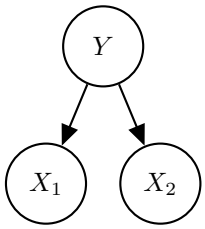
CS 188
Summer 2023

Final Review ML Solutions

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).

$P(Y = 1|X_1 = 1, X_2 = 1) =$ $\frac{p_1p_2q}{p_1p_2q+(1-p_1)(1-p_2)(1-q)}$

$$
\begin{aligned}
P(Y = 1, X_1 = 1, X_2 = 1) &= P(X_1 = 1|Y = 1)P(X_2 = 1|Y = 1)P(Y = 1) \\
&= p_1p_2q \\
P(Y = 0, X_1 = 1, X_2 = 1) &= P(X_1 = 1|Y = 0)P(X_2 = 1|Y = 0)P(Y = 0) \\
&= (1-p_1)(1-p_2)(1-q) \\
P(Y = 1|X_1 = 1, X_2 = 1) &= \frac{P(Y = 1, X_1 = 1, X_2 = 1)}{P(X_1 = 1, X_2 = 1)} \\
&= \frac{P(Y = 1, X_1 = 1, X_2 = 1)}{P(Y = 1, X_1 = 1, X_2 = 1) + P(Y = 0, X_1 = 1, X_2 = 1)} \\
&= \frac{p_1p_2q}{p_1p_2q + (1-p_1)(1-p_2)(1-q)}
\end{aligned}
$$

The model is trained with the following data:

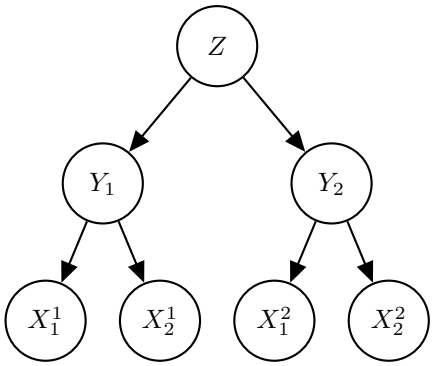
sample number	1	2	3	4	5	6	7	8	9	10
X_1	0	0	1	0	1	0	1	0	1	1
X_2	0	0	0	0	0	0	0	1	0	0
Y	0	0	0	0	0	0	0	1	1	1

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

$p_1 =$ $\frac{3}{5}$ $p_2 =$ $\frac{4}{5}$ $q =$ $\frac{3}{10}$

The maximum likelihood estimate of p_1 is the fraction of counts of samples in which $X_1 = Y$. In the given training data, samples $1, 2, 4$ and 6 have $X_1 = Y = 0$ and samples 9 and 10 have $X_1 = Y = 1$, so 6 out of the 10 samples have $X_1 = Y$ and thus $p_1 = \frac{6}{10} = \frac{3}{5}$. Analogously, 8 out of the 10 samples have $X_2 = Y$ and thus $p_2 = \frac{8}{10} = \frac{4}{5}$. The maximum likelihood estimate of q is the fraction of counts of samples in which $Y = 1$, thus $q = \frac{3}{10}$.

(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).

$$P(Z = 1 \mid X_1^2 = 1, X_2^2 = 1, Y_1 = 1) = \frac{rq_1(q_1p_1p_2 + (1-q_1)(1-p_1)(1-p_2))}{rq_1(q_1p_1p_2 + (1-q_1)(1-p_1)(1-p_2)) + (1-r)q_0(q_0p_1p_2 + (1-q_0)(1-p_1)(1-p_2))}$$

(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).

$$P(Y_2 = 1 \mid X_1^2 = 1, X_2^2 = 1, Y_1 = 1) = \frac{(1-r)q_0^2p_1p_2 + rq_1^2p_1p_2}{(1-r)q_0^2p_1p_2 + rq_1^2p_1p_2 + (1-r)q_0(1-q_0)(1-p_1)(1-p_2) + rq_1(1-q_1)(1-p_1)(1-p_2)}$$

(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?

- [x] $L_{nb} \leq L_{hnb}$
- [ ] $L_{nb} \geq L_{hnb}$
- [ ] $L_{nb} = L_{hnb}$
- [ ] Insufficient information, the above relationships rely on the particular training data.
- [ ] None of the above.

The hierarchical naive bayes model can represent all of the same probability distributions that naive bayes model can, but also some more (when $q_0 \neq p_2$ or $r \neq 0.5$). So in general the hierarchical model allows for more fitting of the training data, which results in a higher likelihood.

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?

For the first point, $y = g(w_1x_1 + w_2x_2) = g(2 \cdot -1 + -2 \cdot 2) = g(-5) = -1$, which is incorrectly classified. To update the weights, we add the first data point: $w_1 = 2 + (-1) = 1$ and $w_2 = -2 + 2 = 0$.

(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.

The data is not seperable, so it will never converge.

(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?

$\frac{\partial Loss}{\partial w_1} = 2(g(w_1x_1 + w_2x_2) - y^*)g'(w_1x_1 + w_2x_2)x_1$

(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: $w_i = w_i - \alpha \frac{\partial Loss}{\partial w_1}$

Because the gradient of g is zero, the weights will stay $w_1 = 2$ and $w_2 = -2$.

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

The gradient of that activation function is zero, so the weights will not update.