1) Consider the two different generative model-based algorithms. **1 point**

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.
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 parameters to estimate?

In Model 1:
no. of parameters depends on all possible joint probabilities of features
So for n features that can take k distinct values
no. of parameters $k^n - 1$

In Model 2:
The number of parameters are reduced significantly
n → no. of features
m → no. of possible labels
k → possible values for each feature
$m \cdot n \cdot (k-1) + (m-1)$

- [x] Model 1
- [ ] Model 2

2) 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. **1 point**

- [ ] If $\hat{p}_j^y = 0.2$ for $y = 0$, then $\hat{p}_j^y = 0.8$ for $y = 1$
- [ ] $\sum_{j=1}^{d} \hat{p}_j^y = 1$ for any y
- [ ] If $\hat{p}_j^y = 0$ for $y = 0$, then $\hat{p}_j^y = 0$ for $y = 1$
- [x] If $\hat{p}_j^1 = 0$, no labeled 1 example in the training dataset takes j^{th} feature values as 1.
- [ ] None of the above

3) A naive Bayes model is trained on a dataset containing d features $f_1, f_2, ..., f_d$. 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?

- [ ] $P(y = 1) > P(y = 1)$
- [ ] $\prod_{i=1}^{d} P(f_i|y = 1) > \prod_{i=1}^{d} P(f_i|y = 0)$
- [x] $\left(\prod_{j=1}^{d} (\hat{p}_j^1)^{f_j}(1 - \hat{p}_j^1)^{1-f_j}\right) P(y = 1) > \left(\prod_{j=1}^{d} (\hat{p}_j^0)^{f_j}(1 - \hat{p}_j^0)^{1-f_j}\right) P(y = 0)$
- [ ] None of the above

Decision boundary is determined by equating posterior probabilities $P(y=0/x) = P(y=1/x)$

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

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

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

$P(y=0)\cdot\frac{1}{\sqrt{2\pi\cdot 2}}e^{\left(\frac{-x^2}{2\cdot 2}\right)} = P(y=1)\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{\left(\frac{-(x-2)^2}{2\sigma^2}\right)}$

Take logarithm & simplify

$\ln(P(y=0)) - \ln(P(y=1)) - \frac{x^2}{4} = \frac{-(x-2)^2}{2\sigma^2} + \ln(\sqrt{\sigma^2}) - \ln(\sqrt{2})$

coeff of x^2 are equal

$-\frac{1}{4} = -\frac{1}{2\sigma^2} \quad : \quad 2\sigma^2 = 4 \quad \sigma^2 = 2$

2

5) 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 $[1,1]^T$? **1 point**

- ○ Label 0
- ● Label 1
- ○ Insufficient information to predict.

Common data for questions 6 and 7

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

$\hat{p} = \frac{2}{4} = \frac{1}{2} = 0.5$

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

0.5

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

- ● $(1.8, 2.2)$
- ○ $(0.6, 1.2)$
- ○ $(2.0, 2.0)$
- ○ $(0.8, 1.2)$

$\mu_{f_1}/y=0$

$f_1 = \frac{1.3+2.3}{2} = 1.8$

$f_2 = \frac{2.0+2.4}{2} = 2.2$

8) 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 parameters must be estimated if we apply the naive Bayes algorithm to the same dataset? Assume that covariance is same for both the distributions $f_2|(\text{class } 0)$ and $f_2|(\text{class } 1)$.

8

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

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

0.98

$P(f_2=0/y=1) = 1 - P(f_2=1/y=1)$
$= 1 - 0.02$
$= 0.98$

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

0