Tut 5: Naive Bayes Classifier

Feb 2025

The general mathematical formulation of a generative model:

$$\begin{aligned} h_D(\boldsymbol{x}) &= \underset{j\in\{1,\dots,K\}}{\operatorname{argmax}}\ \mathbb{P}(Y=j|\boldsymbol{X}=\boldsymbol{x}) \\ &= \underset{j\in\{1,\dots,K\}}{\operatorname{argmax}}\ \frac{\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=j)\mathbb{P}(Y=j)}{\mathbb{P}(\boldsymbol{X}=\boldsymbol{x})} \\ &= \underset{j\in\{1,\dots,K\}}{\operatorname{argmax}}\ \mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=j)\mathbb{P}(Y=j) \\ &= \underset{j\in\{1,\dots,K\}}{\operatorname{argmax}}\ [\ln\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=j)+\ln\mathbb{P}(Y=j)] \end{aligned} \tag{5.1}$$

Naive Bayes:

$$\boxed{\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}|Y=j) \approx \prod_{i=1}^{p}\mathbb{P}(X_i=x_i|Y=j)}$$

1. (Jan 2022 Final Q4(a), 10 marks) The training data for part (a) is given in Table 4.1.

Table 4.1: Training data for credit card application approval.

Age	PriorDefault	Employed	Approved
59.67	Yes	False	+
27.25	No	True	-
20.67	No	False	-
16.50	No	False	-
26.67	Yes	True	+
37.50	Yes	False	-
36.25	Yes	True	+
21.17	No	False	-
32.33	Yes	False	+
58.42	Yes	True	+

Use the Naïve Bayes classifier model without Laplace smoothing to predict if the credit card approval is positive or negative for the person is of age 38.17, has a prior default and is employed.

Solution. Let Y=Approved, X_1=Age, X_2=PriorDefault, X_3=Employed.

$$\begin{aligned} &P(Y=+|X_1=38.17, X_2=Yes, X_3=True) \\ \propto &P(X_1=38.17|Y=+)\times P(X_2=Yes|Y=+)\times P(X_3=True|Y=+)P(Y=+) \end{aligned}$$

[1 mark]

Y	$P(Y)$	$X_1=38.17$	$X_2=Yes$	$X_3=True$	Product	Prob
$+$	$\frac{5}{10}=0.5$	0.02491317	$\frac{5}{5}=1$	$\frac{3}{5}=0.6$	0.0074740	0.9681
$-$	$\frac{5}{10}=0.5$	0.01230699	$\frac{1}{5}=0.2$	$\frac{1}{5}=0.2$	0.0002461	0.0319
	[1.5 marks]	[3 marks]	[1.5 marks]	[1.5 marks]	[0.5 mark]	

Using scientific calculator, we can obtain the estimate:

$$\mu_+ = \frac{59.67 + 26.67 + 36.25 + 32.33 + 58.42}{5} = 42.668$$

$$\sigma_+ = \sqrt{\frac{(59.67 - \mu_+)^2 + (26.67 - \mu_+)^2 + ... + (58.42 - \mu_+)^2}{5-1}} = 15.33945$$

$$P(X_1 = 38.17|Y = +) = \frac{1}{\sqrt{2\pi}(15.33945)} \exp(-\frac{(38.17 - 42.668)^2}{2(235.2986)}) = 0.02491317$$

Similarly,

$$\mu_- = 24.618$$
$$\sigma_- = 8.158544805$$

Since the product $P(X_1 = 38.17|Y = +) \times P(X_2 = Yes|Y = +) \times P(X_3 = True|Y = +)P(Y = +) > P(X_1 = 38.17|Y = -) \times P(X_2 = Yes|Y = -) \times P(X_3 = True|Y = -)P(Y = -)$, the credit card approval is **positive**. ....[1 mark] □

2. Ahmad would like to construct a model to decide if a day is suitable to play tennis. The table below shows the results whether to play tennis, based on Outlook, Temperature and Wind, collected by Ahmad.

Day	Outlook	Temperature	Wind	PlayTennis
D1	Sunny	34	Weak	No
D2	Sunny	32	Strong	No
D3	Overcast	28	Weak	Yes
D4	Rain	22	Weak	Yes
D5	Rain	16	Weak	Yes
D6	Rain	8	Strong	No
D7	Overcast	12	Strong	Yes
D8	Sunny	20	Weak	No
D9	Sunny	10	Weak	Yes
D10	Rain	23	Weak	Yes
D11	Sunny	19	Strong	Yes
D12	Overcast	21	Strong	Yes
D13	Overcast	31	Weak	Yes
D14	Rain	25	Strong	No

Using Naïve Bayes approach with Laplace smoothing, predict whether a sunny day with strong wind, 27°C, is suitable to play tennis.

Solution. Let $y = PlayTennis(Yes = 1; No = 0)$

X_1=Outlook; X_2=Temperature; X_3=Wind

New observation: $x_1^* = sunny$; $x_2^* = 27$; $x_3^* = strong$

Steps for finding the posterior $\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}^*)$.

