Amirkabir University of Technology
(Tehran Polytechnic)

Assignment 2
Bayes Theorem: A Simple, Yet Powerful Technique

Homeworks Guidelines and Policies

- ***What you must hand in.*** It is expected that the students submit an assignment report (HW2_[student_id].pdf) as well as required source codes (.m or .py) into an archive file (HW2_[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 25th**.
- ***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 me through **ali.the.special@gmail.com**.

1. You Can Win Lamborghini When You Know Bayes Theorem! (12 Pts.)

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

Suppose you are in a game show in which there are three doors and your task is to choose one of them. Behind one door is a Lamborghini, and behind the others are goats. The host knows what is behind the doors. You choose a door, say door 1. The host will not reveal the car, instead he opens another door, say door 2, which has a goat. He then says to you, "Do you want to switch your selection to door 3?". 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.

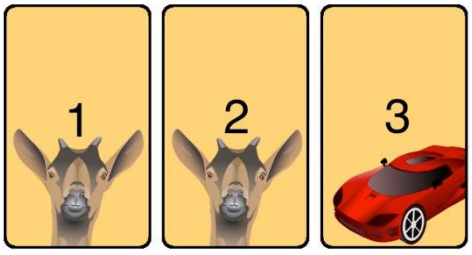


Figure 1 The concept of Monty Hall problem. There are three doors among which only one opens to the grand prize, which is a Lamborghini!

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 Link: Monty Hall problem

2. Optical Character Recognition Through 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 in the 'P2' folder of the 'inputs' directory. The dataset is divided into three groups, as depicted in Figure 2.

(a)

(b)

(c)

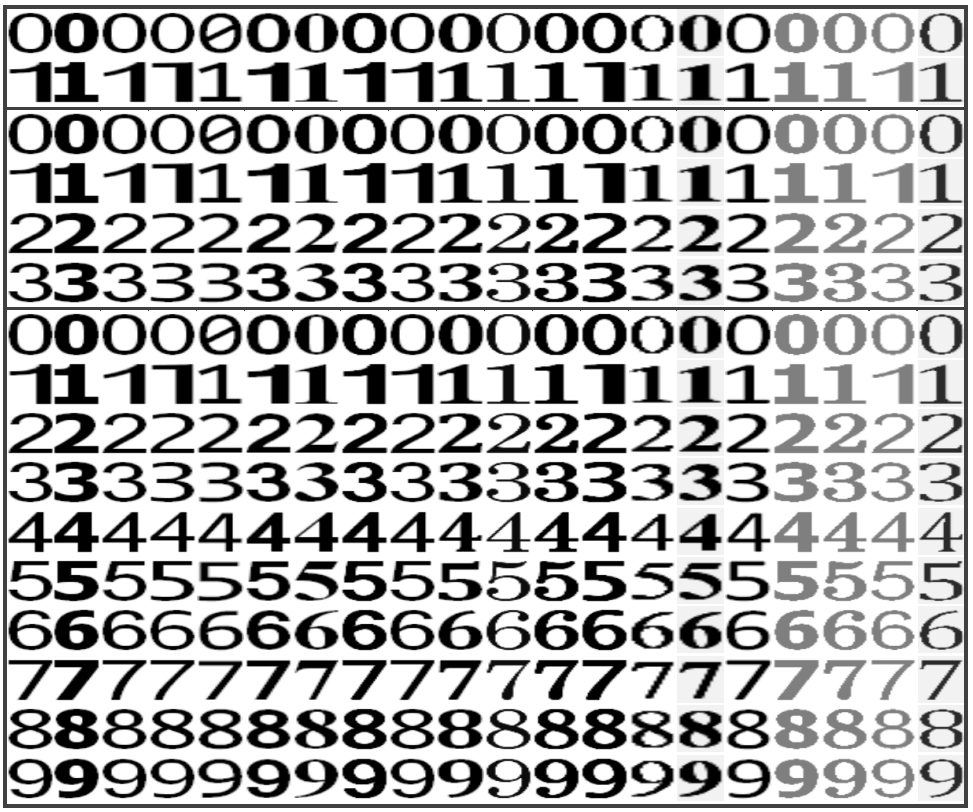


Figure 2 The problem is divided into three different parts with three different datasets (a) first dataset, including only two digits with structurally distinctive appearance (b) second dataset, which leads to a more complicated task with four categories (c) last dataset, including all digits.

