Lahore University of Management Sciences

CS 535 - Machine Learning

Assignment 3

Due Date : October 24, 2017 (11:55 p.m)

Instructions

- You are **NOT** allowed to use in-built MATLAB functions.
- You need to upload the scanned copy of your assignment and script files on LMS before the deadline and hand over the hard copy(if its hand-written) on Wednesday, 25^{rd} October after the lecture.
- Write neatly and clearly highlight the final answers. Upload *ONE* zipped folder containing the assignment and MATLAB files.
- Question 1, 2 (Part b) and 3 require you to write a MATLAB script. You should name your script files as 'Question Number-Part-Roll Number.mat'.
- These assignments are to be done individually. Any plagiarism found would be reported to disciplinary committee.
- You'll need to use Poisson Distribution in Question 5. Poisson Random Variable is a discrete variable whose pdf is given by,

$$P_X(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, 2, ...$$

- The delta function $\delta(x)$ is defined as function that takes a value of '1' when $x = 0$ and '0' otherwise,

$$\delta(x) = \begin{cases} 1 & if\ x = 0 \\ 0 & otherwise \end{cases}$$

Problem 1. [10 Points]
The data for a 3-class classification problem is provided in $'Q1Data.mat'$. There are 4 attributes x_1, x_2, x_3, x_4 and 30 instances. The order of data is described in the table below,

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Instance No.	$x_1 \in \{1,2,3\}$	$x_2 \in \{1,2\}$	$x_3 \in \{1,2\}$	$x_4 \in \{1,2\}$	Class Label $Y \in \{1,2,3\}$

Prior to the experiment it is known that Y follows the following distribution,

$$P_Y(y) = \frac{3}{10}\delta(y-3) + \frac{1}{10}\delta(y-2) + \frac{6}{10}\delta(y-1)$$

Now you are provided a new unlabeled instance with attribute values $x_1 = x_2 = x_3 = x_4 = 2$. Write a MATLAB script to compute the likelihood and MAP estimate of the class label Y for this new instance using the prior information and data provided.
Your code should output the likelihood probability and estimated class of new instance. Name your script file as $'Q1 - RollNumber.m'$

Problem 2. In this question, you need to estimate the density parameters of a normal distribution using the data provided.

Part (a) [15 Points]
If $\vec{x}$ represents data drawn from a normal distribution. *Derive* the maximum likelihood estimate of mean and variance of the distribution.

Part (b) [10 Points]
Write a MATLAB script to implement the above results for parameter estimation of normal random variable. Test your script on data provided and report your values of mean and variance.

- If your Roll Number is odd, use *Q2odd.mat*
- If your Roll Number is even, use *Q2even.mat*

Problem 3. Consider a single attribute binary-class problem($C_1 = 1$, $C_2 = 0$). It is known that the likelihood function follows a Normal distribution with different parameters for different class label. Since, the parameters are unknown, we'll try to learn them.

Part(a) [7 Points]
Using the data set given in *Q3Learn.mat*, where first column represents attribute and second column represents class label, write a MATLAB script to learn prior and likelihood probabilities. Your script should also test the accuracy of estimates using *Q3Test.mat*. Your code should output prior probability of C_1 and C_2, means and variances of likelihood probability and SSE on test set.

Part (b) [13 Points]
In this part we observe the effect of sample size on test error.
If $ss = [15000, 20000, 25000, 28000, 3000, 35000, 40000]$ represent different sample sizes, you should write a code that *randomly* selects data for each sample size, computes test error, repeats the process 500 times and reports *average* test error. Your script should output the plot of Average Test error (on y-axis) vs. Sample Size(on x-axis).
What conclusion can you draw from your plot?

Problem 4. Suppose that an experiment is performed on three courses being taught at SBASSE which are Machine Learning(ML), Probability Theory(PT) and Linear Systems(LS). The class of ML contains 30 MS students, 40 undergraduate students(U), and 10 PhD students, class of PT contains 10 MS , 10 U, and 0 PhD students, and class of LS contains 30 MS, 30 U, and 4 PhD students.

Part (a) [10 Points]
If a class is chosen at random with probabilities p(ML) = 0.2, p(PT) = 0.2, p(LS) = 0.6, and a student is chosen from the class (with equal probability of selecting any of the students in the class and each student is enrolled in only one course), then what is the probability of selecting an MS student?

Part (b) [10 Points]
If we observe that the selected student is in fact an undegrad, what is the probability that he is taking LS?

Problem 5. You have started using the app *Sarahah* where you get messages from people with their identities hidden. Since, you observed that once you post anything on Facebook, you'll definitely receive some messages on the app which could be either from your friends(F) or from anonymous people(A), you want to apply probability theory to decide if the message was from your friend or someone anonymous. For simplicity, you assume that you post only once in a day and mark the time of post as $t = 0$. Luckily, you get to meet a domain expert who is working on a similar problem and he provides you the following information,

- The number of messages you receive from your friends in a day (after posting on Facebook) follow a Poisson distribution with $\lambda = a$. The number of messages from anonymous users also follows a Poisson distribution but with $\lambda = b$.
- The arrival time t (number of seconds after you posted at $t = 0$) of each message follow an exponential distribution. The means of the distribution for messages from your friends and for anonymous users is c and d respectively.
- Each message has 25% chance of being from your friend. If it's not from your friend then it is labeled as being from anonymous user.

Part (a) [10 Points]
What is the distribution of total number of messages received in a day after posting on Facebook?

Part (b) [5 Points]
If N messages are received from anonymous users in a day, what is the probability that they were received in the time interval interval $[t_1, t_2]$, i.e received t_1 seconds after posting but within the first t_2 seconds?

Part (c) [15 Points] If you receive 10 messages in a day and all of them were received during the interval $\tau = [t_1, t_2]$. You applied machine learning techniques for text analysis and figured out that all those messages were from the same user. Using MAP estimation, figure out if the messages were sent by your friend or some anonymous user.
Note: Consult Table 1 for values of a,b,c,d,$[t_1, t_2]$ according to your roll number.

Table 1: Variable Values

Parameters	Odd Roll Number	Even Roll Number
a	5	7
b	1	2
c	2	4
d	10	20
t_1	5	8
t_2	10	13