- Prior, $\mathbb{P}(Y = 1) = \frac{9}{14}$
- Density functions,

  $\mathbb{P}(X_1 = sunny|Y = 1) = \frac{2+1}{9+3} = \frac{1}{4}$

  $\mathbb{P}(X_2 = 27|Y = 1) = \frac{1}{\sqrt{2\pi(s^2_{x2:y=1})}} e^{-\frac{(x_2^* - \overline{x}_{2:y=1})^2}{2s^2_{x2:y=1}}} = \frac{1}{\sqrt{2\pi}(6.8880)} e^{-\frac{(27-20.2222)^2}{2(47.4445)}} = 0.0357$

where $\overline{x_{2:y=1}} = 20.2222$; $s_{x_{2:y=1}} = 6.8880$
$\mathbb{P}(X_3 = strong|Y = 1) = \frac{3+1}{9+2} = \frac{4}{11}$

- Hence, posterior probability for PlayTennis=Yes is

$$
\begin{aligned}
&\mathbb{P}(\hat{Y} = 1|\boldsymbol{X} = \boldsymbol{x}^*) \\
&\propto P(Y = 1) \cdot \mathbb{P}(X_1 = sunny|Y = 1) \cdot \mathbb{P}(X_2 = 27|Y = 1) \cdot \mathbb{P}(X_3 = strong|y = 1) \\
&= \frac{9}{14} \cdot \frac{1}{4} \cdot 0.0357 \cdot \frac{4}{11} \approx 0.0021
\end{aligned}
$$

Steps for finding the posterior $\mathbb{P}(Y = 0|\boldsymbol{X} = \boldsymbol{x}^*)$.

- Prior, $\mathbb{P}(Y = 0) = \frac{5}{14}$
- Density functions,

$$\mathbb{P}(X_1 = sunny|Y = 0) = \frac{3+1}{5+3} = \frac{1}{2}$$

$$\mathbb{P}(X_2 = 27|Y = 0) = \frac{1}{\sqrt{2\pi(s^2_{x_{2:y=0}})}} e^{-\frac{(x_2^* - \overline{x_{2:y=0}})^2}{2s^2_{x_{2:y=0}}}} = \frac{1}{\sqrt{2\pi(10.4499)}} e^{-\frac{(27-23.8)^2}{2(109.20)^2}} = 0.0364$$

where $\overline{x_{2:y=0}} = 23.8000$; $s_{x_{2:y=0}} = 10.4499$
$\mathbb{P}(X_3 = strong|y = 0) = \frac{3+1}{5+2} = \frac{4}{7}$
Hence, posterior probability for (PlayTennis = No) is

$$
\begin{aligned}
&\mathbb{P}(Y = 0|\boldsymbol{X} = \boldsymbol{x}^*) \\
&\propto \mathbb{P}(y = 0) \cdot \mathbb{P}(X_1 = sunny|Y = 0) \cdot \mathbb{P}(X_2 = 27|Y = 0) \cdot \mathbb{P}(X_3 = strong|Y = 0) \\
&= \frac{5}{14} \cdot \frac{1}{2} \cdot 0.0364 \cdot \frac{4}{7} \approx 0.0037
\end{aligned}
$$

Since $\mathbb{P}(Y = 0|\boldsymbol{X} = \boldsymbol{x}^*) > \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}^*)$, the day is not suitable to play tennis. □

3. (Jan 2021 Final Q4(b)) Suppose the mood (M) of a student is affected by two features, the weather (W) and his result (R) and the Table 4.2.

Table 4.2: Observed Data.

Weather (W)	Result (R)	Mood (M)
Bad	Poor	Unhappy
Good	Poor	Unhappy
Good	Poor	Unhappy
Good	Poor	Unhappy
Bad	Good	Unhappy
Bad	Good	Happy
Bad	Good	Happy
Good	Good	Happy

(a) Using Table 4.2 and a Naive Bayes classifier to predict the mood if today's situation is that the weather is good, the result is good. Show your computations clearly and write down the classifier's prediction. (1.5 marks)

Solution. Let Unhappy=U, Happy=H, G=Good. Then

$$
\begin{aligned}
&P(M = U|W = G, R = G) \\
&\propto P(W = G|M = U) \times P(R = G|M = U) \times P(M = U) = \frac{3}{5} \times \frac{1}{5} \times \frac{5}{8} = 0.075
\end{aligned}
$$

[0.6 mark]

$$P(M = H|W = G, R = G)$$

$$\propto P(W = G|M = H) \times P(R = G|M = H) \times P(M = H) = \frac{1}{3} \times \frac{3}{3} \times \frac{3}{8} = 0.125$$

[0.6 mark]

The classifier's prediction of the mood is **Happy**. .... [0.3 mark] ☐

(b) Using Table 4.2 and a Naive Bayes classifier to predict the mood if today's situation is that the weather is bad, the result is poor. Show your computations clearly and write down the classifier's prediction. (1.5 marks)

Solution. Let Unhappy=U, Happy=H, B=Bad, P=Poor. Then

$$P(M = U|W = B, R = P)$$

$$\propto P(W = B|M = U) \times P(R = P|M = U) \times P(M = U) = \frac{2}{5} \times \frac{4}{5} \times \frac{5}{8} = 0.2$$

[0.6 mark]

$$P(M = Happy|W = R, R = P)$$

$$\propto P(W = B|M = H) \times P(R = P|M = H) \times P(M = H) = \frac{2}{3} \times \frac{0}{3} \times \frac{3}{8} = 0$$

[0.6 mark]

The classifier's prediction of the mood is **Unhappy**. .... [0.3 mark] ☐

(c) Suppose an additional feature, exercise (E), which indicates that the student will carry out outdoor exercise or not, is added to the Table 4.2 to form Table 4.3.

