Amirkabir University of Technology
(Tehran Polytechnic)

Assignment 2
Bayes Theorem: A Simple, Yet Powerful Classification Tool

Homeworks Guidelines and Policies

- ***What you must hand in.*** It is expected that the students submit an assignment report (.pdf) as well as – if necessary – required source codes (.m or .py) into an archive file named according to the following template: HW2_XXXXXXXX.zip
  where XXXXXXXX must be replaced by their student ID.
- ***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 November 20th**.
- ***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 **dalirani@aut.ac.ir**. You may also find us in pattern recognition and image processing lab, 3rd floor, CEIT building.

1. You Think Bayes Theorem is Rudimentary? Well, Think Again (12 Pts.)

Keywords: *Theory of Probability, Conditional Probability, Bayes Theorem, Prior Probability, Posterior Probability, Likelihood Function*

Bayes Theorem is a deceptively simple tool which follows from the axioms of **Conditional Probability**, providing a way to update our beliefs based on the arrival of new, relevant pieces of evidence. More specifically, according to Bayes theorem, our **Posterior** belief $\mathrm{P}(A \mid B)$ is calculated by multiplying our **Prior** belief $\mathrm{P}(A)$ by the **Likelihood** $\mathrm{P}(B \mid A)$ that B will occur if A is true.

The power of Bayes theorem, aka **Bayes' Rule**, is that in many problems where the goal is to compute $\mathrm{P}(A \mid B)$, it turns out that it is difficult to do so directly, yet there might be direct information about $\mathrm{P}(B \mid A)$. It in fact enables us to compute $\mathrm{P}(A \mid B)$ in terms of $\mathrm{P}(B \mid A)$.

Bayes rule is said to be one of the most important theorem in the field of mathematical statistics and probability theory. It is also widely used in the field of pattern recognition. Let's practice some warm-up problems prior to dealing with more practical ones.

In the first scenario, assume that the probability of Donald Trump winning the US presidential election given that Bernie Sanders is running for the election is 0.35, and that the probability of Trump winning the election given that Sanders is not running is 0.65. Also suppose that Bernie Sanders will decide whether or not to run for presidential election based on a coin flip.

Figure 1 Donald Trump and Bernie Sanders. You're going to use Bayes theorem to predict 2020 US presidential election.

a. What is the probability that Bernie Sanders runs?
b. What is the probability that Donald Trump wins?
c. What is the probability that Bernie Sanders ran, given that Donald Trump lost?

Second, suppose you're trying to evaluate the effects of sleep deprivation on academic performance of university students. It is often claimed that enough sleep time helps students get better results. To test this theory, a survey of university students was conducted and the following results were obtained:

- *44% sleep before midnight (which is considered enough).*
- *52% have a GDP more than 16 (which is considered good).*
- *37% sleep after midnight AND have a GDP less than 16.*

d. Use this data to justify the above-mentioned claim.

Next, assume in Amirkabir University, 57% of the students are male. Also, 14.6% of males and 2.4% of females smoke cigars, respectively (disappointing!). One student is randomly picked for a survey.

e. What is the prior probability that the selected student is a female?
f. The student is smoking. Find the probability that the selected student is a female using this additional information.

Finally, let's deal with a major concern in every sport competition, that is, using banned performance-enhancing drugs by athletes. The World Anti-Doping Agency (WADA) conduct routine drug tests on random athletic competitors in order to fight against drugs in sports. However, the usage of other allowed drugs may have cross-reactive results, yielding a high false-positive rate. Assume a false positive rate of 9.8% and a false-negative rate of 1.4%,

g. Find the probability of a positive test given that an athlete has used performance-enhancing drugs.
h. If 2% of athletes in a certain competition have used unallowable drugs, what fraction of this population will test positive?
i. If an athlete tests positive, find the probability that this athlete has used illegal drugs.
j. The World Anti-Doping Agency (WADA) suggests aggressive monitoring so that half of those tested have used illegal drugs. In this case, what is the probability that an athlete testing positive actually has used performance-enhancing drugs?

Figure 2 Iranian weightlifter Saeid Alihosseini received an eight years ban in 2009, while he has always claimed that the conducted test had been a false-positive one

k. So that WADA can decide on the aggressiveness of the monitoring, provide a plot of p, the prior probability that an athlete has used illegal drugs, versus the posterior probability for athletes that test positive for doping.

