Assignment 2
Decisions, decisions, decisions!

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: HW2_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 April 13th.**
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. Bayesian Inference for Some Well-known Probability Theory Paradoxes (9 Pts.)

Keywords: *Bayesian Inference (Reasoning), Prior Probability, Posterior Probability, Likelihood Function, Probability Theory Paradoxes*

Bayesian Inference is the application of Bayes theorem to update the probability for a hypothesis when more evidence or information is being provided. It derives the **Posterior Probability** as a consequence of a **Prior Probability** and a **Likelihood Function** derived from a statistical model for the observed data.

In this problem, your task is to explain some famous probability theory puzzles using Bayesian inference. These puzzles are called '*paradox*', because in each one of them the initial reasoning you may have is probably different from the true answer of the problem.

a. Assume you need two pens and you search your bag, which has three separate pockets. Each one of these pockets has two pens inside; one has two blue pens, the other has two black pens and the last has one blue pen and one black pen. You randomly select a pocket, and realize it has one blue pen. The probability that the other pen is also blue in the selected pocket is intuitively $1/2$, but it is actually $2/3$. Explain why.
b. Suppose you are given two indistinguishable bags of money, one contains twice as much as the other. You may choose one bag and keep the money it contains. You choose a bag, but before opening it, you are given the chance to change your decision. Although it may seem that there is no point in changing your decision, it is possible to argue that it is more beneficial for you to switch the bags. Explain why.
c. Imagine a family has two children. There are two scenarios, 'the older child is a boy' and 'at least one of them is a boy'. In each scenario, the probability that the other child is also boy might seem to be $1/2$, but it's not. Calculate the probabilities and explain the reason.

Note: Although these puzzles are also explainable using other probabilistic methods, please stick to the Bayesian analysis only.

2. Minimum Distance Classifier: A Basic Yet [Sometimes] Powerful Classifier (10 Pts.)

Keywords: *Classification Problem, Minimum Distance Classifier*

When prior probabilities are the same for all classes, the optimum decision rule can be stated very simply: to classify a new sample by finding the class that has a *prototype* with the minimum Euclidean distance to the new sample. Such a classifier is called a **Minimum Distance Classifier**. Although a basic algorithm, **MDC** is a very fast classification method which in some cases works pretty well.

In this problem, you are to practice MDC in two different problems. First, assume the following toy dataset:

$$\omega_1 : \begin{bmatrix} -1.0 \\ 0.8 \end{bmatrix}, \begin{bmatrix} -1.0 \\ 0.9 \end{bmatrix}, \begin{bmatrix} -1.0 \\ 1.3 \end{bmatrix} \qquad \omega_2 : \begin{bmatrix} -1.0 \\ -0.5 \end{bmatrix}, \begin{bmatrix} -1.2 \\ -1.5 \end{bmatrix}, \begin{bmatrix} -0.8 \\ -1.0 \end{bmatrix} \qquad \omega_3 : \begin{bmatrix} 1.3 \\ 0.5 \end{bmatrix}, \begin{bmatrix} 0.9 \\ -0.2 \end{bmatrix}, \begin{bmatrix} 0.8 \\ -0.3 \end{bmatrix}$$

a. What are the prototype vectors for a minimum distance classifier?
b. Determine the label of the test vector $\begin{bmatrix} -0.5 & 1.0 \end{bmatrix}^{\mathrm{T}}$.

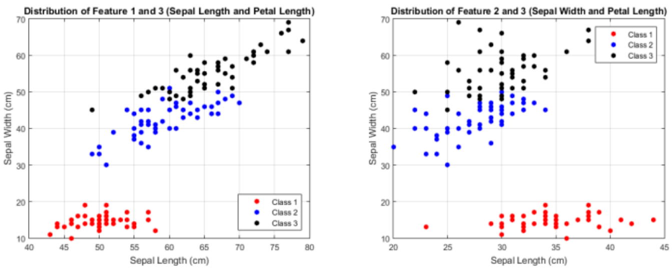
c. Repeat the previous part assuming the decision function $d_j(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{m}_j - \frac{1}{2}\mathbf{m}_j^{\mathrm{T}}\mathbf{m}_j$, in which $\mathbf{m}_j$ is the prototype vector for class j.

d. Write down the equation of the decision boundary separating classes 1 and 2 for a MDC classifier.

e. Repeat the previous part for classes 1 and 3.