Table 4.3: Observed Data with New Feature.

Weather (W)	Result (R)	Exercise (E)	Mood (M)
Bad	Poor	No	Unhappy
Good	Poor	Yes	Unhappy
Good	Poor	Yes	Unhappy
Good	Poor	Yes	Unhappy
Bad	Good	No	Unhappy
Bad	Good	No	Happy
Bad	Good	No	Happy
Good	Good	Yes	Happy

Using Table 4.3 and the Naive Bayes Classifier to the mood if W=Good, R= Good, E=Yes. Show your computations and the classifier's prediction. Will the new feature improve the performance of the Naive Bayes classifier from the one built based on Table 4.2? Justify your answer. (2 marks)

Solution. Let Unhappy=U, Happy=H, G=Good, Y=Yes. Then

$P(M = U|W = G, R = G, E = Y)$
$\propto P(W = G|M = U) \times P(R = G|M = U) \times P(E = Y|M = U) \times P(M = U)$
$= \frac{3}{5} \times \frac{1}{5} \times \frac{3}{5} \times \frac{5}{8} = 0.045$ [0.3 mark]
$P(M = H|W = G, R = G, E = Y)$
$\propto P(W = G|M = H) \times P(R = G|M = H) \times P(E = Y|M = H) \times P(G = H)$
$= \frac{1}{3} \times \frac{3}{3} \times \frac{1}{3} \times \frac{3}{8} = 0.04166667$ [0.3 mark]

The classifier's prediction of the mood is **Unhappy**. .... [0.2 mark]

No. .... [0.2 mark]

The new feature E will not improve the performance of the Naive Bayes classifier's prediction because the new feature E is correlated with the feature W and violates the assumption in Naive Bayes classifier. .... [1 mark] ☐

4. (Final Exam Jan 2023, Q5(a)) The data in Table 5.1 is from a study of car evaluation. The values of the predictors are listed below:

- X_1=`maint` (price of the maintenance): vhigh, high, med, low
- X_2=`persons` (capacity in terms of persons to carry): 2, 4, more
- X_3=`lugboot` (the size of luggage boot): small, med, big
- X_4=`safety` (estimated safety of the car): low, med, high
- Y=`class` (car acceptability): unacc, acc, good;

Obs.	`maint`	`persons`	`lugboot`	`safety`	`class`
1	med	more	big	high	good
2	low	more	small	high	good
3	low	4	big	high	good
4	low	4	small	high	acc
5	med	4	small	high	acc
6	low	4	med	med	acc
7	low	2	small	low	unacc
8	vhigh	more	small	med	unacc
9	high	4	big	med	unacc
10	high	2	big	high	unacc
11	low	2	big	high	unacc

Table 5.1: Attributes of car evaluation.

(a) Write down all the parameters of the **categorical naive Bayes model with Laplace smoothing** based on the data in Table 5.1. You may leave the parameters in fractional form. (9 marks)

Solution. The posterior probability of the Naïve Bayes classifier model for the problem has the form

$$P(Y|X_1, X_2, X_3, X_4) \propto P(Y) \cdot P(X_1|Y) \cdot P(X_2|Y) \cdot P(X_3|Y) \cdot P(X_4|Y) \quad \text{[1 mark]}$$

The parameters are the prior probabilities summarised in the tables below.

Y	$P(Y)$	`maint`, $P(X_1\|Y)$ vhigh	high	med	low	`persons`, $P(X_2\|Y)$ 2	4	more
good	$\frac{3}{11}$	$\frac{0+1}{3+4}=\frac{1}{7}$	$\frac{0+1}{3+4}=\frac{1}{7}$	$\frac{1+1}{3+4}=\frac{2}{7}$	$\frac{2+1}{3+4}=\frac{3}{7}$	$\frac{0+1}{3+3}=\frac{1}{6}$	$\frac{1+1}{3+3}=\frac{2}{6}$	$\frac{2+1}{3+3}=\frac{3}{6}$
acc	$\frac{3}{11}$	$\frac{0+1}{3+4}=\frac{1}{7}$	$\frac{0+1}{3+4}=\frac{1}{7}$	$\frac{1+1}{3+4}=\frac{2}{7}$	$\frac{2+1}{3+4}=\frac{3}{7}$	$\frac{0+1}{3+3}=\frac{1}{6}$	$\frac{3+1}{3+3}=\frac{4}{6}$	$\frac{0+1}{3+3}=\frac{1}{6}$
unacc	$\frac{5}{11}$	$\frac{1+1}{5+4}=\frac{2}{9}$	$\frac{2+1}{5+4}=\frac{3}{9}$	$\frac{0+1}{5+4}=\frac{1}{9}$	$\frac{2+1}{5+4}=\frac{3}{9}$	$\frac{3+1}{5+3}=\frac{4}{8}$	$\frac{1+1}{5+3}=\frac{2}{8}$	$\frac{1+1}{5+3}=\frac{2}{8}$

