Question 1

a) Suppose we are given the following dataset, where A,B,C are input binary random variables, and y is a binary output whose value we want to predict.

A	B	C	y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

How would a naive Bayes classifier predict y given this input: A = 0,B = 0,C = 1.

Answer: The classifier will predict 1

2.5 grades

$P(y=0)=3/7; P(y=1)=4/7$

$P(A=0|y=0)=2/3; P(B=0|y=0)=1/3; P(C=1|y=0)=1/3$

$P(A=0|y=1)=1/4; P(B=0|y=1)=1/2; P(C=1|y=1)=1/2$

Predicted y maximizes $P(A=0|y)P(B=0|y)P(C=1|y)P(y)$

$P(A=0|y=0)P(B=0|y=0)P(C=1|y=0)P(y=0)=0.0317$

$P(A=0|y=1)P(B=0|y=1)P(C=1|y=1)P(y=1)=0.0357$

Hence, the predicted y is 1.

b) Briefly describe the difference between a *maximum likelihood* hypothesis and a *maximum a posteriori* hypothesis.

Solutions:

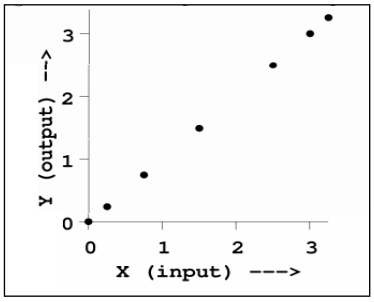
ML: maximize the data likelihood given the model, i.e., $\arg\max_W P(Data|W)$

MAP: $\arg\max_W P(W|Data)$

2 grade

Question 2

a) Consider the following data set with one input and one output



1) What is the mean squared training set error of running linear regression this data (using the model $y = \theta_0 + \theta_1 X$)?

Zero

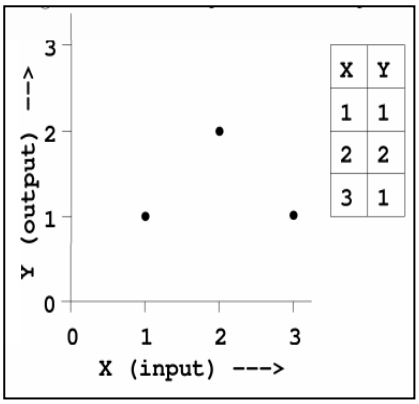
0.5 grade

2) What is the mean squared test set error of running linear regression this data. Assuming the rightmost three points are in the test set.and the other in the training data.

0.5 grade

zero

b) Consider the following data with one input and one output



X	Y
1	1
2	2
3	1

1- What is the mean squared training set error of running regression on this data (using the model $h_\theta(x) = \theta_0 + \theta_1 X$)
Hint : by symmetry it is clear that the best fit to the three data points is a horizontal line)

2 grade

$\Theta_0 = 1$, $\Theta_1 = 0$ $h_\theta(x) = 1$

MSE= $\frac{1}{n}\sum_{i=1}^{n}(Y_i - h_\theta(x_i))^2 = \sum_{i=1}^{m}\hat{\varepsilon}_i^2$

MSE = 1/3(0+1+0)= 1/3

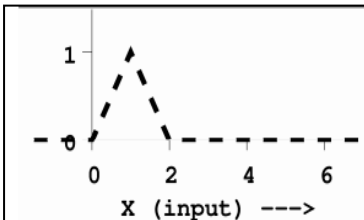
Or

$\Theta_0 = 1.5$, $\Theta_1 = 0$ $h_\theta(x) = 1$

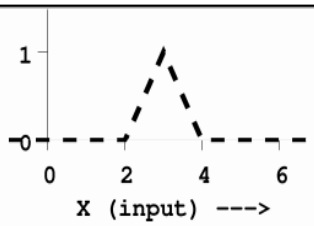
MSE= $\frac{1}{n}\sum_{i=1}^{n}(Y_i - h_\theta(x_i))^2 = \sum_{i=1}^{m}\hat{\varepsilon}_i^2$

