Naïve Bayes

Made Satria Wibawa, M.Eng.

2020

PENDAHULUAN

Teorema Bayesian

Teorema Bayes atau Hukum Bayes menjelaskan probabilitas dari sebuah peristiwa berdasarkan kejadian terdahulu (*prior knowledge*) yang berhubungan dengan peristiwa tersebut.

Berdasarkan probabilitas kejadian terdahulu (*prior probability*) yang terkait dengan sebuah peristiwa, kita dapat memprediksikan probabilitas sebuah peristiwa yang akan terjadi (*posterior probability*)

Bayesian

Ditemukan oleh Thomas Bayes pada tahun 1740-an

Pernah digunakan untuk menemukan ‘harta karun’ yang karam pada SS Central America (1857) oleh Tommy Thompson pada tahun 1988

Sekarang, kerap digunakan untuk filter spam dalam bentuk email

TEOREMA BAYES

Teori Probabilitas

Probability

Misalkan kita melempar 1 dadu sebanyak 1 kali, maka :

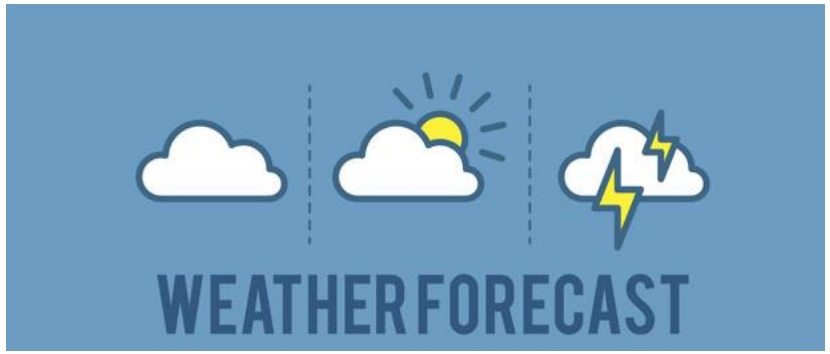
a. Peluang munculnya angka 2

- Banyaknya event/kejadian = 1
- Jumlah anggota dalam sampel = 6
- Maka probabilitasnya adalah 1/6 = 0.177

b. Peluang munculnya angka ganjil

- Banyaknya event/kejadian = 3
- Jumlah anggota dalam sampel = 6
- Maka probabilitas adalah 3/6 = 0.5

Conditional Prob.



Conditional probability adalah probabilitas kejadian A terjadi jika kejadian B juga terjadi.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$P(A|B)$ = Peluang A jika event B terjadi

$P(A \cap B)$ = Peluang A dan B terjadi secara bersamaan

$P(B)$ = Peluang B terjadi

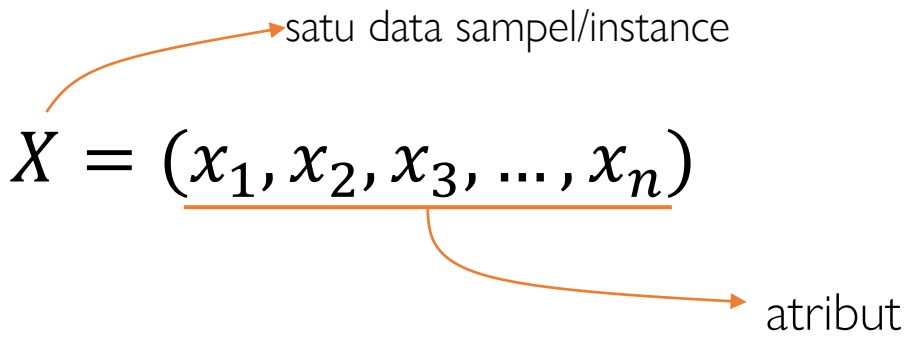
Misal: jika hari ini suhu udara sejuk, kelembapan tinggi dan kecepatan angin kencang berapakah peluang hari ini hujan?