.... [1+2+2=5 marks]

Y	`lugboot`, $P(X_3\|Y)$ small	med	big	`safety`, $P(X_4\|Y)$ low	med	high
good	$\frac{1+1}{3+3}=\frac{2}{6}$	$\frac{0+1}{3+3}=\frac{1}{6}$	$\frac{2+1}{3+3}=\frac{3}{6}$	$\frac{0+1}{3+3}=\frac{1}{6}$	$\frac{0+1}{3+3}=\frac{1}{6}$	$\frac{3+1}{3+3}=\frac{4}{6}$
acc	$\frac{2+1}{3+3}=\frac{3}{6}$	$\frac{1+1}{3+3}=\frac{2}{6}$	$\frac{0+1}{3+3}=\frac{1}{6}$	$\frac{0+1}{3+3}=\frac{1}{6}$	$\frac{1+1}{3+3}=\frac{2}{6}$	$\frac{2+1}{3+3}=\frac{3}{6}$
unacc	$\frac{2+1}{5+3}=\frac{3}{8}$	$\frac{0+1}{5+3}=\frac{1}{8}$	$\frac{3+1}{5+3}=\frac{4}{8}$	$\frac{1+1}{5+3}=\frac{2}{8}$	$\frac{2+1}{5+3}=\frac{3}{8}$	$\frac{2+1}{5+3}=\frac{3}{8}$

.... [1.5+1.5=3 marks]

Average: 4.95 / 9 marks in Jan 2023; 32% below 4.5 marks. □

(b) Use the parameters in part (i) to estimate the posterior probabilities of the `class` to be good, acc, and unacc given that price of maintenance is med, the capacity of persons is 4, the size of luggage boot is big and the estimated safety of the car is high. (4 marks)

Solution. From part (i), we have

$$P(Y = good|X_1 = \text{med}, X_2 = 4, X_3 = \text{big}, X_4 = \text{high}) \propto \frac{3}{11} \times \frac{2}{7} \times \frac{2}{6} \times \frac{3}{6} \times \frac{4}{6} = 0.008658009$$
$$P(Y = acc|X_1 = \text{med}, X_2 = 4, X_3 = \text{big}, X_4 = \text{high}) \propto \frac{3}{11} \times \frac{2}{7} \times \frac{4}{6} \times \frac{1}{6} \times \frac{3}{6} = 0.004329004$$
$$P(Y = unacc|X_1 = \text{med}, X_2 = 4, X_3 = \text{big}, X_4 = \text{high}) \propto \frac{5}{11} \times \frac{1}{9} \times \frac{2}{8} \times \frac{4}{8} \times \frac{3}{8} = 0.002367424$$

[3 marks]

The posterior conditional probabilities are

$$P(Y = good|X) = 0.5638767, \quad P(Y = acc|X) = 0.2819383, \quad P(Y = unacc|X) = 0.154185,$$

[1 mark]

Average: 1.5 / 4 marks in Jan 2023; 43% below 2 marks.

□

5. (Final Assessment May 2020 Q2) The testing dataset of an insurance claim is given in Table 2.1. The variables "gender", "bmi", "age_bracket" and "previous_claim" are the predictors and the "claim" is the response.

Table 2.1: The testing data of an insurance claim (randomly sampled with repeated entry).

gender	bmi	age_bracket	previous_claim	claim
female	under_weight	18-30	0	no_claim
female	under_weight	18-30	0	no_claim
male	over_weight	31-50	0	no_claim
female	under_weight	50+	1	no_claim
male	normal_weight	18-30	0	no_claim
female	under_weight	18-30	1	no_claim
male	over_weight	18-30	1	no_claim
male	over_weight	50+	1	claim
female	normal_weight	18-30	0	no_claim
female	obese	50+	0	claim

The "gender" is binary categorical data, the "bmi" is a four-value categorical data with values under_weight, normal_weight, over_weight and obese, the "age_bracket" is a three-value categorical data with value "18-30", "31-50" and "50+", the "previous_claim" is a binary categorical data with 0 indicating "no previous claim" and 1 indicating "having a previous claim". The "claim" is a binary response with values "no_claim" (negative class, with value 1) and "claim" (positive class, with value 0).

(b) Write down the mathematical formula for the Naive Bayes model with the predictors and response in Table 2.3. Use the Naive Bayes model trained on the training data from Table 2.3 to **predict** the "claim" of the insurance data in Table 2.1 as well as **evaluating** the performance of the model by calculating the confusion matrix, accuracy, sensitivity, specificity, PPV, NPV of the Naive Bayes model.

