Assignment 2
Decisions, decisions, decisions!

Please note:

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 November 10th.**
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. Applying Bayesian Inference to Clarify Stupendous 'Monty Hall' Problem (12 Pts.)

Keywords: *Bayesian Inference (Reasoning), Prior Probability, Posterior Probability, Likelihood Function, Monty Hall Problem*

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.

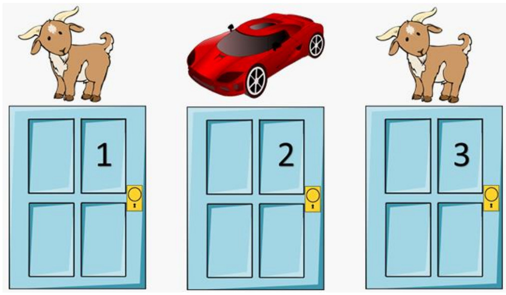

Figure 1 The concept of Monty Hall problem

In this problem, your task is to examine this method on a famous puzzle; the Monty Hall problem. Suppose in a game show there are three doors and you are asked to choose one of them. Behind one door is a car, and behind the others are goats. The host knows what's behind the doors. You choose a door, say door 1. The host will not reveal the car, instead he opens another door, say door 3, which has a goat. He then says to you, "Do you want to switch your selection to door 2?". Quite strangely, you should switch to the other door. If you don't change your choice, you have only a $1/3$ chance of winning, while if you change your mind your chance of winning the car increases to $2/3$. Many people refuse to accept that the switching is beneficial. Even *Paul Erdős*, a well-known mathematician in 20th century, remained unconvinced until he was shown a computer simulation verifying the predicted result.

a. Determine a Bayesian model for this problem. You have to specify the random variables and the input data, as well as the meaning of the prior and the posterior probabilities.
b. Calculate the probability values for the prior.
c. Calculate the probability values for the likelihood.
d. Compute the posterior probability (include intermediate steps).
e. Why is it in your advantage to switch your selection?

Now consider a different scenario. The host doesn't remember what is behind each of the doors, so it cannot be guaranteed that he will not accidentally reveal the car by opening the correct door. The only thing we know is the he won't open the door that you have picked. Therefore, if he accidentally opens the door for the car, you win.

f. How does this twist change the analysis?
g. Is it still in your advantage to switch you choice? Justify your answer by re-calculating your probabilities.

Useful Links: [Monty Hall problem]

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.

Assume X is a two-dimensional feature vector and $p(\mathrm{x}\,|\,\omega_i) \sim N(\mu_i, \Sigma_i)$, $i = 1, 2, \ldots, 7$, as its class conditional distribution.

a. Sketch decision regions of the minimum distance classifier assuming the following class means:

$$\mu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad \mu_3 = \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \quad \mu_4 = \begin{bmatrix} 0 \\ 4 \end{bmatrix}, \quad \mu_5 = \begin{bmatrix} 0 \\ -4 \end{bmatrix}, \quad \mu_6 = \begin{bmatrix} -2 \\ 2 \end{bmatrix}, \quad \mu_7 = \begin{bmatrix} -2 \\ -2 \end{bmatrix}$$

Now let's take a closer look to MDC using MATLAB. Consider the following three-class sets of samples:

$$\omega_1 = \{(3,1), (3,3), (4,1), (4,4), (5,0), (5,1), (5,2), (6,3)\}$$

$$\omega_2 = \{(-1,1), (-1,5), (-2,0), (-2,2), (-3,1), (-3,2), (-5,0), (-5,1)\}$$

$$\omega_3 = \{(-3,0), (-1,-1), (-1,0), (2,-1), (2,0), (3,0), (3,2), (4,-1)\}$$

b. Plot samples belonging to each class, and highlight them using different markers/colors based on their labels.
c. Classify the following points using MDC.

$$X_1 = (-2,1) \qquad X_2 = (4,2) \qquad X_3 = (0,0) \qquad X_4 = (1,-4) \qquad X_5 = (1,3)$$

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

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 a three-category classification problem, where the classes have class-conditional probability distributions $p(\mathrm{x}\,|\,\omega_i) = N(\mu_i, \Sigma_i)$ with

$$\mu_1 = \begin{bmatrix} 4 \\ 4 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} 4 \\ -5 \end{bmatrix} \qquad \mu_3 = \begin{bmatrix} -4 \\ 0 \end{bmatrix}$$

and identical covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma_3 = \sigma^2\mathbf{I}$ where $\sigma^2 = 0.6$. Considering equal priors for all classes,

a. Specify the discriminant functions corresponding to the Bayes decision rule.
b. Determine the decision regions and decision boundaries associated with the Bayes decision rule.
c. Generate 1000 random samples from each of the three class-conditional distributions and plot them in a two-dimensional feature space. Also display the decision boundaries you obtained, on the same plot.

d. Using the generated samples, find the estimates of μ_i and Σ_i, $i = 1,2,3$.
e. Substitute the estimates you found in the previous part to the Bayes decision rule and determine the decision boundaries. Sketch them along with the plots in part c.
f. Now generate a new set of random samples from each class, find the estimates of the parameters, and determine the resulting decision boundaries. Sketch a new plot like the one in part e., which overlays the new decision boundaries on the new generated samples.
g. Compare the results of part f. to the previous case. Are the decision boundaries you obtained by applying different samples the same? Why / why not? Does it match your expectations?

