Amirkabir University of Technology
(Tehran Polytechnic)

Assignment 1
A Brief 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). Please combine all your reports just into a single .pdf file.
- ***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.
- ***MATLAB is also allowable.*** By default, we assume you implement your codes in Python. If you are using MATLAB, you have to use the equivalent functions when it is asked to use specific Python 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 .py file for part b. of question 3, which must be named 'p3b.py'. (or .ipynb)
- ***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 only made through Moodle.

- ***Assignment Deadline.*** Please submit your work **before the end of November 4th**.
- ***Delay policy.*** During the semester, students are given only 7 free late days which they can use them in their own ways. Afterwards, there will be a 20% penalty for every late day, and no more than four 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 Telegram group chat or following email addresses: **m.ebadpour@aut.ac.ir** and **atiyeh.moghadam@aut.ac.ir**.

Amirkabir University of Technology
(Tehran Polytechnic)

1. Give me some images I'll tell you identification (13 Pts.)

Keywords: *Pattern Recognition System, Feature Extraction, Prediction Problems, Classification, Regression, Clustering, Human Re-Identification*

Human re-identification (Re-ID) with CCTV (Closed-Circuit Television) is a computer vision task that involves identifying and tracking individuals across multiple camera views within a surveillance network. The goal of this task is to associate a person detected in one camera frame with the same person detected in another camera's frame or across a series of frames. Human Re-ID is particularly valuable in surveillance and security applications, as it allows for the monitoring and tracking of individuals as they move through different areas covered by various cameras.

Figure 1 : One of the scarce usages of Human-ReID is the detection and tracking of protesters in urban spaces.

You are asked to suggest a practical and feasible **Pattern Recognition System** capable of human detection and Re-Identification. This system is expected to assign a **'reliability score'** to each shopping center as monitoring room.

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 and extract it from inputs?
g. Is there any pre-processing stage needed? Explain.
h. Express the challenges and difficulties that may affect the outcome of your system.
i. Which equipment do you need in deploy and operation step? How beneficial do you think it is to design such a system? Express the pros and cons of applying these systems instead of other monitoring and security 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. Statistics Warm-up (17 Pts.)

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

As a fundamental aspect of the course, Statistical Pattern Recognition implies that it is essential to develop a strong understanding of statistical techniques for analysing data measurements to extract meaningful information and make informed decisions. Therefore, mastering the basic statistical properties and being able to comprehend and utilize them is of utmost importance. In this problem, you are required to review your knowledge in this field. First, assume that birthdays are equally likely to fall on any day of the year. Consider a group of 'n' people, of which you are not a member

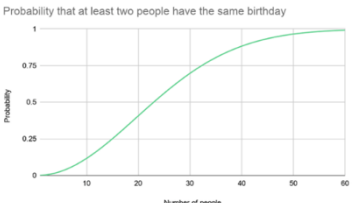

Figure 2: The probability of a birthday date being shared by at least two people in real life.

(a) Define the probability function *P* for the sample space Ω.

(b) Consider the following events:

- ***A***: "someone in the group shares your birthday"
- ***B***: "some two people in the group share a birthday"
- ***C***: "some three people in the group share a birthday"

Carefully describe the subset of Ω that corresponds to each event.

(c) Find an exact formula for *P*(*A*). What is the smallest *n* such that *P*(*A*)>0.5?

(d) Justify why *n* is greater than 365 without doing any computation.

(e) Find an exact formula for *P*(*B*).

(f) In real life, Why the probability of a birthday match is slightly higher than in the equal-probability model above? (figure 2)

Next, Using the joint probability distribution table below, answer the following questions:

*	X = 1	X = 2	X = 3
Y=1	0.05	0.13	0.17
Y=2	0.12	0.05	0.08
Y=3	0.01	0.03	0.36

(g) Calculate probability for P(X < 2, Y≤2).

(h) Calculate probability for P(X ≤2, Y≥ 3).

(i) What is the support of X, S_x?

(j) Determine the probability mass functions (PMF) of X and Y, $P_x(x)$ and $p_y(y)$

(k) Find the cumulative distribution of X.

(l) Find P(2.5 <X ≤3).

(m) Find E(X) and E(Y).

(n) Find Var(X) and Var(Y).

