# Graded assignment

## Question 1

### Statement

Consider the two different generative model-based algorithms.

1. Model $1$: chances of occurring a feature are affected by the occurrence of other features and the model does not impose any additional condition on conditional independence of features.
2. Model $2$: chances of occurring a feature are not affected by the occurrence of other features and therefore, the model assumes that features are conditionally independent of the label.

Which model has more independent parameters to estimate?

### Options

#### (a)

Model $1$

#### (b)

Model $2$

### Answer

(a)

### Solution:

In the first model, features are not independent, therefore, we need to find the probabilities (or density) for each and every possible example given the labels whereas in model 2, the features are independent, therefore we need to find the pmf (or pdf) of the features only.

That is the model 1 has more parameters to estimate.

---

## Question 2

### Statement

Which of the following statement is/are always correct in context to the naive Bayes classification algorithm for binary classification with all binary features? Here, $\hat{p}_j^y$ denotes the estimate for the probability that the $j^{th}$ feature value of a data point is $1$ given that the point has the label $y$.

## Options

### (a)

If $\hat{p}_j^y = 0.2$ for $y = 0$, then $\hat{p}_j^y = 0.8$ for $y = 1$

### (b)

$\sum_{j=1}^{d} \hat{p}_j^y = 1$ for any $y$

### (c)

If $\hat{p}_j^y = 0$ for $y = 0$, then $\hat{p}_j^y = 0$ for $y = 1$

### (d)

If $\hat{p}_j^1 = 0$, no labeled $1$ example in the training dataset takes $j^{th}$ feature values as $1$.

### (e)

None of the above

## Answer

(d)

## Solution

In general, $\hat{p}_j^y =$ estimate for $P(f_j = 1|y)$.

It means that $\hat{p}_j^y$ denotes the parameters of different distributions $f_j|y$ for different $y$ and for different $j$.

Therefore, If $\hat{p}_j^y = 0.2$ for $y = 0$, then it is not mandatory that $\hat{p}_j^y = 0.8$ for $y = 1$ as they come from different distributions.

For different $j$, distributions $f_j|y$ are different distributions. Therefore, it is not necessary that $\sum_{j=1}^{d} \hat{p}_j^y = 1$. What we can say is that $\sum_{i \in R(f_j)} P(f_j = i|y) = 1$

If $\hat{p}_j^y = 0$ for $y = 0$, It implies that there is no labeled zero examples such that $j^{th}$ feature value is $1$. It doesn't mean that all $j^{th}$ feature value is $1$ for all labeled one examples.

If $\hat{p}_j^1 = 0$, no labeled $1$ example in the training dataset takes $j^{th}$ feature values as $1$.

---

# Question 3

## Statement

A naive Bayes model is trained on a dataset containing $d$ features $f_1, f_2, \ldots, f_d$. Labels are $0$ and $1$. If a test point was predicted to have the label $1$, which of the following expression should be sufficient for this prediction?

## Options

### (a)

$P(y=1) > P(y=0)$

### (b)

$\prod_{i=1}^{d} P(f_i|y=1) > \prod_{i=1}^{d} P(f_i|y=0)$

### (c)

$\left(\prod_{j=1}^{d} (\hat{p}_j^1)^{f_j}(1-\hat{p}_j^1)^{1-f_j}\right) P(y=1) > \left(\prod_{j=1}^{d} (\hat{p}_j^0)^{f_j}(1-\hat{p}_j^0)^{1-f_j}\right) P(y=0)$

### (d)

None of the above

## Answer

(c)

## Solution

A test example is predicted label $1$, it implies that

$$
\begin{aligned}
& P(y=1|x) > P(y=0|x) \\
\Rightarrow & \frac{P(x|y=1).P(y=1)}{P(x)} > \frac{P(x|y=0).P(y=0)}{P(x)} \\
\Rightarrow & P(x|y=1).P(y=1) > P(x|y=0).P(y=0) \\
\Rightarrow & \left(\prod_{j=1}^{d} (\hat{p}_j^1)^{f_j}(1-\hat{p}_j^1)^{1-f_j}\right) P(y=1) > \left(\prod_{j=1}^{d} (\hat{p}_j^0)^{f_j}(1-\hat{p}_j^0)^{1-f_j}\right) P(y=0)
\end{aligned}
$$

# Question 4

## Statement

Consider a binary classification dataset contains only one feature and the data points given the label follow the given distribution

$$
\begin{aligned}
x|(y=0) &\sim \mathrm{N}(0,2) \\
x|(y=1) &\sim \mathrm{N}(2,\sigma^2)
\end{aligned}
$$

If the decision boundary learned using the gaussian naive Bayes algorithm is linear, what is the value of $\sigma^2$?

## Answer

$2$

## Solution

Since the decision boundary is linear, both theiances will be the same. That is $\sigma^2 = 2$

# Question 5

## Statement

Consider a binary classification dataset with two binary features $f_1$ and $f_2$. The $f_2$ feature values are $0$ for all label '$0$' examples but the label '$1$' examples take both values $1$ and $0$ for the feature $f_2$. If we apply the naive Bayes algorithm on the same dataset, what will be the prediction for point $[1, 1]^T$?