Table 2.3: The training dataset of an insurance claim data for Naive Bayes model.

Obs.	gender	bmi	age_bracket	previous_claim	claim
1	female	obese	50+	1	no_claim
2	female	under_weight	31-50	0	no_claim
3	male	under_weight	31-50	1	no_claim
4	female	over_weight	18-30	1	no_claim
5	female	normal_weight	31-50	0	no_claim
6	female	under_weight	31-50	0	no_claim
7	female	obese	18-30	0	no_claim
8	male	under_weight	50+	1	no_claim
9	female	normal_weight	31-50	0	no_claim
10	male	over_weight	31-50	0	no_claim
11	female	normal_weight	50+	0	claim
12	male	over_weight	31-50	1	claim
13	male	under_weight	31-50	1	claim
14	male	over_weight	31-50	1	claim
15	male	obese	50+	0	claim
16	male	under_weight	50+	0	claim
17	female	obese	31-50	1	claim
18	female	under_weight	50+	1	claim
19	female	normal_weight	50+	1	claim
20	female	under_weight	18-30	1	claim

Note: The default cut-off is 0.5.

Solution. Let X be the predictors; g be the predictor "gender" with F (female) and M (male); b be the predictor "bmi" with UW (under weight), OW (over weight), NW (normal weight), OB (obese); a be the predictor "age bracket" with a18 (18-30), a31 (31-50) and a50 (50+); p be the predictor "previous claim"; Y be the "actual" response "claim". The Naive Bayes model is

$$\mathbb{P}(Y|X) \propto \mathbb{P}(Y) \cdot \mathbb{P}(g|Y) \cdot \mathbb{P}(b|Y) \cdot \mathbb{P}(a|Y) \cdot \mathbb{P}(p|Y) = \text{prop.} \qquad [0.5 \text{ mark}]$$

Let $\widehat{Y}$ be the predicted response. Note that in the question, "no_claim" has a value 1 (negative) and "claim" has a value 0 (positive) which we will follow here. For the given training data, we have

$$\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = \frac{10}{20} = 0.5. \qquad [0.5 \text{ mark}]$$

Since it will not contribute to our calculation, we can actually ignore it. However, it will be maintained to match textbook algorithm.

From Table 2.1, we need to calculate

$\mathbb{P}(g = F|Y = 1) = 0.7$ $\quad$ $\mathbb{P}(g = M|Y = 1) = 0.3$
$\mathbb{P}(g = F|Y = 0) = 0.5$ $\quad$ $\mathbb{P}(g = M|Y = 0) = 0.5$
$\mathbb{P}(b = UW|Y = 1) = 0.4$ $\quad$ $\mathbb{P}(b = NW|Y = 1) = 0.2$
$\mathbb{P}(b = OW|Y = 1) = 0.2$ $\quad$ $\mathbb{P}(b = OB|Y = 1) = 0.2$
$\mathbb{P}(b = UW|Y = 0) = 0.4$ $\quad$ $\mathbb{P}(b = NW|Y = 0) = 0.2$
$\mathbb{P}(b = OW|Y = 0) = 0.2$ $\quad$ $\mathbb{P}(b = OB|Y = 0) = 0.2$
$\mathbb{P}(a = a18|Y = 1) = 0.2$ $\quad$ $\mathbb{P}(a = a31|Y = 1) = 0.6$ $\quad$ $\mathbb{P}(a = a50|Y = 1) = 0.2$
$\mathbb{P}(a = a18|Y = 0) = 0.1$ $\quad$ $\mathbb{P}(a = a31|Y = 0) = 0.4$ $\quad$ $\mathbb{P}(a = a50|Y = 0) = 0.5$
$\mathbb{P}(p = 1|Y = 1) = 0.4$ $\quad$ $\mathbb{P}(p = 0|Y = 1) = 0.6$
$\mathbb{P}(p = 1|Y = 0) = 0.7$ $\quad$ $\mathbb{P}(p = 0|Y = 0) = 0.3$

