جامعة خليفة
Khalifa University

College of Engineering

ECCE732: Machine Learning and Applications

Fall 2022

Instructor: Dr Panos Liatsis

PROBLEM #	SCORE	MAXIMUM
Q1		10
Q2		30
Q3		30
Q4		30
Extra Credit*		20
TOTAL		100

Homework 1

Due: 26/09/2018, 11pm

Homework 1: Bayes Theorem and Classification

Q1. Smith owns a retail store for selling phones. The phones are manufactured at three different factories: A, B, C. Factories A, B, and C respectively produce 20%, 30%, and 50% of the phones being sold at Smith's store. The probabilities of defective phones from factories A, B, and C are 2%, 1%, and 0.05%, respectively. The total number of phones being sold at Smith's store is 10000. One day, a customer walks up to Smith's store, and ask for a refund for a defective phone. You are required to calculate the following:

(a) What is the probability of a phone being defective?

(1 points)

(b) What is the probability that this defective phone is manufactured at factory A?

(3 points)

(c) What is the probability that this defective phone is manufactured at factory B?

(3 points)

(d) What is the probability that this defective phone is manufactured at factory C?

(3 points)

Q2. Assume the table below with nine samples, characterized by three features, belonging to two classes (positive or negative). Features a_1 and a_2 are categorical (true/false), while feature a_3 is numerical.

Sample	a_1	a_2	a_3	Class label
1	T	T	5.0	P
2	T	T	7.0	P
3	T	F	8.0	N
4	F	F	3.0	P
5	F	T	7.0	N
6	F	T	4.0	N
7	F	F	5.0	N
8	T	F	6.0	P
9	F	T	1.0	N

Note that for numerical features, the likelihood for each class can be calculated through:

$$P(x_j|w_i) \propto f(x_j|\mu_i, \sigma_{ij}^2) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} exp\left\{-\frac{(x_j-\mu_{ij})^2}{2\sigma_{ij}^2}\right\}$$

Assume a test point (T, F, 2.0) and the naïve Bayes classifier. Calculate and show the workings for all of the associated:

(i) Prior probabilities,

(3 points)

(ii) Likelihoods, and

(10 points)

(iii) Posterior probabilities.

(4 points)

(iv) Classify the test point (T,F, 2.0), using the naïve Bayes classifier. Which class label should be assigned to it and why?

(3 points)

Hint: In the case of categorical features, you need to calculate the associated frequency tables for each of the features.

Q3. The class-conditional density functions of a discrete random variable ***x*** for four pattern classes are shown below:

x	$p(\boldsymbol{x}\|c_1)$	$p(\boldsymbol{x}\|c_2)$	$p(\boldsymbol{x}\|c_3)$	$p(\boldsymbol{x}\|c_4)$
1	1/3	2/3	1/4	3/5
2	2/3	1/3	3/4	2/5

The loss function is as follows, where action α_i means "decide pattern class c_i".

	c_1	c_2	c_3	c_4
α_1	0	2	3	4
α_2	1	0	1	8
α_3	3	2	0	2
α_4	5	3	1	0

You are required to calculate the following:

(a) A decision rule *a(**x**)* lets us decide which action to take for any observation. Build a table stating all possible decision rules. For example, one possible decision rules is:

If *x = 1*, Take action α_1
If *x = 2*, Take action α_2

(6 points)

(b) Now, calculate the risk function $R(a_i/\boldsymbol{x})$ for all the decision rules, where:

$$R(a_i|\boldsymbol{x}) = \sum_{j=1}^{c} L(a_i/c_j) P(c_j/\boldsymbol{x})$$

Can we devise a uniformly best rule?