Teorema Bayesian

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

- $P(H|X)$, peluang hipotesa H benar berdasarkan kondisi X (posterior probability)
- $P(X)$, peluang dari X yang diamati (evidence)
- $P(X|H)$, peluang X, berdasarkan kondisi pada hipotesa H (likelihood)
- $P(H)$, peluang dari hipotesa H (prior probability)

Naïve Bayes



Naïve Bayes berasumsi bahwa tiap atribut pada X saling bebas (conditionally independent) atau tidak mempengaruhi satu sama lain. Oleh karena itu disebut naïve.

Asumsi ini pada kehidupan nyata sering tidak benar. Namun, pada praktikalnya mampu memberikan prediksi yang relative akurat.

Oleh karena conditionally independent, maka:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

Prediction dengan MAP

Untuk dapat membuat prediksi dari hasil posterior probability, kita dapat menggunakan Maximum A Posteriori dengan perumusan sebagai berikut:

$$\hat{C} = \arg\max_{C} P\,(C) \prod_{k=1}^{n} P(\,x_k \mid C\,)$$

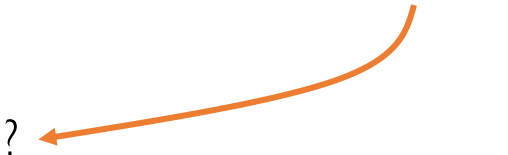
Dalam kata lain, kita memilih kelas dengan posterior probability tertinggi

CONTOH PERHITUNGAN

Naïve Bayes pada Data Kategorikal

index	Outlook	Temperature	Humidity	Windy	Play
0	overcast	hot	high	false	yes
1	overcast	cool	normal	true	yes
2	overcast	mild	high	true	yes
3	overcast	hot	normal	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	rainy	mild	normal	false	yes
8	rainy	mild	high	true	no
9	sunny	hot	high	false	no
10	sunny	hot	high	true	no
11	sunny	mild	high	false	no
12	sunny	cool	normal	false	yes
13	sunny	mild	normal	true	yes

Outlook	Temperature	Humidity	Windy	Play
Sunny	Cool	High	True	?



Tabel Likelihood

Outlook (o)		
	yes	***no***
sunny	2	3
overcast	4	0
rainy	3	2
sunny	2/9	3/5
overcast	4/9	0/5
rainy	3/9	2/5

Temperature (t)		
	yes	***no***
hot	2	2
mild	4	2
cool	3	1
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity (h)		
	yes	***no***
high	3	4
normal	6	1
high	3/9	4/5
normal	6/9	1/5

Windy (w)		
	yes	***no***
false	6	2
true	3	3
false	6/9	2/5
true	3/9	3/5

Play	
yes	***no***
9	5
9/14	5/14

Perhitungan Posterior Probability

1

$$P(C_{yes}) = \frac{9}{14} = 0.643$$

$$P(X|C_{yes}) = P(x_o|C_{yes}) \times P(x_t|C_{yes}) \times P(x_h|C_{yes}) \times P(x_w|C_{yes})$$

$$= \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = 0.008$$

2

$$P(C_{no}) = \frac{5}{14} = 0.357$$

$$P(X|C_{no}) = P(x_o|C_{no}) \times P(x_t|C_{no}) \times P(x_h|C_{no}) \times P(x_w|C_{no})$$

$$= \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} = 0.058$$

3

$$P(X) = \left(P(X|C_{yes}) \times P(C_{yes})\right) + \left(P(X|C_{no}) \times P(C_{no})\right)$$

$$= (0.008 \times 0.643) + (0.058 \times 0.357)$$

$$= 0.005 + 0.020 = 0.025$$

4

$$P(C_{yes}|X) = \frac{P(X|C_{yes})P(C_{yes})}{P(X)} = \frac{0.008 \times 0.643}{0.025} = 0.199$$

$$P(C_{no}|X) = \frac{P(X|C_{no})P(C_{no})}{P(X)} = \frac{0.058 \times 0.357}{0.025} = 0.801$$

$$C = argmax\left(\left(P(C_{yes}|X)\right), \left(P(C_{no}|X)\right)\right) = 0.801$$

