Amirkabir University of Technology
(Tehran Polytechnic)

Assignment 1
A Short Warm-Up Before the Real Deal!

Homeworks Guidelines and Policies

- ***What you must hand in.*** It is expected that the students submit an assignment report (HW1_[student_id].pdf) as well as required source codes (.m or .py) into an archive file (HW1_[student_id].zip).
- ***Pay attention to problem types.*** Some problems are required to be solved *by hand* (shown by the icon), and some need to be implemented (shown by the icon).
  Please do not use implementation tools when it is asked to solve the problem by hand, otherwise you will be penalized and lose some points.
- ***Don't bother typing!*** You are free to solve by-hand problems on a paper and include their pictures in your report. Here, cleanness and readability are of high importance. Images should also have appropriate quality.
- ***Reports are critical.*** Your work will be evaluated mostly by the quality of your report. Do not forget to explain your answers clearly, and provide enough discussions when needed.
- ***Appearance matters!*** In each homework, 5 points (out of a possible 100) belong to compactness, expressiveness, and neatness of your report and codes.
- ***Python is also allowable.*** By default, we assume you implement your codes in MATLAB. If you are using Python, you have to use the equivalent functions when it is asked to use specific MATLAB functions.
- ***Be neat and tidy!*** Your codes must be separated for each question, and for each part. For example, you have to create a separate .m file for part b. of question 3, which must be named 'p3b.m'.
- ***Use bonus points to improve your score.*** Problems with bonus points are marked by the icon. These problems usually include uncovered related topics, or those that are only mentioned briefly in the class.
- ***Moodle access is essential.*** Make sure you have access to Moodle, because that is where all assignments as well as course announcements are posted. Homework submissions are also made through Moodle.

- ***Assignment Deadline.*** Please submit your work **before the end of November 1st**.
- ***Delay policy.*** During the semester, students are given 7 free late days which they can use them in their own ways. Afterwards, there will be a 25% penalty for every late day, and no more than three late days will be accepted.
- ***Collaboration policy.*** We encourage students to work together, share their findings, and utilize all the resources available. However you are not allowed to share codes/answers or use works from the past semesters. Violators will receive a zero for that particular problem.
- ***Any questions?*** If there is any question, please do not hesitate to contact us through the following email addresses: **ali.the.special@gmail.com** and **ebp.mohsen@gmail.com**.

1. Can Artificial Intelligence Expose Their Lies? (12 Pts.)

Keywords: *Pattern Recognition System, Feature Extraction, Prediction Problems, Classification, Regression, Clustering, Fake News Detection*

These days we are surrounded by rumors from various sources. Many of these stories are supported by a group, yet denied by another. From the death of Mahsa Amini to Evin Prison fire, there have been lots of debates over the claims made by different parties. Under such chaotic circumstances, can we expect AI to help us evaluate the authenticity of the news?

Figure 1 News reports are full of surprises these days!

You are asked to suggest a practical and feasible **Pattern Analysis System** capable of detecting fake news. This system is expected to assign a 'reliability score' to each piece of news it receives. Please specify the following:

a. Which types of prediction problems (classification, regression, etc.) does it belong to?
b. What are the inputs?
c. What sensors (if any) are needed?
d. What is your training set?
e. How do you gather your data?
f. Which features do you select?
g. Is there any pre-processing stage needed? Explain.
h. Express the challenges and difficulties that may affect the outcome of your system.
i. How beneficial do you think it is to design such a system? Express the pros and cons of applying these systems instead of other news evaluation methods.

Note 1: There is no limitations on the method you choose. As an example, a system capable of grouping students by their personalities could be based on surveillance cameras (**Computer Vision** techniques), paper-based surveys (**Sentiment Analysis** techniques), etc.

Note 2: Your design must be as practical as possible. For instance, features must be discriminative and measurable.

2. Basic Statistics Warm-Up (I) (15 Pts.)

Keywords: *Probability Theory, Random Variable, Discrete Variable, Conditional Probability, Marginal Probability , Probability Distribution, Density Function, Continuous Variable, Cumulative Distribution Function, Independent Variables, Correlated Variables , Expected Value*

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, assume a fair 6-sided die with the number 1 on three sides, the number 2 on two sides, and the number 3 on one side. After rolling it 10 times,

a. Calculate the probability that we get five 1s, three 2s, and two 3s in no particular order.
b. Find the distribution of the number of 2s we could get. Include the expected value and variance.

c. In another experiment, the die is rolled two times. Let X be the sum of the two numbers that we rolled. Calculate the standard deviation of X.

Next, suppose we have a standard 3×3×3 Rubik's Cube, which breaks apart into 1×1×1 small cubes after being dropped to the floor.

