Assignment-2(Machine Learning)

Name: Rajkumar Lenin Pillai

Question 1:

Solution:-

a) No of different Hypostheses = 2^{2^n} where n is the no of features. In file q1.csv the no of features is 6.

Hence, no of different Hypostheses = $2^{2^6} = 2^{64}$ = 18446744073709551616

b) Probability which guarantee a classifier that is 90% accurate =

$\delta \geq | H | . (1 - \epsilon)^{m} = 2^{64} . (1 - 0.1)^{200}$ = 13014323873.71398218910381120471= $1.301432387371\times10^{10}$

The above probability is greater than one hence, it does not give us any information as we are calculating probability $0 < \delta < 1$

Probability which guarantee a classifier that is 80% accurate =

$\delta \geq | H | . (1 - \epsilon)^{m} = 2^{64} . (1 - 0.2)^{200}$ = 0.76545051729020975577310162521901

c) Additional samples needed to get a classifier with 90% accuracy, 80% of the time

$$m \geq 1/\epsilon . [\ \ln(|H|) + \ln(1/\delta) = 1/0.1\ . [\ \ln(2^{2^6}) + \ln(1/0.2)\ ]$$

$$= 1/0.1\ . [\ 2^6 . \ln(2) + \ln(5)\ ]$$

$$= 1/0.1\ . [\ 64 . (0.693) + 1.609\ ]$$

$$= 1/0.1\ . [\ 44.352 + 1.609\ ]$$

$$= 1/0.1\ . [\ 45.961\ ]$$

$$= 459.61$$

$$m \geq 460$$

We need at least 460 more samples to get a classifier with 90% accuracy, 80% of the time

d) Model with a single conjunction, with each variable appearing either positively, negatively, or not at all

$| H | = 3^6$

In this model samples needed to get a classifier with 90% accuracy, 80% of the time.

$$m \geq 1/\epsilon . [\ \ln(|H|) + \ln(1/\delta) = 1/0.1\ . [\ \ln(3^6) + \ln(1/0.2)\ ]$$

$$= 1/0.1\ . [\ 6 . \ln(3) + \ln(5)\ ]$$

$$= 1/0.1\ . [\ 6 . (1.0986) + 1.609\ ]$$

$$= 82.006$$

$$m \geq 82$$

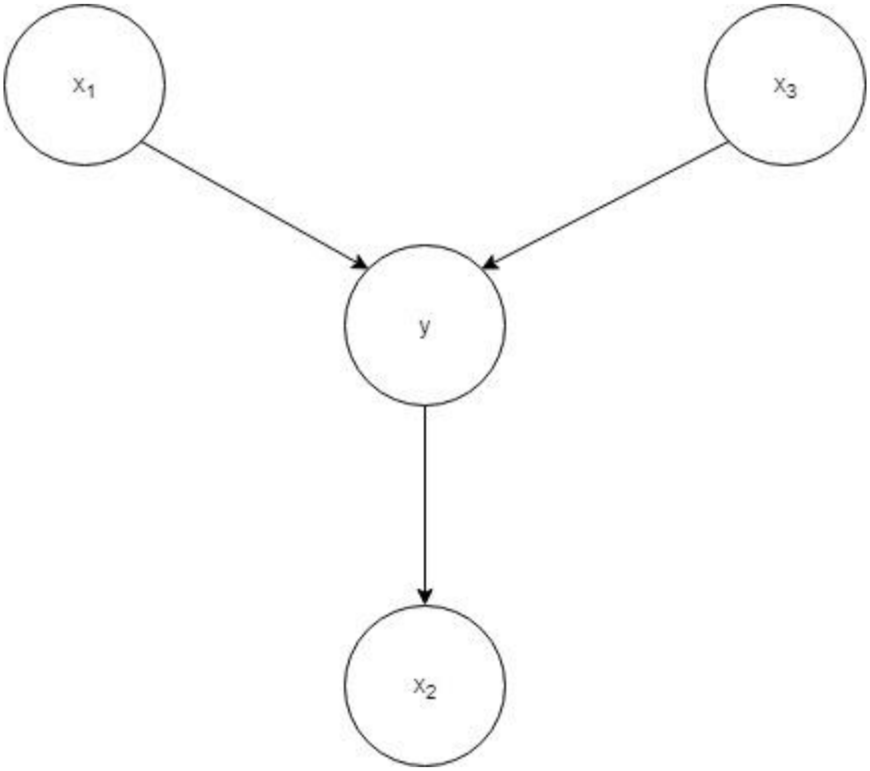
We need at least 82 more samples to get a classifier with 90% accuracy, 80% of the time

