BÀI TẬP VỀ PHÂN LỚP BAYES

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 bi nhiễu

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

BÀI LÀM

(a) Tính Prior và Likelihood: Prior: P(Yes) = 9/14, P(No) = 5/14 (14 mẫu).

* Likelihood tính cho từng thuộc tính:

Outlook			Temperature			Humidity			Windy			Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	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	High	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) Phân lớp mẫu mới: (Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = True)

P(Yes | X ={Sunny,Cool,High,True}) ~ P(X={Sunny,Cool,High,True}| Yes) x P(Yes) = 0.0053

P(No | X ={Sunny,Cool,High,True}) ~ P(X={Sunny,Cool,High,True} | No) x P(No) = 0.0206

- Likelihood(Yes) = P(X={Sunny,Cool,High,True}| Yes)

= 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

- 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

- Xác suất:

P(Yes) = 0.0053 / (0.0053 + 0.0206) = 0.205

P(No) = 0.0206 / (0.0053 + 0.0206) = 0.795

==> Kết quả phân lớp là: "No".

(c) Dữ liệu bị nhiễu: bỏ thuộc tính Outlook, chỉ sử dụng 3 thuộc tính còn lại

Likelihood(Yes) = 9/14 * 3/9 * 3/9 * 3/9 = 1/42

Likelihood(No) = 5/14 * 1/5 * 4/5 * 3/5 = 6/175

P(Yes) = (1/42)/(1/42+6/175) = 0.41

P (No) = (6/175)/(6/175 + 1/42) = 0.59

==> Kết quả vẫn là "No" dù bị nhiễu.

BÀI TẬP VỀ PHÂN LỚP BAYES

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) Prior và likelihood:14 mẫu

* Prior: P(Yes) = 9/14, P(No) = 5/14.

* Likelihood cho từng thuộc tính: dùng công thức phân phối Gaussian

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		90	90	true	3	3		
rain	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		
rain	3/9	2/5											

BÀI TẬP VỀ PHÂN LỚP BAYES

Công thức Gaussian:

$$P(x|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}}\exp\left(-\frac{(x-\mu_c)^2}{2\sigma_c^2}\right)$$

(b) Phân lớp mẫu dữ liệu:

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

$$f(Temperature = 66|yes) = \frac{1}{\sqrt{2\pi \cdot 6.2^2}}\exp(-\frac{(66-73)^2}{2 \cdot 6.2^2}) = 0.034$$

$$f(Temperature = 66|no) = \frac{1}{\sqrt{2\pi \cdot 7.9^2}}\exp(-\frac{(66-74.6)^2}{2 \cdot 7.9^2}) = 0.028$$

$$f(Humidity = 90|yes) = \frac{1}{\sqrt{2\pi \cdot 10.2^2}}\exp(-\frac{(90-79.1)^2}{2 \cdot 10.2^2}) = 0.0221$$

$$f(Humidity = 90|no) = \frac{1}{\sqrt{2\pi \cdot 9.7^2}}\exp(-\frac{(90-86.2)^2}{2 \cdot 9.7^2}) = 0.038$$

Likelihood(yes) = 2/9 * 0.034 * 0.0221* 3/9 * 9/14 = 0.000036

Likelihood(no) = 3/5 * 0.028 * 0.038 * 3/5 * 5/14 = 0.000136

P(yes) = 0.000036/(0.000036 + 0.000136) = 20.9

P(no) = 0.000136/(0.000036 + 0.000136) = 79.1

==> Kết quả phân lớp cho thuộc tính Play = "no".

Câu 3. Cho bảng dữ liệu sau:

	Email	w_1	w_2	w_3	w_4	w_5	w_6	w_7	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) Tính prior và likelihood

*Prior Probability: P(N) = 2/3, P(S) = 1/3. (3 mẫu)

- Có tổng cộng 17 từ, tổng từ trong lớp (N, S): N = 10, S = 5.

*Likelihood: có bảng từ vựng [w1, w2, w3, w4, w5, w6, w7] ==> Vocab size = 7 ==> |V| = 7. Sử dụng Laplace Smoothing với alpha = 1, ta tính được xác suất xuất hiện của từng từ trong văn bản sau:

- Lớp = N: 10 từ, sau Smoothing = 17 từ (+7 từ vocab size).

+ Ban đầu:

P(w1 | N) = 1/10; P(w2 | N) = 4/10; P(w3 | N) = 1/10; P(w4 | N) = 0/10; P(w5 | N) = 2/10;

P(w6 | N) = 1/10; P(w7 | N) = 1/10.

+ Sau Smoothing: (alpha = 1 => tử số +1)

P(w1 | N) = 2/17; P(w2 | N) = 5/17; P(w3 | N) = 2/17; P(w4 | N) = 1/17; P(w5 | N) = 3/17;

P(w6 | N) = 2/17; P(w7 | N) = 2/17.

- Lớp = S: 5 từ, sau Smoothing = 12 từ.

+ Ban đầu:

P(w1 | S) = 1/5; P(w2 | S) = 0/5; P(w3 | S) = 1/5; P(w4 | S) = 1/5; P(w5 | S) = 0/5;

P(w6 | S) = 2/5; P(w7 | S) = 0/5.

+ Sau Smoothing: (alpha = 1 => tử số +1)

P(w1 | S) = 2/12 ; P(w2 | S) = 1/12; P(w3 | S) = 2/12; P(w4 | S) = 2/12; P(w5 | S) = 1/12;

P(w6 | S) = 3/12; P(w7 | S) = 1/12.

(b) Phân loại cho dữ liệu E4

Likelihood(N) = P(N) * P(w_i | N) = P(N) * [P(w1 | N) * P(w7 | N)] = 2/3 * (2/17 * 2/17) = 0.0092

Likelihood(S) = 1/3 * (2/12 * 1/12) = 0.0046

P(N) = 0.0092/(0.0092 + 0.0046) = 0.666

P(S) = 0.0046/(0.0092 + 0.0046) = 0.334

==> Kết quả phân lớp cho E4 là Not Spam (N).

---o0o---

(Hết)