(o) Find conditional distribution of X given Y = 1, P(X = x |Y = 1), where x = 1,2,3.

3. Why is Normal Distribution so Normal? (20 Pts.)

Keywords: *Normal distribution, Gaussian distribution, Linear transformation, Whitening transformation, Expected Value, Variance, Mahalanobis distance.*

Imagine a bustling city with a population of millions. In this city, a group of epidemiologists is tirelessly monitoring the spread of a contagious disease. They collect data on the health of individuals, meticulously recording various health indicators like temperature, white blood cell count, and more. As they compile and analyze this data, they notice a curious pattern: when graphed, many of these health indicators follow a bell-shaped curve. This is no coincidence; it's the normal distribution in action.

Figure 3: Iranian epidemiologists researched on COVID vaccine.

The epidemiologists quickly realize that the normal distribution provides them with a powerful tool. By understanding the characteristics of this distribution, they can determine the average health indicators for the population and predict how likely it is for an individual to fall within a certain health range. This knowledge helps them make critical decisions, allocate resources efficiently, and respond effectively to the disease outbreak. In this problem, we will review some of the key characteristics of the normal distribution. First, Let X and Y be independent standard normal random variables as health indicators, that is $X \sim N(01)$ and $Y \sim N(01)$

Consider the following linear transformations:

$$U = aX + bY \; and \; V = cX + dY \qquad a\,b\,c\,d \in R$$

(a) Find the joint density of U and V, denoted by f_{UV}.

(b) For what choices of a, b, c, d are U and V independent?

(c) Show that if U and V are uncorrelated, then they are independent. (Note that uncorrelatedness does not imply independence in general.)

(d) Suppose a = 1, b = 0, c = ρ and $d = \sqrt{1 - \rho^2}$ for $\rho \in (01)$

Find:

- Correlation coefficient of U and V.
- $E(U|V)$ and $Var(U|V)$.

A In the world of finance, a team of risk analysts is tasked with assessing the volatility and correlations of various assets within a vast investment portfolio. This formidable task requires understanding the intricate relationships between these assets.

To navigate this complexity, the analysts employ the Whitening Transform, which, when applied to the Gaussian covariance matrix of asset returns, simplifies the covariance structure. It

transforms the data such that the transformed covariance matrix becomes diagonal, with equal values along the diagonal.

This remarkable property allows the analysts to untangle the web of interdependencies between assets and identify the true underlying trends and risks within the portfolio. The Whitening Transform not only simplifies the analysis but also enhances the ability to make informed investment decisions, ensuring the portfolio is resilient to market fluctuations.

Figure 4: In the realm of risk management and finance, the Stop Loss Bound is defined as a Gaussian distribution, which serves as a fundamental framework for quantifying potential losses within portfolios and trading strategies

Consider the multivariate normal distribution $p(X|\omega)\sim\mathcal{N}(\mu\,\Sigma)$ where x_0, x_1, and x_2 refers to emotional, political and fundamental risks in finance respectively.

$$\mu = \begin{bmatrix}1\\2\\2\end{bmatrix}, \Sigma = \begin{bmatrix}1 & 0 & 0\\0 & 5 & 2\\0 & 2 & 5\end{bmatrix}$$

(a) Find the probability density for emotional risk as $x_0 = (05\ 0\ 1)^t$

(b) Construct the whitening transformation A_ω. Next, transform the distribution to one centered on the origin with covariance matrix equal to the identity matrix, $p(x|\omega)\sim\mathcal{N}(0\,\mathrm{I})$.

(c) Apply the same overall transformation to $x_0$to yield a transformed point x_ω.

(d) By explicit calculation, confirm that the Mahalanobis distance from x_0 to the mean μ in the original distribution is the same as for x_ω to 0 in the transformed distribution.

(e) Does the probability density remain unchanged under a general linear transformation? Explain.

4. Can We convince each other? (17 Pts.)

Keywords: *eigenvalues, eigenvectors, eigenvalue decomposition, eigenspace, invertible matrix, diagonalizable matrix, discrete dynamical systems.*

If you have ever been in a waiting room, on a bus, or in the metro, you would invariably find a group of two or more people discussing political and economic problems or simply analyzing recent news—something we colloquially refer to as a "taxi driver's analysis."