(10 points)

(c) Compute the *Bayes* risk for the following class prior probabilities: $P(c_1) = 1/4$, $P(c_2) = 1/4$, $P(c_3) = 3/8$, $P(c_4) = 1/8$,
Using the formula,

$$R^* = \sum_x R(a_i|\boldsymbol{x})p(\boldsymbol{x})$$

Use equal probabilities p(**x**=1)=p(**x**=2)=0.5.
Note: The ***Bayes*** risk can be calculated using the above formula by calculating the conditional risk for all possible actions and selecting the action a_i for which $R(a_i|\boldsymbol{x})$ is minimum.

(14 points)

Q4. You are required to write a MATLAB/Python script to use the MATLAB Naïve Bayes classifier functions in the classification of the Iris dataset. Please note that this data is available in MATLAB as fisheriris. Alternatively, they can be downloaded from the UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/iris. Ensure that the script is appropriately commented to ensure program readability. The Iris dataset involves the classification of three different species of Iris, i.e., Iris Setosa, Iris Versicolor and Iris Virginica, based on ***4 features***, i.e., petal and sepal length/width. To evaluate the performance of the Naïve Bayes classifier, you will need to ***implement*** (i.e., write a MATLAB/Python script, ***rather than use the existing MATLAB/Python function***) the process of ***k-fold cross validation***. Cross-validation is used in evaluating the performance of a machine learning model on unseen data. A limited sample of the data is used to estimate how a classifier performs (i.e., generalizes), when used on data not previously seen in the training of the classifier. During k-fold cross-validation, the dataset is randomly shuffled and split into k groups. Next, k-1 groups are used for training and the remaining one group is used for testing. The performance of the classifier in terms of training and testing error is recorded and then the classifier model is discarded. The process is repeated k times until each of the groups is independently used for testing purposes. In this application, you are required to implement 10-fold cross-validation. This translates to splitting the Iris dataset into 10 random groups of patterns. You will use 9 of these groups, e.g., 1, 2, 3, 4, 5, 6, 7, 8, and 9 for training and group 10 for testing. In the next training iteration, you will use another set of 9 groups, e.g., 1, 2, 3, 4, 5, 6, 7, 8, and 10 for training, and group 9 for testing, and so on. For this assignment, you should provide:

(a) the MATLAB/Python code for the Bayes classifier and 10-fold cross-validation

(20 points)

(b) a table reporting the performance of the Naïve Bayes classifier for each of the 10 cross-validation rounds and the mean performance in terms of training and testing error.

(10 points)

Extra Credit (20 points)

Consider the pattern distribution of Table 1 corresponding to two classes:

Class 1	(1,5,4)	(1,6,6)	(2,4,5)	(2,5,3)	(2,6,4)	(3,4,5)	(3,5,7)	(4,5,6)
Class 2	(1,1,-1)	(1,2,1)	(2,3,0)	(3,1,-1)	(3,2,2)	(4,3,1)		

Table 1: Two class pattern distribution

You are required to develop a classifier based on discriminant functions as follows:

(i) Write a MATLAB/Python script to implement the application of discriminant functions given by:

$$g_i(\boldsymbol{x}) = lnp(\boldsymbol{x}/w_i) + \ln P(w_i)$$

where $p(\boldsymbol{x}/w_i)$ follows a multivariate Gaussian density:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{1/2}} exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

The corresponding discriminant function for each class is given by:

$$g_i(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{W}_i\boldsymbol{x} + \boldsymbol{w}_i^T\boldsymbol{x} + w_{i,0}$$

Ensure that the script is appropriately commented to ensure program readability. The script should calculate the following parameters:

(a) The weight of the quadratic term

$$\boldsymbol{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$$

(4 points)

(b) The weight of the first-order term

$$\boldsymbol{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

(4 points)

(c) The weight of the constant term (i.e., bias)

$$w_{i,0} = -\frac{\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i}{2} - ln|\Sigma_i| + \ln P(w_i)$$

(4 points)

(ii) Provide the equations for the discriminant functions $g_1(\boldsymbol{x})$ and $g_2(\boldsymbol{x})$ for each of the classes.

(4 points)

(iii) Provide the equation of the resulting decision boundary.

(4 points)