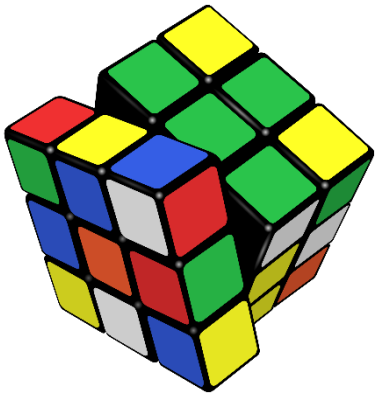

Figure 2 A standard 3×3×3 Rubik's Cube

d. Assume you pick up one of the cubes which has five unpainted faces. Calculate the probability that the remaining face (on the bottom) is also unpainted.
e. Two of the cubes are randomly selected. Find the probability that each one of them has exactly two sides painted.
f. Two of the cubes are randomly selected. What is the probability that the total number of painted sides is more than five?

Now, consider there is a bag of coins containing p pennies and d dimes. We draw the coins from the bag one at a time without replacement. Assume X is the number of draws until we get a dime.

g. Find $E[X]$ and $Var[X]$ in terms of p and d.

Next, assume a random variable X is 3 with probability 0.25, -3.5 with probability 0.35, 7.269 with probability 0.3, and 0.2 with probability 0.1.

h. Calculate $E[X]$ and $Var[X]$.
i. Plot the probability mass function of X.
j. Roughly plot the cumulative distribution function.
k. Calculate the mean and variance of $Y = 10X - 1$.

Finally, suppose at a COVID-19 vaccination center, on average, there are 7 people per hour who come to get vaccinated.

l. Calculate the probability that exactly 7 people come to get vaccinated in an hour.
m. We started to count people at 2:00 PM and 4 people have already come to get vaccinated so far. Calculate the probability that by 3:00 PM, there will have been exactly 7 people who have come to get vaccinated.

3. Basic Statistics Warm-Up (II) (20 Pts.)

Keywords: *Probability Theory, Random Variable, Discrete Variable, Conditional Probability, Marginal Probability , Probability Distribution, Density Function, Continuous Variable, Cumulative Distribution Function, Independent Variables, Correlated Variables , Expected Value*

Following the previous problem, we are to deal with more complicated statistics problems, mainly **Multivariate Statistics**.

First, assume the random variable vector $X = (X_1, X_2, X_3, X_4)'$ is multivariate normal with the following mean and covariance matrix:

$$\mu = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \qquad \Sigma_X = \begin{bmatrix} 1 & 2 & 0 & 2 \\ 2 & 1 & 0 & 1 \\ 0 & 0 & 1 & -2 \\ 2 & 1 & -2 & 16 \end{bmatrix}$$

a. Calculate the distribution of X_1.
b. Determine the independent variables.
c. Calculate the conditional distribution of X_2 given $X_3 = 1$.
d. Calculate the distribution $(X_1, X_2)'$.
e. Calculate the distribution $(X_1, X_2)'$ given $X_4 = 1$.
f. Calculate the joint distribution of Y_1 and Y_2 if $Y_1 = X_1 - X_2 + X_4$ and $Y_2 = X_1 - X_2 - X_4$.
g. Calculate $P(X_1 - X_2 + X_4) > 0$.
h. Find $\sum_{i=1}^{4} Var(X_i)$.
i. Find $Var\left(\sum_{i=1}^{4} X_i\right)$.

Next, assume X and Y are independent and $N(0,1)$-distributed.

j. Find $E(X \mid X > Y)$.
k. Find $E(X + Y \mid X > Y)$.

Finally, assume the following distribution for random variables X and Y:

$$f_{X,Y}(x,y) = \begin{cases} \frac{K}{\pi} e^{-\frac{x^2+y^2}{2}} & xy \geq 0 \\ 0 & xy < 0 \end{cases}$$

l. Find the value of K.
m. Are X and Y each Gaussian random variables? Prove your answer.
n. Are X and Y jointly Gaussian? Prove your answer.
o. Check whether X and Y are independent or not.
p. Check whether X and Y are uncorrelated or not.
q. Calculate $f_{X|Y}(x \mid y)$.
r. Is the result you obtained in the previous part a Gaussian distribution? Prove your answer.

4. Simple Yet Complex: A Review on Eigenvalues and Eigenvectors (18+5 Pts.)

Keywords: *Eigenvalues, Eigenvectors, Eigenspace, Invertible Matrix, Diagonalizable Matrix, Hermitian Matrix*

According to the latest statistics, an average person makes three to four Google Searches per day. It means that at least three to four times a day, you make use of the concepts of **Eigenvalues** and **Eigenvectors** in your life, as they are the cornerstone of the well-known Google algorithm, namely *PageRank*. Eigenvalues and Eigenvectors are also vastly used in various scientific areas, from geology and ecology to computer vision and data mining. In this problem, we are going to take a deeper look into these concepts.

Figure 3 Google PageRank, the main algorithm of the Google Search Engine, is based on the concepts of eigenvalues and eigenvectors.

a. Calculate a non-singular matrix P, such that $P^{-1}AP$ is diagonal, given

$$A = \begin{bmatrix} 4 & -4 & 5 \\ 0 & 5 & -4 \\ 0 & -5 & 4 \end{bmatrix}$$

b. Compute a basis B for $\mathbb{R}^2$ such that the matrix of $T(x,y) = (4x + 4y, 4x + 4y)$ relative to B is diagonal.

c. Compute the eigenvalues of the matrix $\begin{bmatrix} 5 & 12 \\ 12 & 5 \end{bmatrix}$. Find the dimension of the corresponding eigenspace for each eigenvalue.