## Options

### (a)

Label $0$

### (b)

Label $1$

### (c)

Insufficient information to predict.

## Answer

(c)

## Solution

Given that the $f_2$ feature values are $0$ for all label '$0$' examples it implies that $\hat{p}_2^0 = 0$.

Therefore, $p(y = 0|x) = 0$

But the label '$1$' examples take both values $1$ and $0$ for the feature $f_2$. It implies that $\hat{p}_2^1 > 0$.

Still the value of $p(y = 1|x)$ can be 0 if the value of $\hat{p}_1^1$ is zero. So, we need the value of $\hat{p}_1^1$ to make any conclusion.

---

# Common data for questions 6 and 7

## Statement

Consider the following binary classification dataset with two features $f_1$ and $f_2$. The data points given the labels follow the Gaussian distribution. The dataset is given as

| $f_1$ | $f_2$ | **label** $y$ |
|---|---|---|
| 0.5 | 1.3 | 1 |
| 0.7 | 1.1 | 1 |
| 1.3 | 2.0 | 0 |
| 2.3 | 2.4 | 0 |

# Question 6

## Statement

What will be the value of $\hat{p}$, the estimate for $P(y = 1)$?

## Answer

$0.5$

## Solution

$$\begin{aligned}\hat{p} &= \frac{\sum_{i=1}^{n} y_i}{n} \\ &= 2/4 = 0.5\end{aligned}$$

# Question 7

## Statement

What will be the value of $\hat{\mu}_0$?

## Options

### (a)

$(1.8, 2.2)$

### (b)

$(0.6, 1.2)$

**(c)**

$(2.0, 2.0)$

**(d)**

$(0.8, 1.2)$

## Answer

(a)

## Solution

$$\begin{aligned}\hat{\mu}_0 &= \frac{\sum_{i=1}^{n} \mathbb{1}(y_i = 0) x_i}{\sum_{i=1}^{n} \mathbb{1}(y_i = 0)} \\ &= \frac{(1.3, 2.0) + (2.3, 2.4)}{2} \\ &= (1.8, 2.2)\end{aligned}$$

# Question 8

## Statement

Consider a binary classification dataset containing two features $f_1$ and $f_2$. The feature $f_1$ is categorical which can take three values and the feature $f_2$ is numerical that follows the Gaussian distribution. How many independent parameters must be estimated if we apply the naive Bayes algorithm to the same dataset?

## Answer

$9$

## Solution

We need one parameter for $P(y = 1)$ as $y$ takes only two values.

For a given label (say $y = 1$)

feature $f_1$ can take three values, therefore we need two estimates for $P(f_1 = 0|y = 1)$ and $P(f_1 = 1|y = 1)$.

Similarly, two estimates if $y = 0$

For feature $f_2$, we need $\mu_0$, $\mu_1$, $\Sigma_0$ and $\Sigma_1$.

Therefore, total number of parameters = $1 + 2 + 2 + 4 = 9$

---

# Common data for questions 9 and 10

## Statement

A binary classification dataset has 1000 data points belonging to $\{0,1\}^2$. A naive Bayes algorithm was run on the same dataset that results in the following estimate:

| $\hat{p}$ **, estimate for** $P(y=1)$ | **0.3** |
|---|---|
| $\hat{p}_1^0$, estimate for $P(f_1=1\|y=0)$ | $0.2$ |
| $\hat{p}_2^0$, estimate for $P(f_2=1\|y=0)$ | $0.3$ |
| $\hat{p}_1^1$, estimate for $P(f_1=1\|y=1)$ | $0.1$ |
| $\hat{p}_1^2$, estimate for $P(f_2=1\|y=1)$ | $0.02$ |

# Question 9

## Statement

What is the estimated value of $P(f_2=0|y=1)$? Write your answer correct to two decimal places.

## Answer

$0.98$ Range: [0.97, 0.99]

## Solution

$$\begin{aligned}\text{estimate for } P(f_2=0|y=1) &= 1-(\text{estimate for } P(f_2=1|y=1)) \\ &= 1-0.02=0.98\end{aligned}$$

# Question 10

## Statement

What will be the predicted label for the data point $[0,1]$?

## Answer

$0$ No range is required

## Solution

$$\begin{aligned}P(y=0|x) &\propto P(x|y=0).P(y=0) \\ &\propto P(f_1=0|y=0).P(f_2=1|y=0).P(y=0) \\ &\propto (1-0.2)(0.3)(1-0.3) \\ &= 0.168\end{aligned}$$

$$\begin{aligned}P(y=1|x) &\propto P(x|y=1).P(y=1) \\ &\propto P(f_1=0|y=1).P(f_2=1|y=1).P(y=1) \\ &\propto (1-0.1)(0.02)(0.3) \\ &= 0.054\end{aligned}$$

Since $P(y=0|x) > P(y=1|x)$, therefore the point will be predicted label $0$.