Assignment 1

Practicing Statistical Pattern Recognition Basics

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 don't use implementation tools when it is asked to solve the problem by hand, otherwise you'll be penalized and lose some points.
- ***Don't bother typing!*** You are free to solve by-hand problems on a paper and include picture of them 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. Don't forget to explain what you have done, and provide enough discussions when it's needed.
- ***Appearance matters!*** In each homework, 5 points (out of a possible 100) belongs 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're using Python, you have to use 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. Please name it like 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's where all assignments as well as course announcements are posted on. Homework submissions are also done through Moodle.

- ***Assignment Deadline.*** Please submit your work **before the end of December 5th**.
- ***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 don't hesitate to contact us through the following email addresses: **ali.the.special@gmail.com** and **b.roshanfekr@yahoo.com**.

1. Designing Simple Pattern Analysis Systems (15 Pts.)

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

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:

a. Predicting your final grade in this course
b. Predicting US Dollar to Iranian Rial exchange rate in the coming year
c. Grouping students in a dorm by their personalities

In each one of the scenarios, please answer the following questions:

1. Which types of prediction problems (classification, regression, etc.) does it belong to?
2. What sensors (if any) are needed?
3. What is your training set?
4. How do you gather your data?
5. Which features do you select?
6. Is there any pre-processing stage needed? Explain.
7. Express the challenges and difficulties that may affect the outcome of your system.
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.

Note 1: There is no limitations on the methods 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, e.g. features must be discriminative and measurable.

2. Getting More Familiar with the Art of Feature Extraction (12 Pts.)

Keywords: *Feature Extraction, Classification Problems, Race Recognition, Facial Expression Detection, Age Detection, Facial Recognition, Optical Character Recognition (OCR), Geometric Transformation*

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} or {A,B,C,E,F,G} and {D,H} (i.e. **Race Recognition** problem)
b. {Happy} and {Neutral} or {A,C,D,G,H} and {B,E,F} (i.e. **Facial Expression Recognition** problem)
c. {Young} and {Adult} or {A,C,D,E,G,H} and {B,F} (i.e. **Age Detection** problem)
d. {Male} and {Female} or {A,B,D,G} and {C,E,F,H} (i.e. **Gender Recognition** problem)
e. {A}, {B}, {C}, {D}, {E}, {F}, {G} and {H} (i.e. **Facial Recognition** problem)

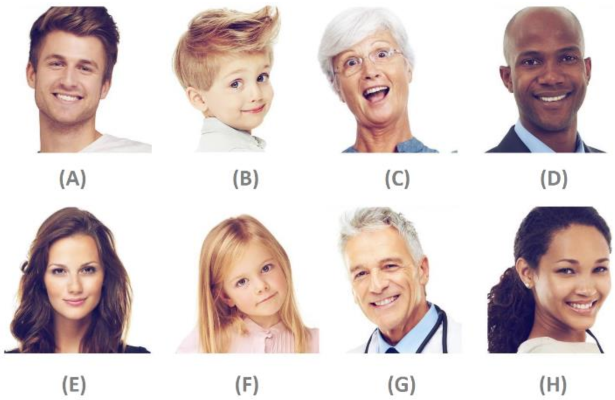

Figure 1 A toy dataset consisting of 8 subjects, given for part a. to part e.

3. Basic Statistics Warm-up (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, find the following quantities for a random variable X with the probability density function:

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

a. c
b. $\mathrm{P}(0\le X\le 0.5)$
c. $\mathrm{E}[X]$
d. $\mathrm{Var}(X)$
e. $\mathrm{E}[2X-2]$
f. $\mathrm{Var}[2X-2]$

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

g. Calculate $P\{-2\le X\le 1\}$.
h. Calculate $\mathrm{E}[X]$ and $\mathrm{Var}(X)$.
i. Find the distribution of $Y=2X-1$. Express what type of random variable it is, and find its parameters.