prior	$\mathbb{P}(g\|Y)$	$\mathbb{P}(b\|Y)$	$\mathbb{P}(a\|Y)$	$\mathbb{P}(p\|Y)$	prop	$\widehat{Y}$	Y
$\mathbb{P}(Y=1)=0.5$	0.7	0.4	0.2	0.6	0.0168	$\checkmark$	no_claim
$\mathbb{P}(Y=0)=0.5$	0.5	0.4	0.1	0.3	0.0030		
$\mathbb{P}(Y=1)=0.5$	0.7	0.4	0.2	0.6	0.0168	$\checkmark$	no_claim
$\mathbb{P}(Y=0)=0.5$	0.5	0.4	0.1	0.3	0.0030		
$\mathbb{P}(Y=1)=0.5$	0.3	0.2	0.6	0.6	0.0108	$\checkmark$	no_claim
$\mathbb{P}(Y=0)=0.5$	0.5	0.2	0.4	0.3	0.0060		
$\mathbb{P}(Y=1)=0.5$	0.7	0.4	0.2	0.4	0.0112		no_claim
$\mathbb{P}(Y=0)=0.5$	0.5	0.4	0.5	0.7	0.0350	$\checkmark$	
$\mathbb{P}(Y=1)=0.5$	0.3	0.2	0.2	0.6	0.0036	$\checkmark$	no_claim
$\mathbb{P}(Y=0)=0.5$	0.5	0.2	0.1	0.3	0.0015		
$\mathbb{P}(Y=1)=0.5$	0.7	0.4	0.2	0.4	0.0112	$\checkmark$	no_claim
$\mathbb{P}(Y=0)=0.5$	0.5	0.4	0.1	0.7	0.0070		
$\mathbb{P}(Y=1)=0.5$	0.3	0.2	0.2	0.4	0.0024		no_claim
$\mathbb{P}(Y=0)=0.5$	0.5	0.2	0.1	0.7	0.0035	$\checkmark$	
$\mathbb{P}(Y=1)=0.5$	0.3	0.2	0.2	0.4	0.0024		
$\mathbb{P}(Y=0)=0.5$	0.5	0.2	0.5	0.7	0.0175	$\checkmark$	claim
$\mathbb{P}(Y=1)=0.5$	0.7	0.2	0.2	0.6	0.0084	$\checkmark$	no_claim
$\mathbb{P}(Y=0)=0.5$	0.5	0.2	0.1	0.3	0.0015		
$\mathbb{P}(Y=1)=0.5$	0.7	0.2	0.2	0.6	0.0084	$\checkmark$	
$\mathbb{P}(Y=0)=0.5$	0.5	0.2	0.5	0.3	0.0075		claim

. . . . [2 marks]

From the table, the confusion matrix is as follows . . . . [0.5 mark]

	claim (0)	no_claim (1)
predict 0	1	2
predict 1	1	6

Accuracy : 0.7, Sensitivity : 0.5, Specificity : 0.75, Pos Pred Value : 0.3333, Neg Pred Value : 0.8571 . . . . [0.5 mark]

□

(c) (Ref: Tut 4 on Logistic Regression) Can we compare the logistic regression model in part (a) to the Naive Bayes model in part (b)? Can we say that the logistic regression model is better than the Naive Bayes model solely based on the performance metrics in part (a) and part (b)? Justify your answers with appropriate theory. (2 marks)

Solution. The two models cannot be compared because they are not trained with the same set of training data. . . . . [0.5 mark]

We cannot say that the logistic regression model is better because the testing data size is too small! . . . . [0.5 mark]

Theoretically, logistic regression model performs better with large number of data and the data is "linear". However, when the number of data is limited, Naive Bayes model will perform better than the logistic regression model based on Bayesian reasoning. [0.5 mark]

We need cross-validation in order to have a better understanding of the generalisation error. A single performance metric does not provide a good estimate for the generalisation error. . . . . [0.5 mark]

□

6. (Final Exam May 2023 Sem, Q4(a)) The Happiness Dataset in Table 4.1 is based on a survey conducted where people rated different metrics of their city on a scale of 5 and answered if they are happy or unhappy. The features are

- `infoavail`: the availability of information about the city services;
- `housecost`: the cost of housing;
- `schoolquality`: the overall quality of public schools.

The response, **happy**, has the values 0 (unhappy) and 1 (happy).

Obs.	`infoavail`	`housecost`	`schoolquality`	`happy`
1	5	3	3	0
2	4	5	5	0
3	4	3	3	0
4	5	2	4	0
5	1	1	1	0
6	5	2	4	1
7	5	2	4	1
8	4	2	3	1
9	3	1	2	1
10	5	5	5	1

Table 4.1: Happiness Dataset.

(i) Write down the mathematical formulation of the posterior probability and find the parameters of the **Gaussian naive Bayes model** based on the Happiness Dataset from Table 4.1. (10 marks)

Solution. Let Y denote the response `happy` and X_1, X_2, X_3 denote `infoavail`, `housecost`, `schoolquality` respectively. The mathematical formulation of posterior probability the Gaussian naive Bayes model for the Happiness Dataset from Table 4.1 is

$$P(Y=k|X_1,X_2,X_3) \propto P(Y=k)\cdot P_G(X_1|Y=k)\cdot P_G(X_2|Y=k)\cdot P_G(X_3|Y=k).$$

[1 mark]

where $k=0$ or $k=1$ and

$$P_G(X_i|Y=k) = \frac{1}{\sqrt{2\pi}\sigma_{i,k}}\exp\left(-\frac{(x-\mu_{i,k})^2}{2\sigma_{i,k}^2}\right) \qquad [0.5 \text{ mark}]$$

The probabilities and parameters are summarised in the tables below.