But have you ever wondered whether or not they can truly persuade each other, or what changes in beliefs they might undergo? We can model political opinion dynamics using discrete dynamical systems.

Figure 5: Since the onset of COVID-19 and the widespread adoption of mask-wearing, there has been a noticeable decrease in the analysis conducted by taxi drivers during trips.

For each of these systems, there exists a steady-state response, which denotes the long-term behavior of that system. Eigenvalues and eigenvectors are crucial for comprehending the long-term behavior or evolution of these types of systems. Let's begin with a review of the concept of eigenvalues and eigenvectors. Given the matrix A as follow:

$$A = \begin{bmatrix} 2 & -2 & -2 \\ 3 & -3 & -2 \\ 2 & -2 & -2 \end{bmatrix}$$

(a) Find the characteristic equation of A.
(b) Find the eigenvalues and eigenvectors of A.
(c) Is A non-singular? invertible?
(d) Write a basis for each eigenspace of A.

Now, consider the following B matrix:

$$B = \begin{bmatrix} 3 & 0 & 0 \\ -3 & 4 & 9 \\ 0 & 0 & 3 \end{bmatrix}$$

(e) Find an invertible matrix P and a diagonal matrix D such that $A = PDP^{-1}$
(f) Calculate B^4 using the diagonalization in the previous part.

Given the characteristic equations below, answer the following questions:

- ✓ $p(\lambda) = \lambda(\lambda+1)^2(\lambda-3)^2$
- ✓ $p(\lambda) = (\lambda - 2)(\lambda - 1)^3$

(g) Calculate the size of assumed matrix.

(h) Is the matrix invertible? if yes, determine the eigenvalues of the inverse matrix.

Finally, let's return to our problem. Assume we have isolated three individuals with varying opinions regarding ceasing the war. We denote b_i as each person's belief, and we represent their opinions using a vector x. We define this discrete dynamical system as follows:

$$x_{k+1} = Ax_k \qquad x_k = \begin{bmatrix} b_{1k} \\ b_{2k} \\ b_{3k} \end{bmatrix} \quad A = \frac{1}{9}\begin{bmatrix} 7 & -2 & 0 \\ -2 & 6 & 2 \\ 0 & 2 & 5 \end{bmatrix}$$

We can demonstrate that as $k \longrightarrow \infty$, the vector of beliefs changes in the following manner, where λ_i represents the eigenvalues and v_i denotes the eigenvectors.

$$x_k = c_1(\lambda_1)^k v_1 + \cdots + c_n(\lambda_n)^k v_n \qquad (k = 123 \dots)$$

(i) Utilizing the explanation above, find the general solution of $x_{k+1} = Ax_k$ if $x_0 = \begin{bmatrix} 1 \\ 11 \\ -2 \end{bmatrix}$.

(j) Assuming beliefs form a spectrum where negative beliefs indicate disagreement and positive beliefs indicate agreement, if we isolate these three individuals, what would be their ultimate opinion as $k \longrightarrow \infty$?

5. Statistical Suitability Assessment for Migration: A Data-Driven Approach (28 +7 Pts.)

Keywords: *Kullback-Leibler (KL) distance, Pearson correlation coefficient, parameter estimation, normalization, heat map generation.*

In a rapidly evolving society where early childhood education plays a pivotal role in shaping the future, the importance of quality childcare services cannot be overstated. The National Database of Childcare Prices (NDCP) has emerged as a crucial resource, offering in-depth insights into childcare costs across the nation. As educators, policymakers, and parents grapple with the ever-increasing demands of childcare, the NDCP database becomes a beacon of hope. It provides data spanning from 2008 to 2018, allowing us to track trends, regarding the welfare of our children.

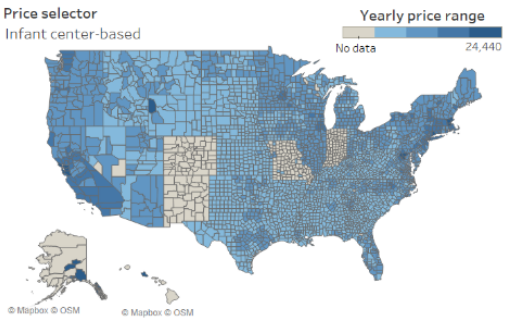

Figure 6: NDCP Map for 2016-2018

The effects of analyzing the NDCP database are far-reaching and profound. Firstly, it enables researchers and educators to uncover critical trends in childcare pricing over the years. Are childcare costs rising or falling? Which regions experience the most significant disparities? Answers to these questions can guide policymakers in crafting effective and equitable childcare policies, ensuring access for all children, regardless of socio-economic backgrounds. Moreover, the analysis can assist parents in making informed decisions about childcare providers, enabling them to provide the best possible early education for their children.