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?

3. Parameter Estimation Can Save Lives! (12 Pts.)

Keywords: *Parameter Estimation, Maximum Likelihood Estimation, Bayes Estimation, Biased Estimator, Estimator Variance, Mean Squared Error (MSE)*

Silicon dioxide, aka *silica*, is an oxide of silicon which is used in many applications, from structural materials and microelectronics to pharmaceutical industries. However, inhaling silica dust may lead to serious diseases such as bronchitis and lung cancer. The European Environment Agency (EEA) conducts occasional reviews on its standards for airborne

(a) **(b)** **(c)**

Figure 3 When people breathe silica dust, they inhale tiny little particles of the mineral silica, which can cause scar tissue in the lungs and even lead to lung cancer. (a) Silica dust released after cutting through rocks. (b) Silicosis. (c) Comparison of a lung with Silicosis (left) with a healthy one (right).

silica. During a review, the EEA investigates data from several studies. Each study takes into account different groups of people, and different groups have different exposures to silica.

Let s be the number studies, n_i be the number of people in the i-th study, x_i be the silica exposure for people in that study, and y_i be the number of people who developed lung cancer in that study. The EEA's model is $Y_i \sim \text{Poisson}(\lambda_i)$, where $\lambda_i = n_i x_i \lambda$ and λ is the typical rate at which silica causes cancer. The n_i's and x_i's are known constants, yet the Y_i's are random variables. Since different studies involve different groups of people in different places, they model the Y_i's from different studies as being independent, but not identically distributed since the λ_i's are different. The EEA goal is to estimate λ.

a. Write down the PDF of the joint distribution of $Y_1,\ldots,Y_s \mid \lambda$, which will also involve the constants x_1, $n_1,\ldots$, x_s, n_s.
b. Calculate the maximum likelihood estimator of λ.
c. Justify whether the maximum likelihood estimator is an unbiased estimator of λ.
d. Calculate the variance of the maximum likelihood estimator.
e. Calculate the mean squared error of the maximum likelihood estimator.
f. Assume the EEA attempts to estimate λ by using this model and combining data from 3 studies with data recorded in the table below. Write down an expression for the maximum likelihood estimate of λ.
   Note: Your answer should only involve numbers, not symbols. However, you don't need to simplify your expression.

Study Number (i)	Sample Size (n_i)	Exposure Level (x_i)	Cancer Case Count (y_i)
1	20	0.4	2
2	50	0.3	5
3	100	0.6	14

g. Suppose the EEA analysts decide to adopt a prior of $\Lambda \sim \text{Gamma}(\alpha,\beta)$, where α and β are known constant they choose to reflect their prior knowledge about λ. Find the posterior distribution for Λ. You should arrive at specific form for the posterior distribution, with parameters involving α, β, $x_1,\ldots,x_s$, $n_1,\ldots,n_s$, and $y_1,\ldots,y_s$.

4. Dealing with a Basic Parameter Estimation Problem in MATLAB (14 Pts.)

Keywords: *Parameter Estimation, Maximum Likelihood Estimation, Maximum A Posteriori (MAP) Estimation, Posterior Distribution, Prior Distribution*

After practicing some problems related to **Parameter Estimation** by applying both **ML** and **MAP** estimation methods, it's time to deal with more realistic problems. The following list of 20 numbers were generated by sampling a binomial distribution with unknown number of trials N, and probability of success p equals to 0.2 for each trial.

$$\{0,2,0,3,1,2,2,2,0,2,3,1,0,0,2,3,3,2,3,0\}$$

a. Plot a bar graph which indicates the values in the list and their number of occurrences.
b. Normalise the plot in part a. and plot the empirical probability mass function.
c. Plot the probability mass function of a binomial distribution corresponding to the following parameters:

$N = 5, p = 0.2$ $N = 10, p = 0.3$ $N = 20, p = 0.1$

d. Which one of the probability mass functions in part c. is more likely to have generated the given list of 20 numbers?
e. Plot the likelihood function of the given numbers as a function of the parameter N.
Note: Your graph must have integer numbers from 1 to 20 as its x-axis, and $p(x|N)$ as its y-axis.
f. Plot the log likelihood function.
g. Determine the value of N that maximises both likelihood and log likelihood.