Then, suppose that in Amirkabir University, 1/5 of the students are going to fail a certain course (not Pattern Recognition!). Seven students are selected randomly.

j. What is the probability that exactly 4 students of them pass this course?

Next, consider a continuous random variable X has the following probability density function:

$$f(x)=\begin{cases} \frac{1}{4}(4-x^2) & 0\le x\le 2 \\ 0 & \text{otherwise} \end{cases}$$

k. Find the median value of X .

Now, assume a DVD disc production company produces discs with a normally distributed diameters with a mean of 10 cm and standard deviation of 0.1 cm.

l. What is the probability of a produced disc having a diameter less than 9.8 cm?

Finally, assume a continuous random variable with the following probability density function:

$$f(x) = \begin{cases} \dfrac{4}{\pi(1+x^2)} & 0 \le x \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

Calculate $\mathrm{E}(X)$.

4. Mastering Eigenvalues and Eigenvectors and Their Properties (18 Pts.)

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

Eigenvalues and **Eigenvectors** are vastly used in various scientific areas, from geology and ecology to computer vision and data mining. They also play an important role in the field of pattern recognition, where it is applied in many applications such as calculating covariance matrix or principle component analysis (PCA).

In this problem, we are going to take a deeper look into these concepts.

a. For each of the following pairs, determine whether v is an eigenvector of $\mathbf{A}$ or not. If so, specify the corresponding eigenvalue.

a1. $\mathbf{A} = \begin{bmatrix} 1 & 4 \\ 3 & 2 \end{bmatrix}, v = \begin{bmatrix} -3 \\ -3 \end{bmatrix}$

a2. $\mathbf{A} = \begin{bmatrix} 5 & 6 & 2 \\ 0 & -1 & -8 \\ 1 & 0 & -2 \end{bmatrix}, v = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$

b. Determine a basis for the eigenspace corresponding to each of the following eigenvalues.

b1. $\mathbf{A} = \begin{bmatrix} 1 & -6 \\ -3 & 4 \end{bmatrix}, \lambda = -2$

b2. $\mathbf{A} = \begin{bmatrix} -2 & 4 & 2 \\ 2 & 1 & -2 \\ 4 & -2 & 5 \end{bmatrix}, \lambda = 6$

c. Calculate the eigenvalues and eigenvectors associated with the following matrices.

c1. $\mathbf{A} = \begin{bmatrix} 3 & 2 \\ -1 & 6 \end{bmatrix}$

c2. $\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$

c3. $\mathbf{A} = \begin{bmatrix} 3 & 1 & -2 \\ 2 & 3 & -2 \\ 2 & 1 & -1 \end{bmatrix}$

d. Find a 2×2 matrix $\mathbf{A}$ with eigenvalues $\lambda_1 = -1$ and $\lambda_2 = 7$, and corresponding eigenvectors $v_1 = \begin{bmatrix} -\frac{3}{2} \\ 1 \end{bmatrix}$ and $v_2 = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$.

e. Let $\mathbf{A}$ be an invertible matrix with eigenvalue λ. Assuming a nonzero x satisfies $\mathbf{A}x = \lambda x$ show that $\mathbf{A}^{-1}$ has an eigenvalue equal to λ^{-1}.

f. Assume $\mathbf{A}$ is a matrix of the size 3×3 with two eigenvalues, and the corresponding eigenspaces are one-dimensional. Justify whether $\mathbf{A}$ is diagonizable or not.

g. Show that if a $n\times n$ matrix $\mathbf{A}$ has an eigenvalue equals to 3, then $\mathbf{A}^2$ has an eigenvalue equals to 9.

h. Consider the sequences x_n and y_n, such that for each $n\geq1$,

$$x_n = 2x_{n-1} - 3y_{n-1}, \quad y_n = -4x_{n-1} + y_{n-1}$$

Assuming $x_0 = 2$ and $y_0 = 3$, specify each of x_n and y_n explicitly in terms of n.