Now let's consider a more real-world problem. Load the iris dataset, and consider <u>only first and third features</u> (sepal length and petal length).

f. Classify the following points using MDC.



(a) (b)

Figure 1 Feature space when considering first (a) 1st and 3rd (b) 2nd and 3rd features of the iris dataset

$$X_1 = (59, 48) \qquad X_2 = (50, 25) \qquad X_3 = (65, 50) \qquad X_4 = (55, 30) \qquad X_5 = (67, 25)$$

g. Sketch the MDC decision regions and decision boundaries for this problem.

h. Repeat the previous part by considering <u>only second and third features</u> (sepal width and petal length).

3. Making Decisions using Bayes Decision Rule (Part I) (12 Pts.)

Keywords: *Classification Problem, Bayes Decision Rule*

Bayes Decision Rule is a decision theory which is informed by **Bayesian Probability**. By using probabilities and costs, Bayes decision rule tries to quantify the trade-off between various decisions. A classifier who applies such a decision theory uses the concepts of Bayesian statistics to estimate the expected value of its decisions.

Assume an **Image Segmentation** problem, in which the goal is to classify pixels of an image into three classes, 'sabzeh', 'ribbon' and 'background'. The training set is as follows:

Pixel	Label	Channel 1	Channel 2	Channel 3
1	ω_1	69	116	14
2	ω_1	88	132	35
3	ω_1	104	157	17
4	ω_1	25	22	17
5	ω_1	99	145	98
6	ω_2	208	30	44
7	ω_2	113	19	20
8	ω_2	159	11	23
9	ω_2	199	19	31
10	ω_2	249	141	167
11	ω_3	226	231	235
12	ω_3	133	131	116
13	ω_3	218	214	211
14	ω_3	186	176	174
15	ω_3	207	202	199

Figure 2 Three different classes defined in an arbitrary image, 'sabzeh', 'ribbon' and 'background'

The dataset contains 15 pixels and their RGB values, picked from the indicated regions in Figure 2. Suppose that <u>you are only allowed to use two channels</u>.

a. By visual inspection using 2D feature space, evaluate which two channels are the most suited.
b. Design a classifier using the Bayes rule by considering the two features you picked in the previous part. The data are assumed to have Gaussian distributions with the same covariance matrix $\Sigma = \mathbf{I}_2$. Find the general form of the discriminant function.
c. Classify the following pixels using the functions you obtained in the previous part.

Pixel	Channel 1	Channel 2	Channel 3
1	196	34	49
2	195	180	185
3	255	88	116
4	125	110	43
5	110	146	108

Next, consider 6 different datasets given in Figure 3. The goal is to design a Bayes classifier assuming Gaussian distribution for the data. The covariance matrices are not equal across classes, but they are diagonal on the form $\Sigma_i = \sigma^2 \ \mathbf{I}$.

d. In each case, sketch the Bayes decision boundary. If the classifier breaks down, explain.

