can see Question3.html
For just report you can check Report_version_Question3.html
To run codes, you can check p3.ipynb. (p3.py is also available) You asked our code to be separated for each section. It is separated properly in the jupyter notebook sections. (If I had separeted them in different files, it would be less readable.)

4. **Basic Statistics Warm-up**
In Statistical Pattern Recognition, the goal is to use Statistical Techniques for analysing data measurements in order to extract meaningful information and make justified decisions. Therefore, mastering basic statistical properties and to be able to understand and use them is highly important.
In this problem, you are to review your knowledge in this area. First, find the following quantities for a random variable X with the probability density function:

$$f(x) = \begin{cases} cx & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

(a)

$$\int_0^1 cx dx = [cx^2/2]_0^1 = c/2 = 1 \rightarrow c = 2$$

(b)

$$P(0 \leq X \leq 0.5) = \int_0^{0.5} 2x dx = [x^2]_0^{0.5} = 1/4$$

(c)

$$E(X) = \int_{-\infty}^{\infty} f(x) x dx = \int_0^1 2x^2 dx = [\frac{2}{3}x^3]_0^1 = \frac{2}{3}$$

(d)

$$Var(X) = E(X^2) - E(X)^2 = \int_{-\infty}^{\infty} f(x)x^2dx - \frac{2}{3}^2 = \int_0^1 2x^3dx - \frac{4}{9} = [\frac{1}{2}x^4]_0^1 - \frac{4}{9} = \frac{1}{18}$$

(e)

$$E(2X-2) = 2E(X) - 2 = -\frac{2}{3}$$

(f)

$$Var(2X-2) = Var(2X) = 4Var(X) = \frac{4}{18}$$

Now suppose a normal random variable X with parameters $\mu = 1$ and $\sigma^2 = 9$

(g)

$$P(-2 \leq X \leq 1) = P((-2-1)/3 \leq Z \leq (1-1)/3) = P(-\frac{1}{1} \leq Z \leq 0) = F_z(0) - F_z(-\frac{1}{1}) = 0.5 - 0.16 = 0.34$$

(h)

$$E(X) \text{ and } Var(X)$$

We define Z as follows:

$$Z = \frac{x-1}{3} \rightarrow x = 3z + 1$$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} (3z+1)f(x(z))3dz = \int_{-\infty}^{\infty} (3z+1)\frac{1}{\sqrt{2\pi} * 3}\exp(-\frac{(3z)^2}{2*3^2})3dz =$$

$$\int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} (3z+1)f(x(z))3dz = \int_{-\infty}^{\infty} (3z+1)\frac{1}{\sqrt{2\pi}}\exp(-\frac{(z)^2}{2})dz =$$

$$\int_{-\infty}^{\infty} 3z\frac{1}{\sqrt{2\pi}}\exp(-\frac{(z)^2}{2})dz + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp(-\frac{(z)^2}{2})dz =$$

$$\frac{3}{\sqrt{2\pi}}[\exp(-\frac{z^2}{2})]_{\infty}^{-\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp(-\frac{(z)^2}{2})dz =$$

$$0 - 0 + F_z(\infty) = 1$$

$$Var(X) = E(X^2) - (E(X))^2$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2f(x)dx = \int_{-\infty}^{\infty} (3z+1)^2f(x)3dz =$$

$$\int_{-\infty}^{\infty} (9z^2+6z+1)\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})dz = \int_{-\infty}^{\infty} 9z^2\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})dz + \int_{-\infty}^{\infty} (6z+1)\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})dz$$

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} z^2\exp(-\frac{z^2}{2})dz = (\frac{1}{\sqrt{2\pi}})([-z\exp(-\frac{z^2}{2})]_{-\infty}^{\infty} + \int_{-\infty}^{\infty}\exp(-\frac{z^2}{2})) = 0 - 0 + 1 = 1$$

$$\rightarrow \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} 9z^2\exp(-\frac{z^2}{2})dz = 9$$