		`infoavail`, $P(X_1\|Y)$		`housecost`, $P(X_2\|Y)$		`schoolquality`, $P(X_3\|Y)$	
k	$P(Y=k)$	$\mu_{1,k}$	$\sigma_{1,k}$	$\mu_{2,k}$	$\sigma_{2,k}$	$\mu_{3,k}$	$\sigma_{3,k}$
0	0.5	3.8	1.6431677	2.8	1.483240	3.2	1.483240
1	0.5	4.4	0.8944272	2.4	1.516575	3.6	1.140175

. . . . [1+3+4.5=8.5 marks]

Here

$$\sigma_{1,0} = \sqrt{\frac{(5-3.8)^2+(4-3.8)^2+(4-3.8)^2+(5-3.8)^2+(1-3.8)^2}{5-1}} = \sqrt{\frac{10.8}{4}} = 1.6431677\ldots$$

□

(ii) Based on the Gaussian naive Bayes model from part (i), find the posterior probabilities for $k=0$ and $k=1$ given `infoavail` is 5, `housecost` is 4 and `schoolquality` is 4. You should round your calculations to six decimal places. (4 marks)

Solution. The products are computed as follows:

k	$P(Y=k)$	$P_G(X_1=5\|Y=k)$	$P_G(X_2=4\|Y=k)$	$P_G(X_3=4\|Y=k)$	product	posterior prob.
0	0.5	0.185959	0.193895	0.232557	0.004193	0.321845
1	0.5	0.356163	0.150783	0.329013	0.008835	0.678155
	[2 marks]				[1 mark]	[1 mark]

□

(iii) State the problem of Naive Bayes with the product of probabilities for a data of large feature space and how can we resolve this issue. (2 marks)

Solution. The problem of Naive Bayes with the product of probabilities is the product will be rounded to when the feature space is large. As can be observed from part (ii)'s calculation, with a feature space of 4 dimension, the product of probabilities get small very quickly. [1 mark]

By taking logarithm of the product of probabilities, we reduce product to sum of (negative value) exponents and avoid rounding to zero problem. [1 mark] □

7. (Final Exam Jan 2024 Sem, Q2) When a bank receives a loan application, the bank has to make a decision whether to go ahead with the loan approval or not based on the applicant's profile. Two types of risks are associated with the bank's decision:

   - If the applicant is a good credit risk, i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank;
   - If the applicant is a bad credit risk, i.e. is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank.

   To minimise loss from the bank's perspective, the bank needs a predictive model regarding who to give approval of the loan and who not to based on an applicant's demographic and socio-economic profiles.

   Suppose the response variable `Y` is 0 when the loan is approved and is 1 when the loan is not approved. Suppose the features of the data are listed below:

   - X_1 (categorical): Status of existing checking account (A11, A12, A13, A14);
   - X_2 (integer): Duration in months
   - X_3 (integer): Credit amount
   - X_4 (integer): Instalment rate in percentage of disposable income
   - X_5 (binary): foreign worker (yes, no)

   (b) When the data is trained with a naive Bayes model with Laplace smoothing, the statistical estimates below are obtained:

```
 A priori probabilities:

    0     1
0.625 0.375

 Tables:

X1             0          1
  A11 0.18518519 0.41176471
  A12 0.18518519 0.35294118
  A13 0.05555556 0.02941176
  A14 0.57407407 0.20588235

X2            0        1
  mean 18.86000 25.30000
  sd   11.29206 15.33117

X3            0        1
  mean 2940.040 3490.167
  sd   2254.614 3213.598

X4            0        1
  mean 3.060000 3.033333
  sd   1.095631 1.098065

X5             0          1
  yes 0.92307692 0.96875000
  no  0.07692308 0.03125000
```

State the naive Bayes model for this problem using conditional probabilities and estimate the posterior probabilities for $Y = 0$ and $Y = 1$ for a foreign worker when the status of existing checking account of the customer is A11, the duration is 6 months, the credit amount is 1169 and the instalment rate of disposable income is 4%. (8 marks)

Solution. The naive Bayes model for the problem with $Y = j$, where $j = 0, 1$ is [1 mark]

$$P(Y=j|X_1,X_2,X_3,X_4,X_5) \propto P(Y=j)P(X_1|Y=j)P(X_2|Y=j)P(X_3|Y=j)\times P(X_4|Y=j)P(X_5|Y=j).$$

From this model, we can build a table for the computation:

j	$P(Y=j)$	$X_1 = A11\|Y=j$	$X_2 = 6\|Y=j$	$X_3 = 1169\|Y=j$	$X_4 = 4\|Y=j$	$X_5 = yes\|Y=j$
0	0.625	0.18518519	0.0184714	12.9972×10^{-5}	0.2520039	0.92307692
1	0.375	0.41176471	0.0117817	9.5638×10^{-5}	0.2466008	0.96875000

. . . . [5 marks]

$$P(X_2 = 6|Y=0) = \frac{1}{\sqrt{2\pi}(11.29206)}\exp\left(-\frac{1}{2}\left(\frac{6-18.86}{11.29206}\right)^2\right) = 0.0184714, \quad \ldots$$

The products are

$$P(Y=0|X) \propto 6.463698 \times 10^{-8}, \quad P(Y=1|X) \propto 4.156474 \times 10^{-8}. \qquad \text{[1 mark]}$$

and the posterior probabilities are

$$P(Y=0|X) = 0.6086246, \quad P(Y=1|X) = 0.3913754 \qquad \text{[1 mark]}$$

Average: 5.24 / 8 marks in Jan 2024; 29.09% below 4 marks. □

8. (Final Exam May 2024 Sem, Q4(a)) The data in Table 4.1 describes factors influencing defect status in a manufacturing environment.