The dataset comprises two distinct .csv files. The first file contains information regarding various regions and countries, while the second file details the statistical data for each country across different years. (information about features)

(a) Begin by conducting an initial exploration of the dataset to gain a comprehensive understanding of its contents. Following this exploration, proceed to load the dataset and apply the necessary normalization to specific features (which ones?). It's essential to understand why normalization is a vital step in this dataset. Explain in the report.

(b) In your perspective, do you believe that the statistics within each state exhibit high variance or low variance, and what leads you to this conclusion? For each state in each year, calculate both the mean and variance of the features derived from different countries. To substantiate your response, employ a visualization or chart that aligns with your rationale. This analysis will serve to reinforce your hypotheses regarding variance, both within the same year and within individual states.

(c) Given the foundational aspects of this task, are the features displaying linear correlations? Calculate the Pearson correlation coefficient and then proceed to reduce the feature set to

half. Additionally, identify the top 5 pairs of features that exhibit the highest degree of linear dependence and indicate you keep which one of them.

Figure 7: Mandel was born in Toronto and raised in the Willowdale area of Toronto, Ontario.

Howard Michael Mandel, a Canadian comedian, television personality, actor, and producer, is planning to relocate to the United States. Being a person who values precision and exactitude, he is keen on selecting a suitable state for his move. To define this suitability, he looks to the statistical behavior of each state. Here's how we can assist him in making this decision. In this step, we aim to assess the suitability of each state based on the normality of its statistical behavior. The Kullback-Leibler (KL) divergence, a fundamental concept in information theory and statistics, is employed to measure the dissimilarity between two probability distributions. It is calculated using the following formula:

$$D_{KL}(P||Q) = \sum P(x) log \frac{P(x)}{Q(x)}$$

In the context of selecting a suitable state for migration, the KL divergence can be used to gauge the distance between the distribution of a state's statistical data and a reference standard normal distribution. By quantifying the information lost when approximating one distribution with another, the KL divergence provides a valuable metric for assessing the similarity or dissimilarity between datasets, assisting in the selection of the most statistically compatible state for migration.

(d) Is the Kullback-Leibler (KL) divergence a symmetric measure? Additionally, when using KL divergence for assessing the suitability of states based on their statistical behavior, what is the appropriate choice for P and Q? Should P represent the statistical data of the states under consideration, or should it be a generated distribution representing the standard normal behavior? Explain your rationale for the choice of P and Q in this context. We search for high KL or low?

Begin by treating the data for each state, considering all years and countries within it, as if it arises from a multivariate normal distribution.

(e) Calculate the mean and variance for each state based on features that obtained from part c. Estimate a normal distribution with the calculated parameters (mean and variance) for each state. Generate equal-sized samples from these distributions.

(f) Compute the Kullback-Leibler (KL) distance between each state's data distribution and the estimated distribution. The KL distance helps determine which state's statistical behavior aligns most closely with a standard normal distribution.

(g) Present the results using a descending bar chart, showcasing the level of similarity to a standard normal distribution for each state. This chart will provide a clear visual representation to help Michael make an informed decision on which state to select.

(h) Based on the analysis conducted in the above, Michael can choose the state that exhibits the statistical behavior closest to a standard normal distribution. Repeat it to choose the country.

(i) To visualize the outcomes of the previous two steps, consider creating a heatmap of the United States (like Fig. 6). In the heatmap, color-code(int or double) the states to indicate their level of similarity to a standard normal distribution. States that are most similar can be represented in one color, while states that deviate from normality can be shown in another. This heatmap will serve as a visual guide for Michael, illustrating the suitability of states based on statistical normality. (You have the option to utilize publicly available libraries and resources to facilitate this task)

Good Luck!

Mohsen Ebadpour, Atiyeh *Moghadam, Romina Zakerian*