Given the matrix $B = \begin{bmatrix} -1 & 4 & -2 \\ -3 & 4 & 0 \\ -3 & 1 & 3 \end{bmatrix}$, find

d. The characteristic equation of B.
e. The eigenvalues and eigenvectors of B.
f. A basis for each eigenspace of B.
g. A matrix P, if possible, that diagonalizes B.
h. $P^{-1}BP$, the corresponding diagonal form of B.

Given the matrix $C = \begin{bmatrix} 10 & -9 & 0 & 0 \\ 4 & -2 & 0 & 0 \\ 0 & 0 & -2 & -7 \\ 0 & 0 & 1 & 2 \end{bmatrix}$, find

i. The characteristic equation of C.
j. The eigenvalues and eigenvectors of C.
k. A basis for each eigenspace of C.

For each of the following characteristic equations of some matrix D, specify:

i. $p(\lambda) = \lambda(\lambda-1)^2(\lambda+2)^2$ ii. $p(\lambda) = (\lambda+1)^3(\lambda+2)(\lambda+4)^5$

l. The size of D.
m. Is D invertible? If yes, determine the eigenvalues of D^{-1}.
n. The number of eigenspaces of D.

Given the matrix $E = \begin{bmatrix} 2 & 1-i \\ 1+i & 3 \end{bmatrix}$,

o. Show that E is Hermitian.
p. Find the eigenvalues of E, and verify that they are real.
q. Find the corresponding eigenvectors of each eigenvalues, and confirm they are orthogonal to each other.
r. Diagonalize E by a unitary eigenvector matrix. In other words, decompose $A = U\Lambda U^T$ with a diagonal matrix Λ and a unitary matrix U.

5. Give Me Your Playlist and I'll Tell You Your Next Favorite Song! (20+5 Pts.)

Keywords: *Feature Extraction, Linear Transformation, Dimensionality Reduction, Mean and Variance, Normal Distribution, Recommender System, Pearson Correlation Coefficient, Spider Chart*

If you are a fan of Spotify, you are probably familiar with its 'Discover Weekly' feature. Every Monday, Spotify releases a playlist of 30 songs personalized for each user based on their music taste. These list are often highly praised by the listeners for their accuracy, and over one-third of all new artist discoveries happen through the recommendations made by this service. But how can people's taste in music be assessed?

Figure 4 Each Monday, over 400 million Spotify users find a new exclusive playlist waiting for them. This playlist contains 30 songs selected based on their music taste.

You are provided with a dataset containing the top 600 most Iranian music tracks played in the year 1399. The goal is to analyse the data and design a simple music **Recommender System**.

a. Calculate the mean of each attribute, and discuss the results. Is it reasonable to conclude that these tracks tend to be *happy* based on 'energy', 'danceability', and 'tempo' features?
b. Calculate the variance of each attribute. Do the listeners have similar patterns in their music preference based on solely the 'liveliness' feature?
c. Assuming a normal distribution for each attribute, display the estimated distribution as well as the histogram associated with each feature, and determine which one actually fits.
d. Calculate the Pearson correlation coefficient for each pair of numerical features. Which features can be safely removed based on their correlation with the others? Remove these features from the dataset.
e. Scale all the numerical features to the same range, then repeat the previous parts. Does it change the previously calculated results? Justify your answer.
f. Assuming that all the features have the same importance to the listeners, design a recommender system and obtain the top five tracks. Explain your method.
g. One approach to recommending new tracks to the users is to multiply each sample by a weight vector and find the top recommendations based on the sorted results. Through this, people's specific preference can be taken into consideration. Implement this strategy and find a list of top tracks that have more danceability and energy with low liveness and acousticness. As expected, the danceability and energy of the recommended tracks have higher values than the mean values, whereas the reverse is the case with liveness and acousticness.
h. Use the algorithm in the previous part to recommend a music track to yourself based on your own music taste. Is the recommendation accurate? Explain the reasons.
i. Spider chart is a graphical method of displaying and comparing attributes in the form of a two-dimensional chart. Display the spider chart associated with the mean vector of the features.
j. Plot the spider chart associated with the top five tracks obtained in part g., together with the one for the mean vector of the features in the same figure, and discuss the results.

6. Some Explanatory Questions (10 Pts.)

Please answer the following questions as clear as possible:

a. Can an eigenvalues be a complex number? Explain.
b. Eigenvalues and eigenvectors can be used to demonstrate the convergence of a linear system. Explain how using an example.
c. Which one plays a more important role in a pattern recognition system: a high variance feature or a low variance one? Which one is more discriminative? Why?
d. Let's assume an OCR system which stores the bitmap of every character as a template and matches these templates with the observed character pixel by pixel. Discuss when this system would fail. Why do you think barcode readers are still used?
e. Imagine a face recognition system, and a 200x200 face image which would be a 40,000-dimensional vector. With one pixel shift to the left, this would be a completely different vector in the 40,000-dimensional space. How would you construct a face recognition algorithm which is robust to these variations?

Good Luck!
Ali Abbasi, Mohsen Ebadpour