This might be problematic in practice because if our goal is prediction rather than knowing concepts then this model will yield wrong answer as 'none' for some cases. Such PAC model is actually not learning anything. So it is not a good model and it cannot be useful in practice.

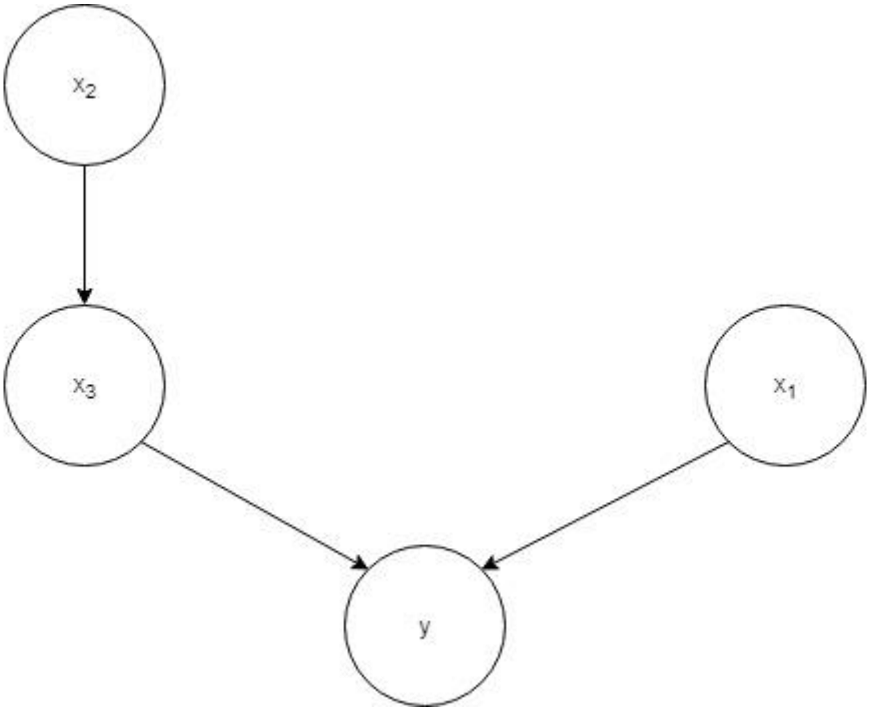
Question 2:

Solution:-

a.)



$$P(B) = P(x_1) . P(x_3) . P(y \mid x_1, x_3) . P(x_2 \mid y)$$



$$P (B) = P (x_2) . P (x_3 \mid x_2) . P (y \mid x_3 , x_1)$$

b.) For BayesNet-1:

P (Y=True | x_1=True ^ x_2=False ^ x_3 = True)

= P (Y=True | x_1=True ^ x_3 = True)