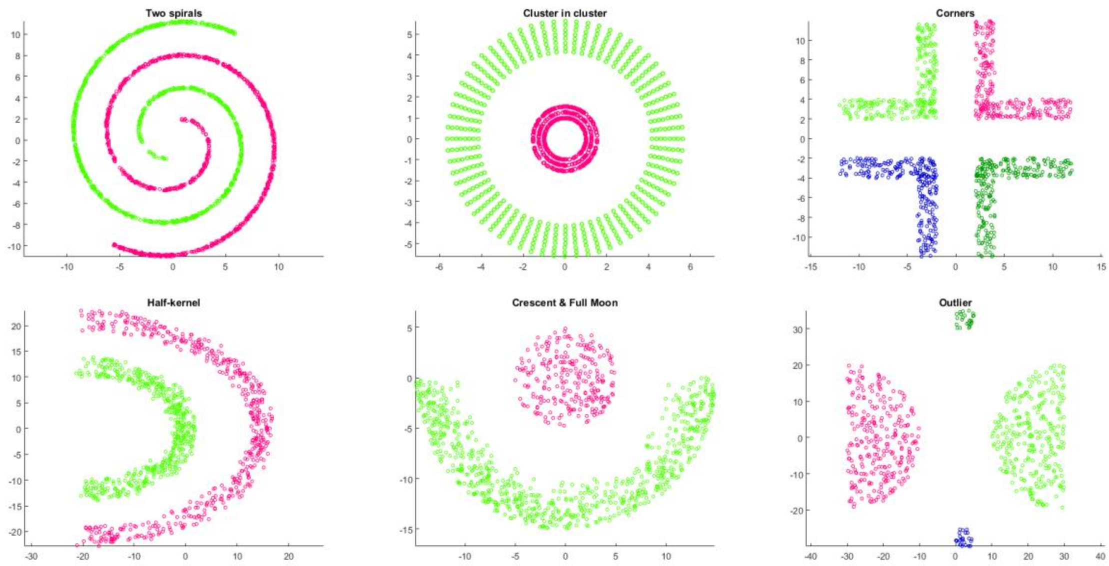


Figure 3 Six different datasets with different distributions

Hint: You can use the provided images of these datasets (attached to this homework) and an image editor software (like Microsoft Paint).

4. Making Decisions using Bayes Decision Rule (Part II) (10 Pts.)

Keywords: *Classification Problem, Bayes Decision Rule*

Following the previous problem, here we are going to face a more complicated case, where covariance matrices belonging to each classes are not identical.

In this problem, X is a two dimensional feature vector and $p(\mathrm{x} \mid \omega_i) \sim N(\mu_i, \Sigma_i)$, $i = 1, 2$, as its class conditional distribution where $\mu_1 = \begin{bmatrix} 2 & 3 \end{bmatrix}^{\mathrm{T}}$, $\mu_2 = \begin{bmatrix} 0 & -2 \end{bmatrix}^{\mathrm{T}}$, $\Sigma_1 = \begin{bmatrix} 2 & -0.8 \\ -0.8 & 2 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}$.

a. Sketch the contours of constant values for two class conditional densities.
b. Sketch the decision boundary for Bayesian classifier and minimum distance classifier.
c. Generate 1000 samples for each class and estimate the classification error for each classifier.
d. Considering the following cost matrix, find the decision boundary for Bayes classifier and compare f-score for generated samples with Bayes classifier in part b.

$$\begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 & 3 \\ 1 & 0 \end{bmatrix}$$

e. Repeat part c. for 20 times, and report an average error for both classifiers.

5. Dealing with Error in Bayes Decision Rule (8 Pts.)

Keywords: *Bayes Decision Rule, Probability of Error, Upper Bounds of Error Probability, Bhattacharyya Error Bound, Chernoff Error Bound*

In general, the **Bayes Decision Rule** – or any other decision rule – does not lead to perfect classification. In order to measure the performance of a decision rule, one must calculate the **Probability of Error**, which is the probability that a sample is assigned to a wrong class.

In practice, calculating the error probability is a difficult task. We may seek either an approximate expression for the error probability, or an upper bound on the error probability. **Bhattacharyya Error Bound** and **Chernoff Error Bound** are some **Upper Bounds of Error Probability**.

Assume $X \in (-1,1)$ is a one dimensional feature which is used to decide between two categories ω_1 and ω_2, with conditional densities as follows:

$$f(x \mid \omega_1) = \begin{cases} \frac{1}{2} & -1 < x < 1 \\ 0 & otherwise \end{cases} \qquad f(x \mid \omega_2) = \begin{cases} x+1 & -1 < x \le 0 \\ -x+1 & 0 < x < 1 \\ 0 & otherwise \end{cases}$$

Assuming $p(\omega_1) = p(\omega_2)$,

a. Find the decision regions for a Bayes decision rule.
b. Compute the probability of error for this rule.
c. Calculate the Bhattacharyya error bound.
d. Calculate the Chernoff error bound.

6. Bayes Risk Estimation: When Small Decisions Have Big Consequences (14 Pts.)

Keywords: *Conditional Risk, Bayes Risk, Decision Cost*

Generally, in decision-theoretic terminology, **Risk** is defined as an expected loss, and $R(\alpha_i \mid \mathrm{x})$ is called **Conditional Risk**. For a particular observation x, the goal is to minimize the expected loss by selecting the action that minimizes the conditional risk. **Bayes Risk**, however, is defined as the best performance that can be achieved by a classifier which applies Bayes decision rule.

