Câu 1. Cho bảng dữ liệu sau:

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

(a) Hãy tính prior và likelihood cho từng thuộc tính

(b) Hãy cho biết lớp của mẫu dữ liệu sau

Outlook	Temp	Humidity	Windy	Play
Sunny	Cool	High	True	?

c) Hãy tính phân lớp khi dữ liệu bị nhiễu

Outlook	Temp	Humidity	Windy	Play
?	Cool	High	True	?

Bài làm

a)

Outlook			**Temperature**			**Humidity**			**Windy**			**Play**	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	Hight	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	Hight	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		

Rainy 3/9 2/5	Cool 3/9 1/5				

b)

= P(Sunny|yes) x P(Cool|yes) x P(High|yes) x P(True|yes)

= 2/9 x 3/9 x 3/9 x 3/9=2/243

P(yes|X={Sunny,Cool,High,True})=

=2/243 x 9/14=1/189=0.0053

Likelihood(no) = P({Sunny,Cool,High,True}|no)

= P(Sunny|no) x P(Cool|no) x P(High|no) x P(True|no)

= 3/5 x 1/5 x 4/5 x 3/5=36/625

P(no|X={Sunny,Cool,High,True}) =P(X={Sunny,Cool,High,True}|no) x P(no)

= 36/625 x 5/14=0.026

Xác suất

P(yes)=0.0053/(0.0053+0.026)=0.205

P(no)= 0.026/(0.0053+0.026)=0.795

Kết quả: No

c)

$$P(\text{yes}|X=\{\text{Cool},\text{High},\text{True}\})=P(X=\text{Cool},\text{High},\text{True}|\text{Yes})P(\text{yes})$$
$$=\frac{3}{9}\times\frac{3}{9}\times\frac{3}{9}\times\frac{9}{14}=0,0238$$

$$P(\text{no}|X=\{\text{Cool},\text{High},\text{True}\})=P(X=\text{Cool},\text{High},\text{True}|\text{Yes})P(\text{no})$$
$$=\frac{1}{5}\times\frac{4}{5}\times\frac{3}{5}\times\frac{5}{14}=0,0343$$

$$P(\text{yes})=\frac{0,0238}{(0,0238+0,343)}=41$$

$$P(\text{no})=\frac{0,0343}{(0,0238+0,343)}=59$$

➔ P(no)>P(yes) nên kết quả là No

Câu 2. Cho bảng dữ liệu sau:

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	78	False	Yes
rain	70	96	false	yes
Rain	68	80	False	Yes
Rain	65	70	True	No
Overcast	64	65	True	Yes
Sunny	72	95	False	No
Sunny	69	70	False	Yes
Rain	75	80	False	Yes
Sunny	75	70	True	Yes
Overcast	72	90	True	Yes
Overcast	81	75	False	Yes

Rain	71	80	True	No

(a) Hãy tính prior và likelihood cho từng thuộc tính

(b) Phân lớp mẫu dữ liệu sau

Outlook	Temp	Humidity	Windy	Play
Sunny	66	90	True	?

Bài làm

a)

Outlook			**Temperature**			**Humidity**			**Windy**			**Play**	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3		83	85		86	85	False	6	2	9	5
Overcast	4	0		70	80		96	90	True	3	3		
Rainy	3	2		68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
				81			75						
Sunny	2/9	3/5	Mean	73	74.6	Mean	79.1	86.2	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Std.dev	6.2	7.9	Std.dev	10.2	9.7	True	3/9	3/5		
Rainy	3/9	2/5											

b)

$$f\left(temperature = 66 \middle| yes\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi} \times 6.2} e^{\frac{(66-73)^2}{2\times 6.2^2}} = 0,034$$

$$f\left(Humidity = 90 \middle| yes\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi} \times 10.2} e^{\frac{(90-79.1)^2}{2\times 102^2}} = 0,0221$$

$$f\left(temperature = 66 \middle| no\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi} \times 7.9} e^{\frac{(66-74.6)^2}{2\times 7.9^2}} = 0,0291$$

$$f\left(Humidity = 90 \middle| no\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi} \times 9.7} e^{\frac{(90-86.2)^2}{2\times 9.7^2}} = 0,038$$

) =

=0.000036

) =

=0.000136

P(yes)= 0.000036/(0.000036+0.000136) = 9/43

P(no)= 0. 000136/(0.000036+0. 000136) = 34/43

Câu 3. Cho bảng dữ liệu sau

	Email	w1	w2	w3	w4	w5	w6	w7	Label
Training data	E1	1	2	1	0	1	0	0	N
	E2	0	2	0	0	1	1	1	N
	E3	1	0	1	1	0	2	0	S
Test data	E4	1	0	0	0	0	0	1	?

Bài toán phân loại mail Spam (S) và Not Spam (N)

Ta có bộ training data gồm E1, E2, E3. Cần phân loại E4

Bảng từ vựng: [w1, w2, w3, w4, w5, w6, w7]

Số lần xuất hiện của từng từ trong từng email tương ứng như bảng dưới

(a) Hãy tính prior và likelihood

(b) Phân lớp cho dữ liệu E4

Bài làm

a)

Label=Spam(S)

	Email	W1	W2	W3	W4	W5	W6	W7
	E3	1	0	1	1	0	2	0
$P(w_i \mid S)$	Trước Smoothing	1/5	0/5	1/5	1/5	0/5	2/5	0/5
$P(w_i \mid S)$	Sau Smoothing	2/12	1/12	2/12	2/12	1/12	3/12	1/12

Label=Not Spam(N)

	Email	W1	W2	W3	W4	W5	W6	W7

	E1	1	2	1	0	1	0	0
	E2	0	2	0	0	1	1	1
Tổng		1	4	1	0	2	1	1
$P(w_i \mid N)$	Trước Smoothing	1/10	4/10	1/10	0/10	2/10	1/10	1/10
$P(w_i \mid N)$	Sau Smoothing	2/17	5/17	2/17	1/17	3/17	2/17	2/17

b)

P(E4 N) = P(w1 N) x P(w7 N) = 2/17 x 2/17 = 4/289 ≈ 0.013840

P(E4 S) =P (w1 S) x P(w7 S)=2/12 x 1/12 ≈0.013889

P(N E4)>P(S E4) -> Label E4: Not Spam(N)