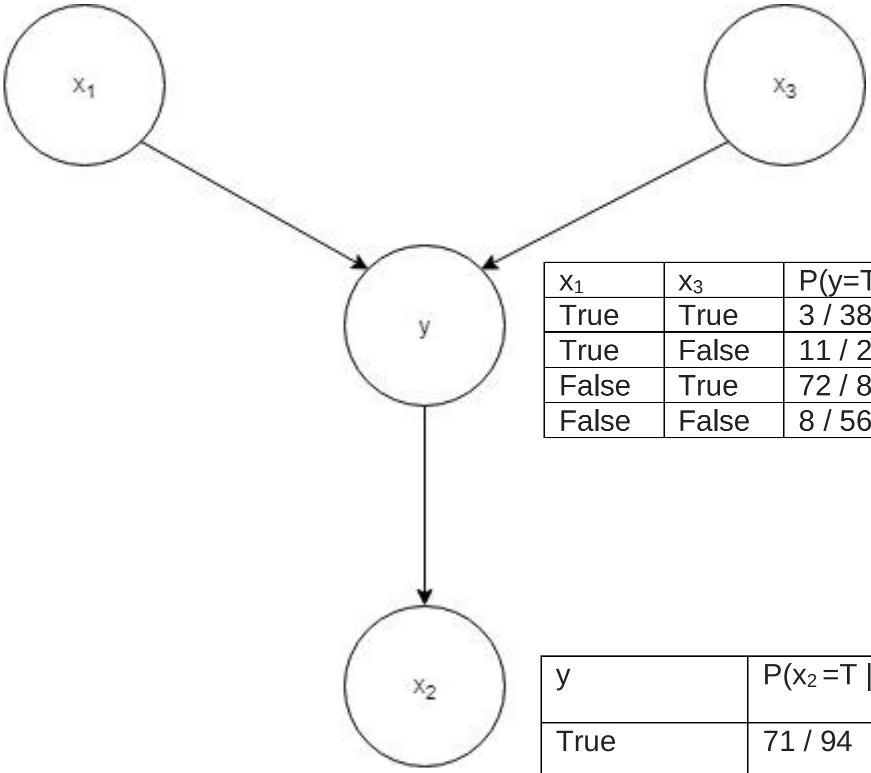
For BayesNet-2:

$P\ (Y=True \mid x_1=True \wedge x_2=False \wedge x_3 = True)$
$=P\ (Y=True \mid x_3 = True \wedge x_1=True)\ .\ P(\ x_3 = True \mid x_2=False\ )$

c.)

1. BayesNet-1

x_1	$P(x_1)$
True	63 / 200
False	137 / 200



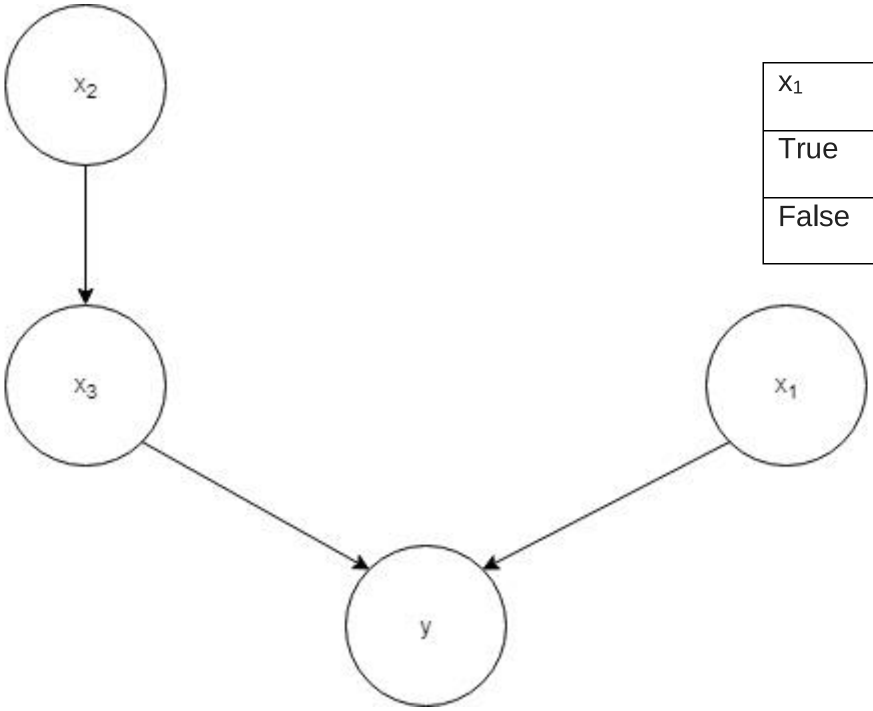
x_3	$P(x_3)$
True	119 / 200
False	81 / 200

x_1	x_3	$P(y=T \mid x_1, x_3)$	$P(y=F \mid x_1, x_3)$
True	True	3 / 38	35 / 38
True	False	11 / 25	14 / 25
False	True	72 / 81	9 / 81
False	False	8 / 56	48 / 56

y	$P(x_2 =T \mid y\ )$	$P(x_2 =F \mid y\ )$
True	71 / 94	23 / 94
False	22 / 106	84 / 106