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} (6z+1)\exp(-\frac{z^2}{2})dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} (6z)\exp(-\frac{z^2}{2})dz + F_z(\infty) = 0 + 1$$

$$\rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (6z+1) \exp(-\frac{z^2}{2}) dz = 1$$

$$E(X^2) = 9 + 1$$

$$\rightarrow Var(X) = E(X^2) - (E(X))^2 = 10 - 1 = 9$$

(i)

$$F_y(y) = P(Y \leq y) = P(2X - 1 < y) = F_X(\frac{y+1}{2}) = \int_{-\infty}^{\frac{y+1}{2}} f_X(u) du$$

$$t = 2u - 1, dt = 2du \rightarrow \int_{-\infty}^{\frac{y+1}{2}} f_X(u) du = \int_{-\infty}^{y} f_X(\frac{t+1}{2}) \frac{dt}{2} =$$

$$\int_{-\infty}^{y} \frac{1}{2\sigma\sqrt{2\pi}} \exp(-\frac{t-(2\mu-1)}{2(2\sigma)^2}) dt = F_{Z \sim Normal(2\mu-1, 4\sigma^2)}$$

(j) Suppose we have n students.

$$P(Event) = \binom{n}{4} (\frac{1}{5})^{n-4} (\frac{4}{5})^4$$

(k)

$$\int_{-\infty}^{t} f(x) dx = 0.5 \rightarrow \int_{-\infty}^{t} f(x) dx = \int_{0}^{t} \frac{1}{4}(4 - x^2) dx = [x - \frac{x^3}{12}]_0^t = 0.5 \rightarrow t^3 - 12t + 6 = 0 \rightarrow t \approx 0.512$$

(l)

$$P(X \leq 9.8) = P(Z = (X - 10)/0.1 \leq -2) = 0.02275013$$

(m)

$$E(X) = \int_0^1 \frac{4x}{\pi(1+x^2)} dx = [\frac{2}{\pi} \ln(1+x^2)]_0^1 = \frac{2}{\pi} \ln(2)$$

5. **Hanging Around with Covariance Matrix and Linear Transformations**

In statistical pattern recognition, Covariance Matrix concept is highly important. In general, it is defined as a matrix whose element in position i, j is the covariance between i th and j th elements of a random vector. It somehow generalises the concept of variance to multiple dimension. In this problem, you are going to examine your knowledge of covariance matrices and their attributes.

(a) Specify the dimensionality of the dataset, i.e. the number of features each sample has.
number of dimensions = 3 because of the dimension of Covariance Matrix

(b) Determine the number of samples in the dataset.
It can not be determined just by covariance matrix.

(c) Find the correlations between different data dimensions.

$$correlation(x, y) = \rho_{x,y} = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

$$\Sigma = \begin{bmatrix} 2 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 2 \end{bmatrix} \rightarrow \rho_{1,1} = \sigma_1 = \sqrt{2}, \rho_{2,2} = \sigma_2 = 1, \rho_{3,3} = \sigma_3 = \sqrt{2}$$

$$\rho_{1,2} = \rho_{2,1} = 0, \rho_{1,3} = \rho_{3,1} = \frac{-1}{2}, \rho_{2,3} = \rho_{3,2} = 0$$

(d) On which dimension are the data scattered more?
feature 1 and feature 3 have bigger variance so in those direction, the data are scattered more

(e) Calculate eigenvalues and eigenvectors associated with the covariance matrix, and then find the angle between each of the eigenvector pairs. What can you infer from the three obtained values? Does it hold in every arbitrary covariance matrix? Justify your answer.
First we need to calculate eigenvalues.

$$|\Sigma - \lambda I| = \begin{vmatrix} 2 - \lambda & 0 & -1 \\ 0 & 1 - \lambda & 0 \\ -1 & 0 & 2 - \lambda \end{vmatrix} = -(\lambda - 3)(\lambda - 1)^2 = 0$$

$$\lambda_1 = 3 \rightarrow v_1 = x_1 \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

$$\lambda_2 = \lambda_3 = 1 \rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow v = x_1 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\rightarrow v_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, v_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Using dot product, we can see easilily that these tree vectors are perpendicular to each other. Also, for this specific matrix, because we have $\lambda_2 = \lambda_3 = 1$, we have an eigenspace corresponding to this eigenvalue.
It is true for all covariance matrix that their eigenvectors are perpendicular. This is true because of the following theorem in Linear Algebra:
Every symmetric and real-valued matrix has orthogonal eigenvectors.
Since every covariance matrix, by definition, is symmetric and real-valued, its corresponding eigenvectors are orthogonal.