Table 4.1: Factors influencing defect status.

Obs.	EnergySupply	ProductionVolume	DefectRate	MaintenanceHours	Y
1	F	600	1.915457	4	0
2	F	659	1.841888	4	0
3	F	299	2.838841	3	0
4	F	568	1.728867	2	0
5	F	276	1.590484	23	1
6	F	492	4.670184	22	1
7	F	803	2.293886	15	1
8	F	319	4.187002	18	1
9	F	277	4.400931	1	1

The target variable in Table 4.1 is Y, the DefectStatus (0 indicates low defects while 1 indicates high defects) and the four features are

- EnergySupply: A binary feature indicating whether Green Energy (denoted by G) or Fossil-Fuel Based Energy (denoted by F) is used in the manufacturing;
- ProductionVolume: Number of units produced per day;
- DefectRate: Defects per thousand units produced;
- MaintenanceHours: Hours spent on maintenance per week.

(i) Find the parameters of the **naive Bayes model with Laplace smoothing** the data in Table 4.1 and then state the mathematical expressions of the naive Bayes model with Laplace smoothing with the information on DefectRate listed below.

```
DefectRate         0         1
      mean  2.0812632 3.4284974
      sd    0.5108489 1.3899979
```

(6 marks)

Solution. By using calculators, the following parameters can be obtained:

Y	0	1
Prior, $P(Y)$	4/9	5/9
EnergySupply=F	(4+1)/(4+2)	(5+1)/(5+2)
EnergySupply=G	(0+1)/(4+2)	(0+1)/(5+2)
ProductionVolume.mean	531.5	433.4
ProductionVolume.sd	159.5170	224.9229
MaintenanceHours.mean	3.25	15.8
MaintenanceHours.sd	0.957427	8.871302

.... [4 marks]

Let the four features be x_1 to x_4 in the order in Table 4.1. The mathematical expressions are

$$P(Y=0|x_1,...,x_4) \propto \frac{4}{9}P(x_1|Y=0)(\frac{1}{\sqrt{2\pi}(159.5170)}\exp(-\frac{(x_2-531.5)^2}{2(159.5170^2)}))$$
$$(\frac{1}{\sqrt{2\pi}(0.5108489)}\exp(-\frac{(x_3-2.0812632)^2}{2(0.5108489^2)}))$$
$$(\frac{1}{\sqrt{2\pi}(0.957427)}\exp(-\frac{(x_3-3.25)^2}{2(0.957427^2)}))$$
$$P(Y=1|x_1,...,x_4) \propto \frac{5}{9}P(x_1|Y=1)(\frac{1}{\sqrt{2\pi}(224.9229)}\exp(-\frac{(x_2-433.4)^2}{2(224.9229^2)}))$$
$$(\frac{1}{\sqrt{2\pi}(1.3899979)}\exp(-\frac{(x_3-3.4284974)^2}{2(1.3899979^2)}))$$
$$(\frac{1}{\sqrt{2\pi}(8.871302)}\exp(-\frac{(x_3-15.8)^2}{2(8.871302^2)}))$$

[2 marks]

□

(ii) Use the naive Bayes model with Laplace smoothing from part (i) to predict the posterior probability of DefectStatus to be high (i.e. $Y = 1$) for the EnergySupply of F (fossil-fuel based energy), the ProductionVolume of 260, the DefectRate of 3.239412, MaintenanceHours of 2. (5 marks)

Solution. By using appropriate scientific calculator, the following table can be constructed.

Y	$P(Y)$	$P(x_1=F\|Y)$	$P(x_2=260\|Y)$	$P(x_2=3.239412\|Y)$	$P(x_3=2\|Y)$	Product
0	4/9	5/6	0.0005876	0.059776	0.177692	2.3116 $\times 10^{-6}$
1	5/9	6/7	0.0013177	0.284366	0.013411	2.3930 $\times 10^{-6}$

.... [4 marks]

The posterior probability when $Y = 1$ given the input is

$$\frac{2.3930\times10^{-6}}{2.3116\times10^{-6}+2.3930\times10^{-6}} = 0.5086511 \qquad [1 \text{ mark}]$$

□

(iii) Evaluate the relevance of the feature EnergySupply to the naive Bayes model with Laplace smoothing with justification. (2 marks)

Solution. The feature EnergySupply is relevant to the naive Bayes model despite no G appears. .... [1 mark]

This is because in formula to obtain posterior probability

$$\frac{...P(x_1 = F|Y = 1)...}{...P(x_1 = F|Y = 0)... + ...P(x_1 = F|Y = 1)...}$$

$P(x_1 = F|Y = 0)$ and $P(x_1 = F|Y = 1)$ have different values and cancellations cannot be obtained. .... [1 mark] □