First, consider the following class-conditional density functions for a discrete random variable X and four classes:

x	$p(x \mid \omega_1)$	$p(x \mid \omega_2)$	$p(x \mid \omega_3)$	$p(x \mid \omega_4)$
1	$\frac{3}{7}$	$\frac{1}{3}$	$\frac{2}{5}$	$\frac{5}{6}$
2	$\frac{4}{7}$	$\frac{2}{3}$	$\frac{3}{5}$	$\frac{1}{6}$

Also, assume the loss function , such that

	ω_1	ω_2	ω_3	ω_4
α_1	0	3	2	1
α_2	4	0	1	8
α_3	7	1	0	5
α_4	5	2	4	0

Note that α_i means "decide class ω_i ". Let the class prior probabilities be $p(\omega_1)=\frac{3}{8}$, $p(\omega_2)=\frac{1}{8}$, $p(\omega_2)=\frac{1}{4}$ and $p(\omega_2)=\frac{1}{4}$.

a. Find the conditional risk for each action α_i.
b. Find the overall risk R.
c. What is the optimal Bayes decision rule?

Now, assume the following conditional densities for two classes ω_1 and ω_2 in a binary classification problem with a scalar feature x:

$$p(x \mid \omega_1)=k_1 \exp\left(-(\mathrm{x}+4)^2/20\right) \qquad p(x \mid \omega_2)=k_2 \exp\left(-(\mathrm{x}-2)^2/24\right)$$

d. Find k_1 and k_2.
e. Plot two densities on a single graph and draw the decision boundary.
f. Write down the expression for the conditional risk, assuming equal prior probabilities for the two classes, equal cost for correct decisions and unequal cost for incorrect decisions, where $C_{21}=\sqrt{3}$ and $C_{12}=\sqrt{2}$ (First index indicates classifier decision, while the second one shows true class).
g. Find and plot the decision regions for Bayes minimum risk.
h. What is the numerical value for the Bayes minimum risk?

7. ROC Curve: A Simple Yet Very Informative Graph for Evaluating a Classifier (7 Pts.)

Keywords: *ROC Curve, True Positive Rate (also Sensitivity or Recall), False Positive Rate (also Fall-out or Probability of False Alarm)*

A **Receiver Operating Characteristic Curve**, aka **ROC Curve**, is a graphical illustration of the ability of a binary classifier when its discrimination threshold is changed. It is created by plotting the **True Positive Rate** (**TPR**), i.e. the **Sensitivity** or **Recall**, with respect to the **False Positive Rate** (**FPR**) at various threshold settings.

Let's get more familiar with it. Figure 4 illustrates the class-conditional probability distributions of observation x given the two classes of a binary classification problem.

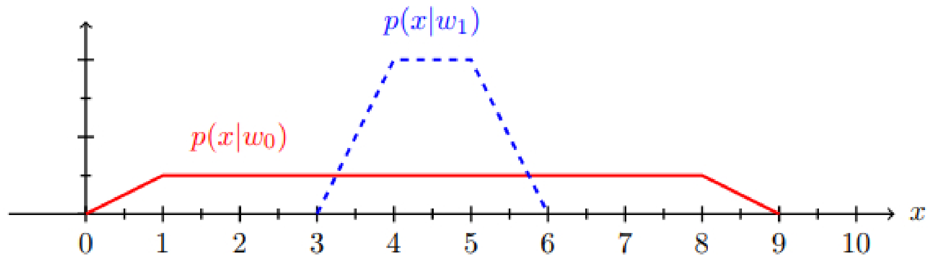


Figure 4 Class-conditional probability distributions corresponding to the two classes in this problem

a. Assuming $p(\omega_0)=0.6$, determine and sketch the decision regions for the minimum-error decision rule. Call them $\Re_0$ and $\Re_1$.
b. Compute the corresponding probability of error.
c. Let $p_{fa}=p(\Re_1|\omega_0)$ as the probability of false alarm, and $p_{md}=p(\Re_0|\omega_1)$ as the probability of misdetection. Plot an ROC curve for the following decision rule

   Assign x to ω_1 if $x>\tau$

   Highlight specific points for at least five different values of τ.