MSE = 1/3(0.25+0.25+0.25)= 1/4

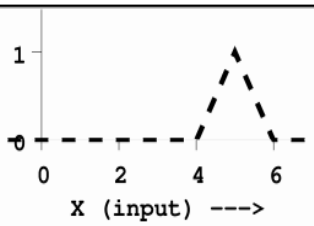
c) Suppose we plan to do regression with the following basis functions



$$\phi_1(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \le x < 1 \\ 2 - x & \text{if } 1 \le x < 2 \\ 0 & \text{if } 2 \le x \end{cases}$$



$$\phi_2(x) = \begin{cases} 0 & \text{if } x < 2 \\ x - 2 & \text{if } 2 \le x < 3 \\ 4 - x & \text{if } 3 \le x < 4 \\ 0 & \text{if } 4 \le x \end{cases}$$



$$\phi_3(x) = \begin{cases} 0 & \text{if } x < 4 \\ x - 4 & \text{if } 4 \le x < 5 \\ 6 - x & \text{if } 5 \le x < 6 \\ 0 & \text{if } 6 \le x \end{cases}$$

Our regression will be $y = \beta_1\phi_1(x) + \beta_2\phi_2(x) + \beta_3\phi_3(x)$.

Assume all our datapoints and future queries have $1 \le x \le 5$. Is this a generally useful set of basis functions to use? If "yes", then explain their prime advantage. If "no", explain their biggest drawback.

NO	1.5 grade
They're forced to predict y=0 at x=2 and x=4 (and forced to be close to zero nearby) no matter what the values of beta.	

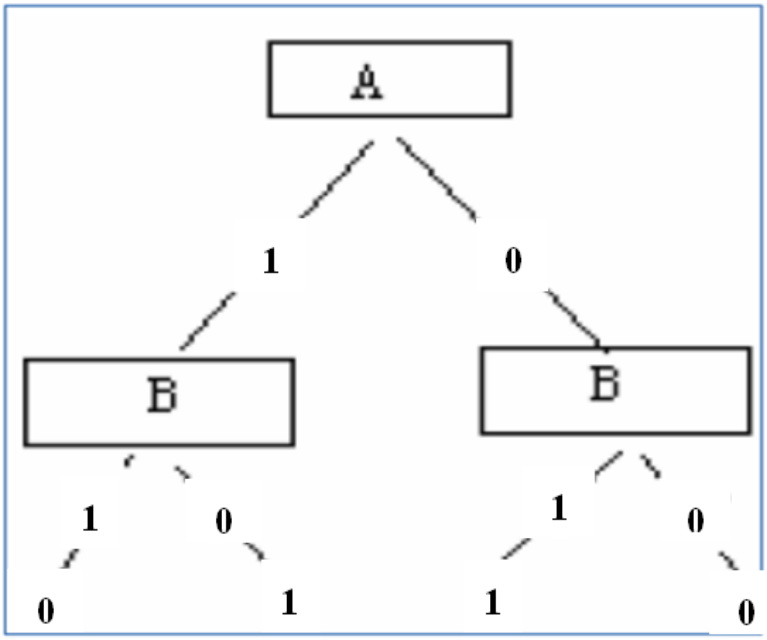
Question 3

a) Give the decision tree that represent XOR function

1.5 grade

A	B	Class
1	1	0
1	0	1
0	1	1
0	0	0

Answer:- the information gain for each feature =0
So you can use any one



b) Suppose that X_1, ..., X_m are categorical input attributes and Y is categorical output attribute. Suppose we plan to learn a decision tree.

For the following sentences, state which one is **true** or **false** with reason(s)

1.5 grade for each

1) If Xi and Y are independent in the distribution that generated this dataset, then Xi will not appear in the decision tree.

Answer: False (because the attribute may become relevant further down

the tree when the records are restricted to some value of another attribute) (e.g. XOR)

2) If $G(Y_j, X_i) = 0$ according to the values of entropy and conditional entropy computed from the data, then X_i will not appear in the decision tree.

Answer: False (because the attribute may become relevant further down the tree when the records are restricted to some value of another attribute) (e.g. XOR)

3) The maximum depth of the decision tree must be less than m+1

Answer: True because the attributes are categorical and can each be split only once