Useful Links: [1]

5. Simple Sample Generation and Beyond (25 Pts.)

Keywords: *Sample Generation, Normal Distribution, Linear Transformations, Simultaneous Diagonalisation, Whitening Transformation*

In many pattern recognition applications, **Sample Generation** plays an important role, where it is necessary to generate samples which are to be normally distributed according to a given expected vector and a covariance matrix.

In this problem, you are going to do this technique yourself. You will also practice some more complicated matrix operations as well.

a. Generate samples from three normal distributions specified by the following parameters:

$$n=1, \quad N=500, \quad \mu=5, \quad \sigma=1,2,3$$

Plot the samples, as well as the histograms associated with each of the distributions. Compare the results.

b. Generate samples from a normal distributions specified by the following parameters:

$$n=2, \quad N=500, \quad M=\begin{bmatrix}2\\1\end{bmatrix}, \quad \Sigma=\begin{bmatrix}2 & 1\\1 & 3\end{bmatrix}$$

Display the samples, as well as the associated contour plot.

c. Consider a normal distribution specified by the following parameters:

$$n=2, \quad N=500, \quad M=\begin{bmatrix}m_1\\m_2\end{bmatrix}, \quad \Sigma=\begin{bmatrix}\sigma_{11} & \sigma_{12}\\\sigma_{21} & \sigma_{22}\end{bmatrix}$$

Determine appropriate values for each of the unknown variables, so that the shape of the distribution becomes:

c1. A circle in the upper left of the Euclidean coordinate system.

c2. A diagonal line (/ shape) in the centre

c3. A horizontal ellipsoid in the lower right of the Euclidean coordinate system

Display the generated samples.

d. Consider a random variable with

$$X=\begin{bmatrix}x_1\\x_2\end{bmatrix}, \quad M=\begin{bmatrix}m_1\\m_2\end{bmatrix}, \quad \Sigma=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\\rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$$

Compute $d^2(X)$ analytically, if the parameters are selected as

$$m_1=2, m_2=3, \sigma_1^2=1, \sigma=4$$

$$\rho=-0.99, -0.5, 0.5, 0.99$$

e. Plot the contour lines for $d^2(X) = 4, 9, 16$.
f. Calculate the sample mean $\hat{M}$, and sample covariance matrix $\hat{\Sigma}$ of the distribution in part b., and comment on the results.
g. Simultaneously diagonalise Σ and $\hat{\Sigma}$, and form a vector $V = [\lambda_1, \lambda_2]^T$.
h. Find a transformation for covariance matrix of the distribution in part b., such that when applied on the data, the covariance matrix of the transformed data becomes $\mathbf{I}$. Transform the data and display the distribution in the new space.
i. Calculate the eigenvalues and eigenvectors associated with the covariance matrix of the distribution in part b. Plot the eigenvectors. What can you infer from them?
j. Again, consider the distribution and samples you generated in part b. Construct a 2×2 matrix $\mathbf{P}$, which has eigenvectors associated with Σ as its columns ($\mathbf{P} = [v_1, v_2]$, such that v_1 is corresponding to the largest eigenvalue). Project your generated samples to a new space using $\mathbf{Y}_i = (\mathbf{X}_i - M) \times \mathbf{P}$, and plot the samples. What differences do you notice?
k. Find the covariance matrix associated with the projected samples in part h. Also calculate its eigenvalues and eigenvectors, and comment on the results.

Recommended MATLAB functions: `meshgrid(), mvnpdf(),mvnrnd(),eig()`

6. Some Explanatory Questions (10 Pts.)

Please answer the following questions as clear as possible:

a. Why do you think Central Limit Theorem is important? Where and how can it be used?
b. What is the difference between a feature and a measurement?
c. Does a covariance matrix need to be symmetric? Why?
d. What does zero eigenvalue mean?
e. When does the whitening transformation come into use?

Good Luck!
Ali Abbasi, Behnam Roshanfekr