(f) Find a transformation to whiten data associated with the given covariance matrix. First, we need to normalize all eigenvectors:

$$A = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Then, the whiten transformation is:

$$WT = \begin{bmatrix} -\frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sqrt{6}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

(g) Show that Σ is a valid covariance matrix.

> **Theorem 1.** *A square matrix is a covariance matrix of some random vector if and only if it is symmetric and positive semi-definite.*

So we need to show that Σ is symmetric and positive semi-definite. It is obvious that Σ is symmetric. We will use theorem 2 to show that it is also positive semi-definite.

> **Theorem 2.** *The Hermitian matrix M is positive semi-definite if and only if all of its eigenvalues are non-negative.*

We know that every symmetric real-values matrix is Hermitian Matrix. Since Σ's eigenvalues are positive, using theorem 2, we can conclude Σ is positive semi definite hence a covariance matrix.

6. **Simple Sample Generation and Beyond**
The answer to this question can be found in different files:
For Code + Report you can see Question6.html
For just report you can check Report_version_Question6.html
To run codes, you can check p6.ipynb. (p6.py is also available) You asked our code to be separated for each section. It is separated properly in the jupyter notebook sections. (You can run each section independently) (If I had separeted them in different files, it would be less readable.)

7. **Some Explanatory Questions**

   (a) Why do you think Central Limit Theorem is important? Where and how can it be used?

   When we have enough random variables from a independent and identically distributed from a specific distribution, their average tends to have a normal distribution. So we can treat to the sum of these random variables as a value drawn from a **normal distribution**.
   We can also use this theorem to calculate **confidence interval** for the mean estimation.
   Another use of Central limit theorem is that we can think of **noise** as a normal distribution. (Because noise consists of a sum of independent factors).
   (I have written a post about central limit theorem and I will be glad to take a look! Why mean squared error)
   After all, the central limit theorem is one of the most important things in statistics! Without it, we can not estimate the average of any random variable with good confidence.

   (b) What is the difference between a feature and a measurement?
   Suppose we have a data such as (name: Ali, age: 24, height:180) Age is a feature. 24 is a measurement for that feature.

   (c) Does a covariance matrix need to be symmetric? Why?
   Yes. Because of its definition and the fact that $E(XY) = E(YX)$.

   (d) What does zero eigenvalue mean?
   It means that the corresponding transformation maps a line (all vectors parallel to eigenvector) to point zero. This means that the transformation matrix is not full rank and maps space to a lower dimensional space.

   (e) When does the whitening transformation come into use?
   When we want our features to have the same scale and to be de-correlated.
   But sometimes, we should not use whitening transformation and instead we should just de-correlate our data by rotation. This situation occurs when we have a noise in our features. By whitening transormation, we make some noise bigger which is not really good.

   (f) Explain how does Google's PageRank algorithm use eigenvalues and eigenvectors concepts.
   In fact, page rank algorithm can bee seen as finding a stationary distribtuion in a graph of the internet whose nodes are webpages plus one extra node. (We have to add an extra node to simulated the functionality of $\mathbf{d}$ in the page rank algorithm $p(state = i, time = t) = d * \sum transition(j, i) * p(j, t-1) + \frac{1-d}{N}$)
   Finding the stationary distribution is equal to finding ranks for all pages. The stationary distribution has the following property:
   $[A^t][\pi] = 1.[\pi]$ So π is equal to the eigenvector corresponding to eigenvalue of 1.

   Also we can use eigenvalue decomposition to compute the n state transition since:

$$A^n = v[diag(eigenvalues)^n]v^{-1}$$