2. An Ancient Coin Counter System Based on Bayes Decision Rule (10 Pts.)

Keywords: *Classification Problem, Bayes Decision Rule, Image Segmentation, CIELAB Color Space*

Bayes Decision Rule is a wonderfully simple approach to the problem of pattern classification. It assumes the ideal case in which the probability structure underlying each class is known perfectly. While this assumption rarely occurs in practice, it allows us to determine the optimal classifier against which we can compare all other classifiers.

Now, let us consider a fiction story. You're living in 550 BC, where Cyrus the Great rules the Persian Kingdom. You – as an AI expert – were asked to implement a system capable of counting royal coins by their images. Perhaps a good way to do so is by **Image Segmentation**, where you can classify the input image into certain colors which represent the coin types. There are three types of coins, 'Silver', 'Bronze' and 'Gold'. You are only allowed to use the given dataset in the table below:

Pixel	Label	R	G	B	L	a	b
1	Silver	73	72	68	31	0	3
2	Silver	123	123	121	52	0	1
3	Silver	43	44	39	18	-1	3
4	Silver	46	42	41	17	2	1
5	Silver	240	239	237	94	0	1
6	Bronze	207	143	69	65	20	48
7	Bronze	254	242	168	95	-4	37
8	Bronze	130	52	3	33	33	43
9	Bronze	253	221	146	90	4	41
10	Bronze	254	211	143	87	9	40
11	Gold	200	170	46	71	2	63
12	Gold	122	86	0	40	11	47
13	Gold	155	129	36	55	3	51
14	Gold	221	192	64	79	1	65
15	Gold	90	61	19	29	10	30

Figure 3 Three different coins form three different classes; 'Silver', 'Bronze' and 'Gold', each with specific RBG and Lab values

The dataset contains 15 pixels and their color values in RGB and CIELAB color spaces, randomly picked from each of the three coins regions in Figure 3. Suppose that you are only allowed to use two features.

a. By visual inspection using 2D feature space, evaluate which two features 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 $\sum = \mathbf{I}_2$. Find the general form of the discriminant function.
c. Classify the following pixels using the functions you obtained in the previous part.

Pixel	R	G	B	L	a	b
1	243	204	87	84	5	62
2	242	225	156	90	-1	36
3	24	23	19	8	0	3
4	140	137	132	57	0	3
5	181	149	62	63	5	49

d. Express some of the challenges this system faces.

Note: You are allowed to use calculator or any other calculation tools such as MATLAB or Python. However, you are not allowed to solve this problem by mere 'programming'.

3. Calculating Prediction Errors in Bayes Decision Rule (8 Pts.)

Keywords: *Bayes Decision Rule, Probability of Error, Upper Bounds of Error Probability, Bhattacharyya Error Bound, Chernoff Error Bound, Neyman-Pearson Test, Minimax Criterion*

Bayes Decision Rule is the best classifier which minimises the **Probability of Error**, that is the probability that a sample is assigned to a wrong class. But like any other classification model, it does not lead to perfect classification. Calculating the probability of error helps us to evaluate the performance of this decision rule.

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**.

In this problem, we are going to study Bayes error and some error bounds, as well as **Neyman-Pearson** and **Minimax** classifiers. Assume density functions given in Figure 4. Considering $p(\omega_0) = 0.4$,

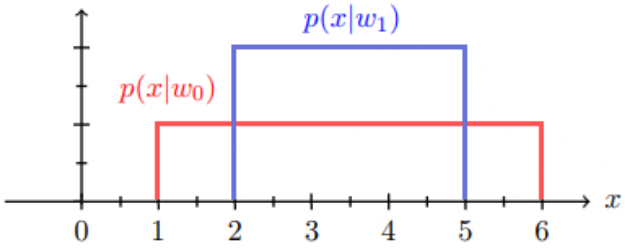

Figure 3 Class-conditional probability distributions

a. Determine and sketch the decision regions for the minimum-error decision rule.
b. Determine the Neyman-Pearson classifier assuming $\varepsilon_1 = 0.01$. Calculate the error for this classifier.
c. Determine the minimax classifier and find the corresponding error.
d. Find the Bhattacharyya error bound and compare it with the Bayes error.