Note: Include all of the details of your calculation in your report.

Recommended MATLAB functions: `mvnrnd(),plot(),fplot()`

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

$$\mu_1 = \begin{bmatrix} -2 \\ -2 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

and

$$\Sigma_1 = \begin{bmatrix} 1 & -0.3 \\ -0.3 & 1 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix}$$

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.

$$\begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}$$

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:

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

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 (10 Pts.)

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

Generally, in decision-theoretic terminology, **Risk** is defined as an expected loss, and $R(\alpha_i | \mathrm{x})$ 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.

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

$$p(x|\omega_1) = k_1 \exp\left(-(\mathrm{x}-4)^2/40\right) \qquad p(x|\omega_2) = k_2 \exp\left(-(\mathrm{x}+6)^2/32\right)$$

a. Find k_1 and k_2.
b. Plot two densities on a single graph and draw the decision boundary.
c. 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).
d. Find and plot the decision regions for Bayes minimum risk.
e. What is the numerical value for the Bayes minimum risk?

7. ROC Curve: A Simple Yet Very Informative Graph for Evaluating a Classifier (8 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 2 illustrates the class-conditional probability distributions of observation x given the two classes of a binary classification problem.

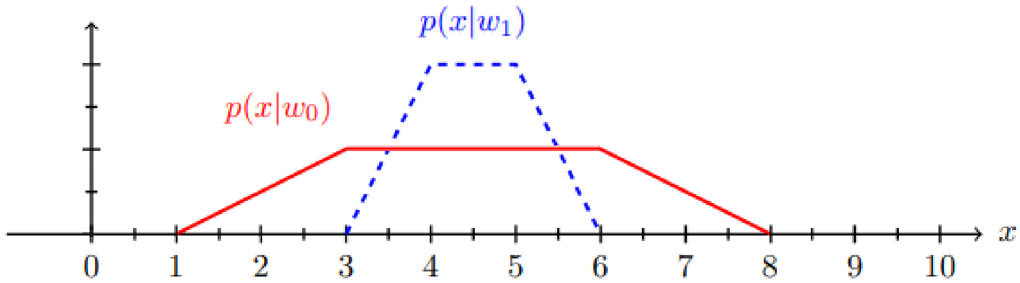

Figure 2 Class-conditional probability distributions corresponding to the two classes in this problem

a. Assuming $p(\omega_0) = 0.3$, 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 Capabilities on a Real-World Problem (15 Pts.)

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

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 based on Bayes decision rule, using Mammographic Mass dataset.

The dataset were extracted from digital mammograms that were taken from 'benign' (class 0) and 'malignant' (class 1) mammographic masses. There are 5 attributes for each sample, which will be mentioned here by their columns numbers (feature 1 to feature 5).

Split the data randomly into training set (75%) and test set (25%). Considering only feature 3, answer the following questions.

a. Find the class priors using the training set. Report the prior probabilities of a mammographic mass being 'benign' or being 'malignant'.
b. Display a feature histogram for each of the two classes, and make sure both plots have the same x-axis range.

c. 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?

d. Display the estimated class-conditional densities in a single diagram. Make a second figure plotting the class posteriors for both classes. Write down the equation you used to compute the class posteriors.

e. Compute the Bayes classification of the training data and calculate the overall training error.

f. Classify the test data and calculate the overall test error.

g. Compute a confusion matrix for this classifier.

h. Calculate the Bayes error.

i. Draw a ROC curve to visualise the performance of the classification.

j. Repeat parts b. to i. for feature 1 and feature 4, and comment on the results. Compare each of the features in terms of classification precision.

Note: The dataset contains missing values, i.e. some samples doesn't have values for some of their features. Please specify the strategy you are employing against them.

Recommended MATLAB functions: `importdata()`, `randsample()`, `randperm()`, `histc()`, `bar()`, `confusionmat()`, `trapz()`

9. Some Explanatory Questions (10 Pts.)

Please answer the following questions as clear as possible:

a. How would a Gaussian-based classifier act upon an XOR classification problem?

b. 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?

c. 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 ?

d. 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.

e. In a binary classification problem, under what circumstances a Bayes classifier and a minimum distance classifier obtain exactly the same results?

f. How would you construct a better classifier if you realize that the FPR is larger than the TPR?

g. How would you construct a better classifier if you realize that the region under the ROC curve is non-convex?
   Hint: A convex region is a set of points where given any two arbitrary points A and B in the set, the line AB joining them also lies entirely within the set.

h. 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?

i. How do you extend Bayes decision rule for a three-class classification problem? Explain and – if necessary – include calculations.

Good Luck!
Ali Abbasi