Now, you are given a dataset containing 5000 random samples generated from Kumaraswamy distribution, which is defined as below:

$$f_X(x) = abx^{a-1}(1-x^a)^{b-1}$$

h. Plot the histogram of the given samples.
i. Write down the likelihood and log-likelihood functions.
j. Estimate a and b by approximately maximizing log-likelihood function over the range [0,12].
k. Calculate the exact values of a and b using an arbitrary optimization approach from a built-in function or an external library.
l. Plot and compare the estimated functions in part (j) and (k) with the histogram in part (h).

Recommended MATLAB functions: `bar()`, `binornd()`, `binopdf()`

5. Bayesian-Based Alcohol Addiction Detection (14 Pts.)

Keywords: *Classification Problem, Bayes Decision Rule, Parameter Estimation, Maximum Likelihood Estimation, Maximum A Posteriori Estimation*

Alcohol is among the most prevalently used addictive substances in the world. In the US, one out of every 12 adults suffers from an alcohol abuse or dependency issue. Distinguishing among regular alcohol drinkers and addicted ones is often a difficult task, however, there are methods by which alcohol addiction can be detected with reasonable accuracy. In this problem, we are going to investigate one of them.

Figure 4 Alcohol addiction test can be performed through various approaches, among which blood test is known to be the most accurate.

Given along with this homework is a dataset in which the first five features denote the indicators obtained from blood tests, whereas the parameter *drinks* indicates daily alcohol consumption for each individual. Participants with daily consumption over 5 are considered as *addicted*. Also, the *selector* parameter is used to construct the training (1) and test (2) sets.

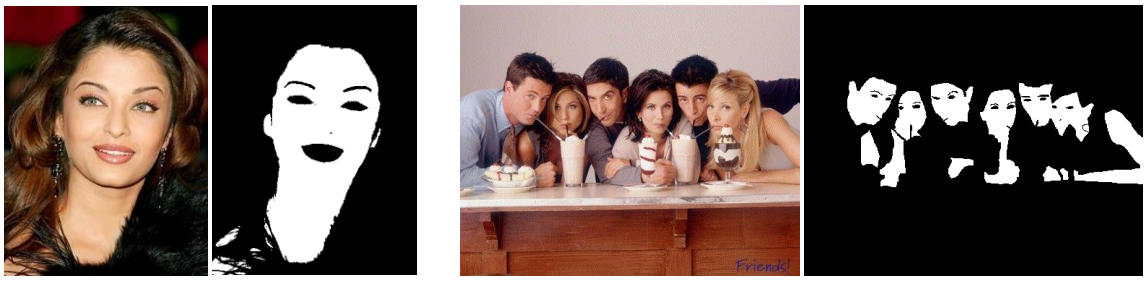
a. Assuming normal distribution for both classes, randomly pick 10, 50, and 100 samples from test set and estimate the distribution parameters using maximum likelihood estimation.
b. Design a classifier for each cases in the previous problem, and find errors using the test set.
c. Repeat the previous parts using Bayesian estimation method, and compare the results.
d. Consider the 100-sample case with maximum likelihood estimation approach. Use trial and error to find two parameters which yield highest classification accuracy, and discard the remaining. Plot the class distribution and Bayes decision boundary for this parameter.
e. Plot the Bayes decision boundary, this time considering the error costs $\lambda_{12} = 2$ and $\lambda_{21} = 3$.

6. Evaluation of Bayes Decision Rule in Skin Detection Problem (20+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.

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?

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

c. Classify the pixels in image 'trump_tie_1.jpg' and 'trump_tie_2.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 all training images ('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()`

7. Some Explanatory Questions (8 Pts.)

Please answer the following questions as clear as possible:

a. 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?
b. Is it possible to apply the Bayesian Decision Rule in a regression problem? If yes, explain how. If no, explain why
c. How does selecting different distance functions affect MDC classification result? Support your answer with simple examples in 2D feature space.
d. How do you explain a Bayes classifier's training phase? What about a MDC classifier?
e. When does MLE estimation lead to a better result than MAP estimation?

Good Luck!
Ali Abbasi