4. An Attempt of Optical Character Recognition by Minimum Distance Classifier (15 Pts.)

Keywords: *Classification Problem, Minimum Distance Classifier, Optical Character Recognition*

A **Minimum Distance Classifier** attempts to classify an unlabelled sample to a class which minimise the distance between the sample and the class in multi-feature space. As minimising distance is a measure for maximising similarity, **MDC** actually assigns data to its most similar category.

While **MDC** might look too basic, it works pretty well in some problems. One of them could be **Optical Character Recognition** (**OCR**), where the goal is to distinguish handwritten or printed text characters inside digital images of documents. You are working with a customised dataset here, which is available at 'P4' folder in the 'inputs' directory. The dataset is divided into three groups, which can be seen in Figure 5.

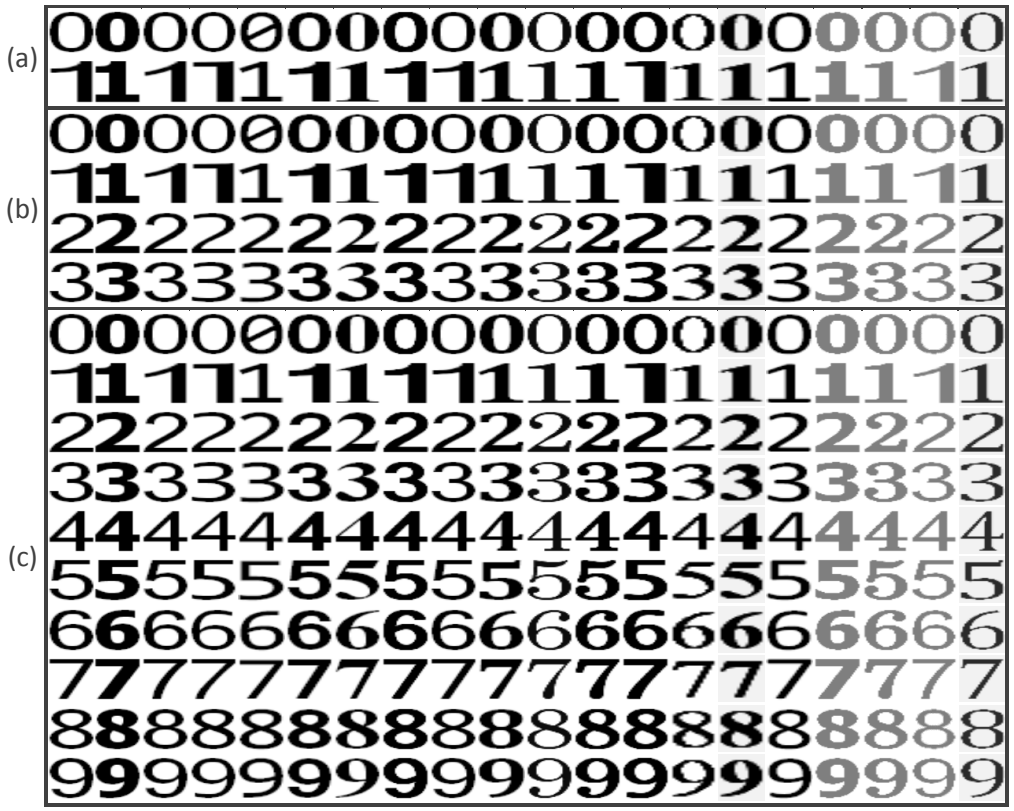

Figure 4 The problem is divided into three different parts with three different datasets (a) first dataset, including only two digits with structurally distinctive appearance (b) second dataset, which leads to a more complicated task with four categories (c) last dataset, including all digits

First, consider group (a).

a. Using the train images, find and display the prototype of each of the available classes.
b. Use test samples to evaluate your MDC classifier. Report the error.
c. Repeat part a. and b. with group (b).
d. Repeat part a. and b. with group (c).
e. Comment on the results obtained from the previous parts. What were your observations?