2. BayesNet-2

x_2	$P(x_2)$
True	93 / 200
False	107 / 200



x_1	$P(x_1)$
True	63 / 200
False	137 / 200

x_2	$P(x_3 = T \mid x_2)$	$P(x_3 = F \mid x_2)$
True	63/ 93	30 / 93
False	56 / 107	51 / 107

x_1	x_3	$P(y=T \mid x_1, x_3)$	$P(y=F \mid x_1, x_3)$
True	True	3 / 38	35 / 38
True	False	11 / 25	14 / 25
False	True	72 / 81	9 / 81
False	False	8 / 56	48 / 56

d.) (i) For BayesNet-1

$P (x_1) = 63 / 200$
$P (x_3) = 119 / 200$

$P (x_1 \mid x_3) = P (x_1 , x_3) / P (x_3)$
$= (38 / 200) / (119 /200)$
$= 38 / 119$

$P (x_3 \mid x_1) = P (x_1 , x_3) / P (x_1)$
$= (38 / 200) / (63 /200)$
$= 38 / 63$

From above probabilities we can see that $P(x_1)$ is equal to $P (x_1 \mid x_3)$. Also, $P(x_3)$ is equal to $P (x_3 \mid x_1)$
Thus for the BayesNet-1 model the variables x_1 and x_3 are specified as independent and it is supported by the data provided.

(ii) For BayesNet-2

$P (x_1) = 63 / 200$
$P (x_2) = 93 / 200$

$P (x_1 \mid x_2) = P (x_1 , x_2) / P (x_2)$
$= (17 / 200) / (93 /200)$
$= 17 / 93$

$P (x_2 \mid x_1) = P (x_1 , x_2) / P (x_1)$
$= (17 / 200) / (63 /200)$
$= 17 / 63$

From above probabilities we can see that $P(x_1)$ is not equal to $P (x_1 \mid x_2)$. Also, $P(x_2)$ is not equal to $P (x_2 \mid x_1)$
Thus for the BayesNet-2 model the variables x_1 and x_2 are specified as independent but it is not supported by the data provided.

Question 3:

Solution:-

a.) In the given dataset in q3.csv file , the feature no 7 (no of sentences) and 8 (no of words) are numeric features.

The value of Mean and Standard deviation for feature no 7 when label = “False” is:
Mean =6. 190821256038648
Standard deviation = 2.533038409777209

The value of Mean and Standard deviation for feature no 8 when label = “False” is:
Mean =70.77053140096618
Standard deviation = 30.24857471611108

The value of Mean and Standard deviation for feature no 7 when label = “True” is:
Mean = 3.9767441860465116
Standard deviation = 1.940143986082091

The value of Mean and Standard deviation for feature no 8 when label = “True” is:
Mean =68.83720930232558
Standard deviation = 8.959858802874177

In the given dataset in q3.csv file , the feature no 1(in html),2 (has emoji),3 (sent to list),4 (from .com) ,5 (has my name) , 6 (has sig) are categoric features.

The value of likelihood probabilities when label = “False” of given features :-
The value of likelihood probabilities for feature 1 when feature1 = “False” is 0.41304347826086957
The value of likelihood probabilities for feature 1 when feature1= “True” is 0.5869565217391305

The value of likelihood probabilities for feature 2 when feature2 = “False” is 0.8526570048309179
The value of likelihood probabilities for feature 2 when feature2 = “True” is 0.1473429951690821

The value of likelihood probabilities for feature 3 when feature3 = “False” is 0.6884057971014492
The value of likelihood probabilities for feature 3 when feature3 = “True” is 0.3115942028985507

