Assignment 1

Practicing Statistical Pattern Recognition Basics

Homeworks Guidelines and Policies

1. What you must hand in includes the assignment report (.pdf) and – if necessary – source codes (.m). Please zip them all together into an archive file named according to the following template: HW1_XXXXXXXX.zip
   Where XXXXXXXX must be replaced with your student ID.
2. Some problems are required to be solved *by hand* (shown by the ✎ icon), and some need to be implemented (shown by the ▲ icon).
3. As for the first type of the problems, you are free to solve them on a paper and include the picture of it in your report. Here, cleanness and readability are of high importance.
4. 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.
5. 5 points of each homework belongs to compactness, expressiveness and neatness of your report and codes.
6. 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.
7. 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.
8. Problems with bonus points are marked by the ★ icon.
9. **Please upload your work in Moodle, before the end of March 16th.**
10. If there is ***any*** question, please don't hesitate to contact me through the following email address: **ali.the.special@gmail.com**
11. Unfortunately, it is quite easy to detect copy-pasted or even structurally similar works, no matter being copied from another student or internet sources. Try to send us your own work, without being worried about the grade! ;)

1. Designing Simple Pattern Analysis Systems (20 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. Labeling a collection of books by their genres
c. Predicting the winner of 2018-19 UEFA Champions League
d. Grouping students in a dorm by their personalities
e. Predicting the price of meat in the coming month

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 (16 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)

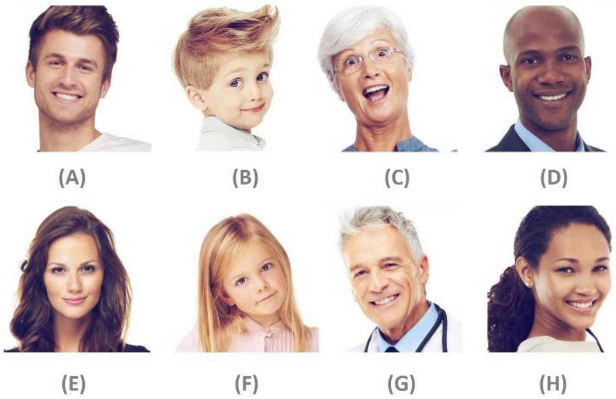

Figure 1 A toy dataset consisting of 8 subjects, given for part a. to part e.

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 (Figure 2):

f. Translation
g. Scaling
h. Rotation
i. Reflection
j. Shear

Hint: Here, a feature is invariant if it can be used to detect the same letter in a normal (without degradations) and a changed (e.g. 2x larger) image.

Useful Links: [1], [2]

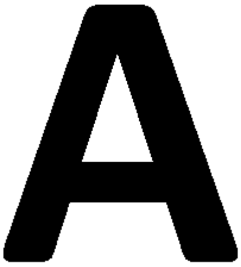

Figure 2 A simple input given in an OCR problem

Note: Your features must be properly measurable.

3. Feature Selection: Evaluating Features to Select '*Good*' Ones (8 Pts.)

Keywords: *Feature Selection, Linear/Nonlinear Separable Data, Data Correlation, Multi-modality Data*

After **Feature Extraction** stage, it is advised to apply a process to select a subset of relevant features. The motivation behind this process, called **Feature Selection**, is that the data may contain redundant on irrelevant features, and can thus be reduced without incurring much loss of information. One can apply this process for many reasons, from simplifying the model and avoiding *the curse of dimensionality* to reducing training times and enhancing generalisation.

In this problem, you are going to practice simple feature selection tasks. You will find out more about **Feature Selection** as the course goes on.

Download the famous Iris dataset and load it in MATLAB. As you can see, the dataset contains 3 classes, and each sample has 4 attributes.

Suppose you want to keep two features for classification. Plot the distributions for all feature pairs (6 figures in total), and answer the previous questions for these feature pairs.

a. Which features are *good*, and which ones are *bad*? Explain your reasons.
b. Investigate each feature in terms of linear/non-linear separability.
c. Investigate each feature in terms of correlation among the samples.
d. Investigate each feature in terms of modality between the samples.

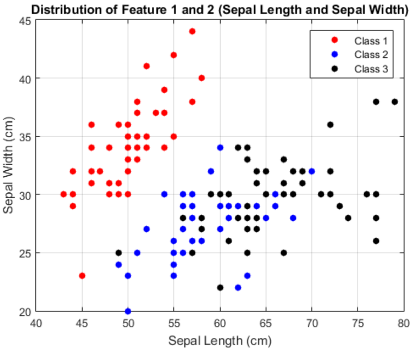

Figure 3 Distribution of the first two features

Hint 1: The first one has been done for you (Figure 3).

Hint 2: In order to clearly highlight the differences, you can plot the features belonging to each class in different vertical levels, as can be seen in Figure 4.

Hint3 : In MATLAB, you can easily load Iris dataset by using `load iris.dat`

Note 1: Remember to highlight samples of each class with different colors.

Note 2: Include the resultant figures of each part in your report.

Recommended MATLAB functions: `plot()`

4. Basic Statistics Warm-up (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*

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:

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

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

5. Hanging Around with Covariance Matrix and Linear Transformations (8+3 Pts.)

Keywords: *Covariance, Covariance Matrix, Data Dimension, Data Correlation, Eigenvalues, Eigenvectors, Linear Transformations, Whitening Transformation*

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.

Consider a dataset with covariance matrix $\Sigma = \begin{bmatrix} 1 & 3 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}$, and answer the following questions.

a. Specify the dimensionality of the dataset, i.e. the number of features each sample has.
b. Determine the number of samples in the dataset.
c. Find the correlations between different data dimensions.
d. On which dimension are the data 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.
f. Find a transformation to whiten data associated with the given covariance matrix.
g. Show that Σ is a valid covariance matrix.

6. Simple Sample Generation and Beyond (18 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()`

7. Some Explanatory Questions (5+2 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?

★ **f.** Explain how does Google's PageRank algorithm use eigenvalues and eigenvectors concepts.

Good Luck!
Ali Abbasi