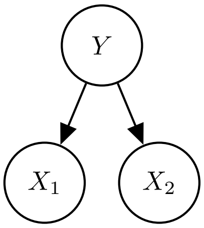
CS 188 Fall 2021 — Final Review: Machine Learning I

Q1. Naive Bayes

You are given a naive bayes model, shown below, with label Y and features X_1 and X_2. The conditional probabilities for the model are parameterized by p_1, p_2 and q.



X_1	Y	$P(X_1\|Y)$
0	0	p_1
1	0	$1-p_1$
0	1	$1-p_1$
1	1	p_1

X_2	Y	$P(X_2\|Y)$
0	0	p_2
1	0	$1-p_2$
0	1	$1-p_2$
1	1	p_2

Y	$P(Y)$
0	$1-q$
1	q

Note that some of the parameters are shared (e.g. $P(X_1=0|Y=0) = P(X_1=1|Y=1) = p_1$).

(a) Given a new data point with $X_1 = 1$ and $X_2 = 1$, what is the probability that this point has label $Y = 1$? Express your answer in terms of the parameters p_1, p_2 and q (you might not need all of them).

$P(Y=1|X_1=1, X_2=1) =$ ____________________

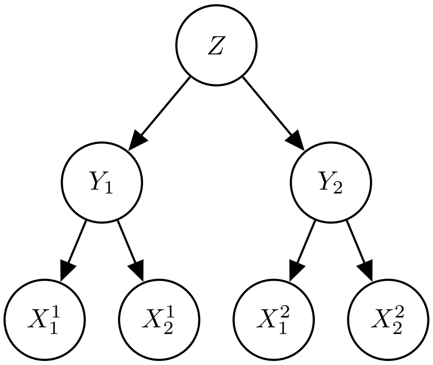
The model is trained with the following data:

sample number	1	2	3	4	5	6	7	8	9	10
X_1	0	0	1	0	1	0	1	0	1	1
X_2	0	0	0	0	0	0	0	1	0	0
Y	0	0	0	0	0	0	0	1	1	1

(b) What are the maximum likelihood estimates for p_1, p_2 and q?

$p_1 =$ ____________________ $p_2 =$ ____________________ $q =$ ____________________

(c) For the next part, the model you are given is no longer simple naive bayes. Now there are two distinct label variables Y_1, Y_2, and there is a super label Z which conditions all of these labels, thus giving us this hierarchical naive bayes model. The conditional probabilities for the model are parametrized by p_1, p_2, q_0, q_1 and r. **Note that some of the parameters are shared as in the previous part**.



X_1^i	Y_i	$P(X_1^i \mid Y_i)$
0	0	p_1
1	0	$1 - p_1$
0	1	$1 - p_1$
1	1	p_1

X_2^i	Y_i	$P(X_2^i \mid Y_i)$
0	0	p_2
1	0	$1 - p_2$
0	1	$1 - p_2$
1	1	p_2

Y_i	Z	$P(Y_i \mid Z)$
0	0	$1 - q_0$
1	0	q_0
0	1	$1 - q_1$
1	1	q_1

Z	$P(Z)$
0	$1 - r$
1	r

(i) What is the probability that $Z = 1$ given the partial data point $X_1^2 = 1, X_2^2 = 1, Y_1 = 1$? Simplify your answer as much as possible and express it in terms of the parameters p_1, p_2, q_0, q_1 and r (you might not need all of them).

$P(Z = 1 | X_1^2 = 1, X_2^2 = 1, Y_1 = 1) =$ ________________________

(ii) Now we are given a partial data point with $X_1^2 = 1, X_2^2 = 1, Y_1 = 1$. What is the probability that $Y_2 = 1$. Simplify your answer as much as possible and express it in terms of the parameters p_1, p_2, q_0, q_1 and r (you might not need all of them).

$P(Y_2 = 1 | X_1^2 = 1, X_2^2 = 1, Y_1 = 1) =$ ________________________

(d) Let L_{nb} and L_{hnb} be the likelihood of the training data under the naive bayes model and the hierarchical naive bayes model, respectively. Assume each of the models use their respective maximum likelihood parameters. Which of the following properties are guaranteed to be true?

○ $L_{nb} \leq L_{hnb}$

○ $L_{nb} \geq L_{hnb}$

○ $L_{nb} = L_{hnb}$

○ Insufficient information, the above relationships rely on the particular training data.

○ None of the above.

Q2. Perceptron

We would like to use a perceptron to train a classifier for datasets with 2 features per point and labels +1 or -1.

Consider the following labeled training data:

Features (x_1, x_2)	Label y^*
(-1,2)	1
(3,-1)	-1
(1,2)	-1
(3,1)	1

(a) Our two perceptron weights have been initialized to $w_1 = 2$ and $w_2 = -2$. After processing the first point with the perceptron algorithm, what will be the updated values for these weights?

(b) After how many steps will the perceptron algorithm converge? Write "never" if it will never converge.

Note: one steps means processing one point. Points are processed in order and then repeated, until convergence.

(c) Instead of the standard perceptron algorithm, we decide to treat the perceptron as a single node neural network and update the weights using gradient descent on the loss function.

The loss function for one data point is $Loss(y, y^*) = (y - y^*)^2$, where y^* is the training label for a given point and y is the output of our single node network for that point.

(i) Given a general activation function $g(z)$ and its derivative $g'(z)$, what is the derivative of the loss function with respect to w_1 in terms of g, g', y^*, x_1, x_2, w_1, and w_2?

$\frac{\partial Loss}{\partial w_1} =$

(ii) For this question, the specific activation function that we will use is:

$$g(z) = 1 \text{ if } z \geq 0 \text{ and } = -1 \text{ if } z < 0$$

Given the following gradient descent equation to update the weights given a single data point. With initial weights of $w_1 = 2$ and $w_2 = -2$, what are the updated weights after processing the first point?
Gradient descent update equation: $w_i = w_i - \alpha \frac{\partial Loss}{\partial w_1}$

(iii) What is the most critical problem with this gradient descent training process with that activation function?