The value of likelihood probabilities for feature 4 when feature4 = “False” is 0.7246376811594203
The value of likelihood probabilities for feature 4 when feature4 = “True” is 0.2753623188405797

The value of likelihood probabilities for feature 5 when feature5 = “False” is 0.39855072463768115
The value of likelihood probabilities for feature 5 when feature5 = “True” is 0.6014492753623188

The value of likelihood probabilities for feature 6 when feature6 = “False” is 0.6763285024154589
The value of likelihood probabilities for feature 6 when feature6 = “True” is 0.32367149758454106

The value of likelihood probabilities when label = “True” of given features :-
The value of likelihood probabilities for feature 1 when feature1 = “False” is 0.2441860465116279
The value of likelihood probabilities for feature 1 when feature1= “True” is 0.7558139534883721

The value of likelihood probabilities for feature 2 when feature2 = “False” is 0.8023255813953488
The value of likelihood probabilities for feature 2 when feature2 = “True” is 0.19767441860465115

The value of likelihood probabilities for feature 3 when feature3 = “False” is 0.9302325581395349
The value of likelihood probabilities for feature 3 when feature3 = “True” is 0.06976744186046512

The value of likelihood probabilities for feature 4 when feature4 = “False” is 0.2558139534883721
The value of likelihood probabilities for feature 4 when feature4 = “True” is 0.7441860465116279

The value of likelihood probabilities for feature 5 when feature5 = “False” is 0.6511627906976745
The value of likelihood probabilities for feature 5 when feature5 = “True” is 0.3488372093023256

The value of likelihood probabilities for feature 6 when feature6 = “False” is 0.3372093023255814
The value of likelihood probabilities for feature 6 when feature6 = “True” is 0.6627906976744186

b.) Overall classification error rate based on a threshold P(Y) of 0.5 :-
Classification error rate = 20 %
Accuracy = 80 %

c.) For the given dataset q3.csv there are numeric and categoric features. To ignore some of the features two subsets were considered one with only the numeric features and the other with only the categoric features.
If there are only categoric variables then these variables have only binary values and the naïve bayes model implemented in such case is the Bernoulli naïve bayes. Using Baye’s theorem the conditional probability can be calculated as:

Posterior Probability = (Prior Probability x Likelihood) / Evidence

In the program, the evidence and Prior Probability are not calculated as they remain the same in the entire equation so, only prior probability is calculated. The classification error obtained with only the categorical feature is :-
Classification Error = 27.5 %
Accuracy = 72.5 %

The Bayesian model with only the numeric feature is Gaussian Naïve bayes model. The likelihood for such a model is calculated using formula:-

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

ImageSource:- https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

The classification error obtained with only the numeric feature is :-

Classification Error is = 24.0 %
Accuracy is = 76.0 %

For the two different subsets there are 3 versions predict function are present in the python program q3.py which are called according to the option specified by user.

For the full feature set the function Predictionofdataset() is used where the likelihood probabilities of all the features were multiplied for each class value and the maximum among these two values were considered to generate the label.

For the subset with category features the function Predictionofdatasetcategory() is used where the likelihood probabilities of all the features were multiplied for each class value and the maximum among these two values were considered to generate the label.

For the subset with numeric features the function Predictionofdatasetnumeric() is used where the likelihood probabilities of all the features were calculated using mean and standard deviation and these probabilities were multiplied for each class value and the maximum among these two values were considered to generate the label.

Running the Program:-

To execute the program with full feature set ,Type y after the program prints this question "Classification of data sample with full feature set?" on screen and after pressing Enter, the Classification error rate and accuracy can be seen.

If you do not wish to run the classifier with all features , type n after the program prints this question "Classification of data sample with full feature set?"

Now, if you want to run the classifier with categoric features , type 2 after the program prints this question "Classification of data sample with numeric feature or categoric feature set (1/2)" on screen and after pressing Enter, the Classification error rate and accuracy can be seen. But if you if you want to run the classifier with numeric features , type 1 after the program prints this question "Classification of data sample with numeric feature or categoric feature set (1/2)" on screen and after pressing Enter, the Classification error rate and accuracy can be seen.