8. Evaluation of Bayes Decision Rule in Skin Detection Problem (15+5 Pts.)

Keywords: *Classification Problem, Bayes Decision Rule, Confusion Matrix, Bayes Error, ROC Curve, Skin Detection*

Up until now, you've encountered some problems regarding to Bayes decision rule and the related topics. It's time to deal with these concepts in a more practical manner.
In this problem, you will get hands-on experience in implementing a classifier for **Skin Detection** problem based on Bayes decision rule. You will work with a human skin detection dataset, known as Pratheepan dataset (Figure 5). You don't need to download it, as a customized version of this dataset is provided for you in the homework directory.

In this dataset, there a 78 images divided into train (66) and test sets (12), and the training images are also divided into two separate sets, one contains single subjects with simple backgrounds, and the other contains multiple subjects with complex backgrounds. Each image has a corresponding groundtruth pair placed in a separate folder, which is actually a binary image with value 255 (white) for the 'skin' and zero (black) for the 'non-skin' pixels.

(a) (b)

Figure 5 Some examples of images in the Pratheepan skin detection dataset and their corresponding groundtruth (a) images with single subject and simple background (b) images with multiple subjects and complex background

First, assume train images placed in ‘set1’ directory.

Figure 6 In addition to the given test set, two images – one with simple background and the other with complex background – are also given to test your algorithm in ‘skin/non-skin’ task

a. Find the class priors using the training set. Report the prior probabilities of a pixel being ‘skin’ or being ‘non-skin’.
b. If we decide to model the class-conditional probability density of each class using a univariate Gaussian, what would be the mean and variance of both class-conditional densities?
c. Classify the pixels in image ‘mad_trump.jpg’ and ‘trump_long_tie.jpg’ (Figure 6) and display the results, i.e. the groundtruth image.
d. Repeat the previous part with a MDC classifier. Display and compare the results.
e. Classify the pixels of the images in ‘test’ folder and report the overall test error.
f. Compute a confusion matrix for this classifier.
g. Calculate the Bayes error.
h. Draw a ROC curve to visualise the performance of the classification.
i. Repeat the previous parts, this time considering <u>all training images</u> (‘set1’ and ‘set2’). Compare the results.

Hint: In RGB space, each pixel has three values (0-255) for each of the red, green and blue channels. Therefore, here we have a two-class three-dimensional classification task.

Note: Groundtruth images are originally in RGB space. It would be easier to convert them to grayscale before using them.

Recommended MATLAB functions: `imread()`, `rgb2gray()`, `confusionmat()`, `dir()`, `fullfile()`,`trapz()`,`trapz()`

9. Some Explanatory Questions (10 Pts.)

Please answer the following questions as clear as possible:

a. How do you improve Minimum Distance Classifier? Suggest at least two modifications.
b. Does a Minimum Distance Classifier have a training phase? What about a minimum-error classifier? Explain.
c. Assume a two-class 1D classification problem with the Gaussian distributions $p(\mathrm{x}|\omega_1) \sim N(-1,1)$ and $p(\mathrm{x}|\omega_2) \sim N(4,1)$, where the probabilities are equal. You are free to choose any classification method you would like, and you are given an infinitely large dataset. What would be the best error you can achieve on the test set, and why?
d. How would a Gaussian-based classifier act upon an XOR classification problem?
e. For a minimum error rate classification, if the penalties for misclassification are different for the two classes, will it affect the decision boundary? If yes, how? And if no, why?
f. How does a monotonic transformation of x_1 and x_2 change Bayes error rate in a two class two-dimensional classification problem with continuous feature x ?
g. What would happen if for a two known distributions $p(\mathrm{x}|\omega_i)$ and priors $p(\omega_i)$ in a *d*-dimensional feature space, we project the distributions to a lower dimensional space before classifying them? How does it affect the true error? Prove your answer.
h. In a binary classification problem, under what circumstances a Bayes classifier and a minimum distance classifier obtain exactly the same results?
i. Discuss whether it is possible to plot ROC curve for a classification problem with more than two categories? If yes, how? And if no, why?
j. How do you extend Bayes decision rule for a three-class classification problem? Explain and – if necessary – include calculations.

Good Luck!
Ali Abbasi