5. Beyond Bayes Decision Rule: When Decision Costs Matter (12 Pts.)

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

Bayes Decision Rule is in fact a classifier which recommend decisions that minimise some total expected **Risk**. The simplest risk is when the **Decision Costs** are equal, which leads to **Bayes Error**. However, this is not always the case in many classification problems, e.g. we prefer to detect leukemia in a healthy person rather than misclassifying a leukemia case as a no risk state. Luckily, Bayes decision rule is also capable of considering decision costs and set the decision boundary accordingly.

Assume a mysterious application, called "Lie 'n' Die!", has been launched which claims it could detect whether a person lies (ω_1) or not (ω_2). By using only two secret features obtained from each individual's face, the distribution of the application training data can be well represented by two Gaussians, as follows:

$$p(\mathrm{x} \mid \omega_1) \sim N(\mu_1, \Sigma_1),\ p(\mathrm{x} \mid \omega_2) \sim N(\mu_2, \Sigma_2)$$

where $\mu_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$, $\mu_2 = \begin{bmatrix} 5 & 0 \end{bmatrix}^T$, $\Sigma_1 = \mathrm{I}$ and $\Sigma_2 = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$.

a. Sketch the contours of constant values for two class conditional densities.
b. Lie 'n' Die! developers think a threshold at $x_1 = 3$ could very well separate their training examples. Design a modified classifier which minimises the Bayes probability of error, and give them an expression in terms of x to prove them they can do better. Assume that in the society, lies are three times as likely as honesty (how sad!).
c. Determine the shape of the optimal decision boundary you found in the previous part. Justify your answer.
d. The developers think it would be unpleasant if the application detects an honest person as a liar. They estimate this misclassification will cost them 5 times as much as the cost of declaring a liar as an honest person. Explain qualitatively how this new rule changes the result you obtained in part b. Plot a sketch of the change.
e. Calculate the Bhattacharya error bound.
f. Calculate the Chernoff error bound.

Note: You are required to plot the desired figures by hand, not using MATLAB or Python.

6. Studying Discriminability by ROC Curves (10 Pts.)

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

ROC Curve or **Receiver Operating Characteristic Curve**, is an incredibly useful tool in evaluating and comparing classification model's performance. With ROC curve, it would be possible to see how any predictive model can distinguish between the **True Positives** and **True Negatives**. To accomplish this task, a model must not only predict a positive as a positive, but also a negative as a negative. ROC is also written as **AUROC**, or **Area Under the Receiver Operating Characteristics**.

In order to get more familiar with it, here you are provided with four separate datasets available at 'P6' directory in the 'inputs' folder. These datasets contain 1000 1-D samples from each of the two available classes c_1 (first column) and c_2 (second column).

a. Compute the discriminability measure for each dataset, defined as below:

$$d' = \frac{|\mu_2 - \mu_1|}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

where μ_i and σ_i are the mean and standard deviation of the distribution of class c_i.

b. Compute and plot the ROC curve for each dataset according to the following approximations:

$$\mathrm{P_{TP}} = \mathrm{P}(x > x^* \mid x \in c_2) \approx \mathrm{N}(x > x^* \mid x \in c_2)/n$$

$$\mathrm{P_{FP}} = \mathrm{P}(x > x^* \mid x \in c_1) \approx \mathrm{N}(x > x^* \mid x \in c_1)/n$$

where n is the number of samples, and N denotes a count function.
Note: In order to make comparisons, plot all four ROC curves in the same figure.

c. Plot the two approximated probability density function for each dataset.
Hint: Use the same method you used in the previous part to get the cumulative distribution, known as **Monte Carlo** method. Note that the probability density function is the derivative of the cumulative density function.

d. Based on your observations, explain how the ROC curve and discriminability relate to each other.

7. Face Detection Problem: How Does Bayesian Decision Theory Handle It? (20 Pts.)

Keywords: *Classification Problem, Bayes Decision Rule, Confusion Matrix, Bayes Error, ROC Curve, Face Detection, RG Chromaticity*

So far you've probably got familiar enough with **Bayesian Decision Theory** to know what to expect from it. If the decision problem is posed in probabilistic terms and all relevant probability values are known, **BDT** allows to take optimal decisions that minimise errors by choosing the least risky class. Although in many practical classification problems, these conditions are not fulfilled and therefore **BDT** won't be effective, there are still some applications where it may come in handy.

Face Detection (not to be mistaken with **Face Recognition**) is the process of identifying and locating human faces in digital images and videos. It is often the first step in many face-related machine vision applications such as face recognition, emotion detection, gender detection, etc. Now we are going to find out how **BDT** deals with this problem in practice.

You are given a customised dataset divided into 'Train' and 'Test' folders at 'P7' directory. The train set consists of 50 face images alongside the corresponding face masks in a separate folder. These masks indicate face pixels with black (or gray-level 0) and non-face pixels with white (or gray-level 255).

a. First assume images in the 'Train' directory. Considering two classes for each pixel, 'face' and 'non-face', use the provided masks to find the class priors.

b. We want to model the class-conditional probability density of each class using a univariate Gaussian. Find the mean and variance of both class-conditional densities.

c. Apply your classifier on the images provided in the 'Test' directory and display the results. Also report the test error using the given masks.

d. Compute a confusion matrix for your classifier.
e. Calculate the Bayes error.
f. Draw a ROC curve to visualise the performance of your classifier.
g. Comment on the above results. In what circumstances has your classifier failed, and why? What do you think the advantages and disadvantages of such a Bayesian face detector are? Suggest at least two modifications to improve it.

(a)

(b)

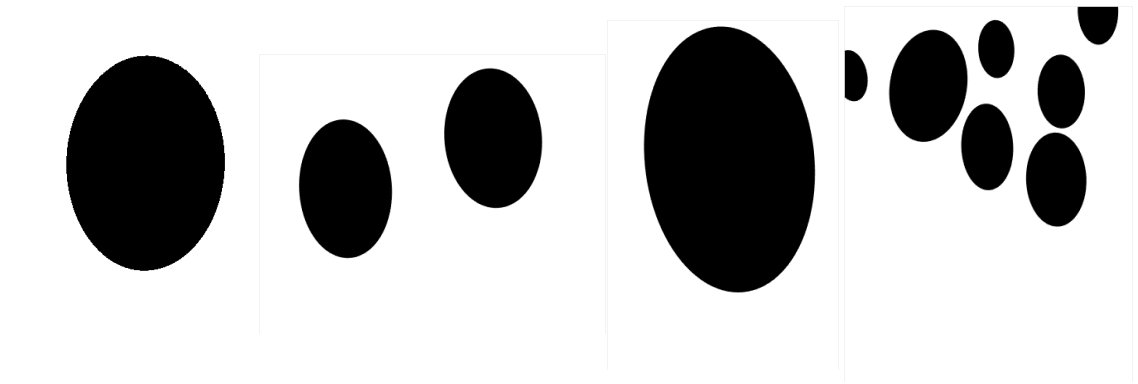

Figure 6 Samples in the given dataset are 'unconstrained', i.e. there is no limitation on the number of faces in an image and their illumination, pose, occlusion and facial expressions (a) original images with the location of face(s) highlighted (b) corresponding masks, which can be used to specify the pixels inside face regions

Hint: RGB is not the best color representation for characterising skin-color, because it represents not only color but also brightness. Therefore, represent skin-color in the chromatic space which is defined as follows:

$$r = \frac{R}{(R+G+B)}, \quad g = \frac{G}{(R+G+B)}, \quad b = \frac{B}{(R+G+B)}$$

8. Some Explanatory Questions (8 Pts.)

Please answer the following questions as clear as possible:

a. Is it possible to apply the Bayesian Decision Rule in a regression problem? If yes, explain how. If no, explain why.
b. A and B are conditionally independent given C, if $\mathrm{P}(AB \mid C) = \mathrm{P}(A \mid C)\mathrm{P}(B \mid C)$. Justify whether or not if A and B are independent, then they're also conditionally independent given any C. Either prove it or give a counterexample.

c. Bayes decision rule is said to be the best decision rule, giving the minimum probability of misclassification. However, according to **No Free Lunch** theorem; "there is no one model that works best for every problem". How can you justify that?
d. How does selecting different distance functions affect MDC classification result? Support your answer with simple examples in 2D feature space.
e. Is the result of the Bayes decision rule unique? Explain.
f. How do you explain a Bayes classifier's training phase? What about a MDC classifier?

Good Luck!
Ali Abbasi, Farhad Dalirani