Naïve Bayes pada Data Numerikal

index	Outlook	Temperature	Humidity	Windy	Play
0	overcast	83	86	false	yes
1	overcast	64	65	true	yes
2	overcast	72	90	true	yes
3	overcast	81	75	false	yes
4	rainy	70	96	false	yes
5	rainy	68	80	false	yes
6	rainy	65	70	true	no
7	rainy	75	80	false	yes
8	rainy	71	91	true	no
9	sunny	85	85	false	no
10	sunny	80	90	true	no
11	sunny	72	95	false	no
12	sunny	69	70	false	yes
13	sunny	75	70	true	yes

Outlook	Temperature	Humidity	Windy	Play
Sunny	65	80	True	?

Tabel Likelihood

Outlook (o)	yes	no
sunny	2	3
overcast	4	0
rainy	3	2
sunny	2/9	3/5
overcast	4/9	0/5
rainy	3/9	2/5

Temperature (t)	yes	no
	83	85
	70	80
	68	65
	64	72
	69	71
	75	
	75	
	72	
	81	
mean	73	74.6
std	6.2	7.9

Humidity (h)	yes	no
	86	85
	96	90
	80	70
	65	95
	70	91
	80	
	70	
	90	
	75	
mean	79.1	86.2
std	10.2	9.7

Windy (w)	yes	no
false	6	2
true	3	3
false	6/9	2/5
true	3/9	3/5

Play	yes	no
	9	5
	9/14	5/14

Probabilitas di Data Numerikal

$$P(x_i|y) = \frac{1}{(\sqrt{2\pi})\sigma} e^{-\left(\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)}$$

$$\mu = \frac{\sum x_i}{n}$$

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

Probabilitas di Data Numerikal

$$P(x_{temp}|y_{yes}) = \frac{1}{(\sqrt{2\pi})\sigma} e^{-\left(\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)} = \frac{1}{(\sqrt{2\times 3.14})6.2} e^{-\left(\left(\frac{(65-73)^2}{2\times 6.2^2}\right)\right)} = 0.0279$$

$$P(x_{temp}|y_{no}) = \frac{1}{(\sqrt{2\pi})\sigma} e^{-\left(\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)} = \frac{1}{(\sqrt{2\times 3.14})7.9} e^{-\left(\left(\frac{(65-74.6)^2}{2\times 7.9^2}\right)\right)} = 0.0241$$

$$P(x_{hum}|y_{yes}) = \frac{1}{(\sqrt{2\pi})\sigma} e^{-\left(\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)} = \frac{1}{(\sqrt{2\times 3.14})10.2} e^{-\left(\left(\frac{(80-79.1)^2}{2\times 10.2^2}\right)\right)} = 0.0389$$

$$P(x_{hum}|y_{no}) = \frac{1}{(\sqrt{2\pi})\sigma} e^{-\left(\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)} = \frac{1}{(\sqrt{2\times 3.14})9.7} e^{-\left(\left(\frac{(80-86.2)^2}{2\times 9.7^2}\right)\right)} = 0.0335$$

Perhitungan Posterior Probability

1

$$P(C_{yes}) = \frac{9}{14} = 0.643$$

$$P(X|C_{yes}) = P(x_o|C_{yes}) \times P(x_t|C_{yes}) \times P(x_h|C_{yes}) \times P(x_w|C_{yes})$$

$$= \frac{2}{9} \times 0.0279 \times 0.0241 \times \frac{3}{9} = 0.00005$$

2

$$P(C_{no}) = \frac{5}{14} = 0.357$$

$$P(X|C_{no}) = P(x_o|C_{no}) \times P(x_t|C_{no}) \times P(x_h|C_{no}) \times P(x_w|C_{no})$$

$$= \frac{3}{5} \times 0.0390 \times 0.0371 \times \frac{3}{5} = 0.0017$$

3

$$P(X) = P(X|C_{yes}) + P(X|C_{no})$$

$$= (0.00005 \times 0.643) + (0.0017 \times 0.357)$$

$$= 0.00003 + 0.0006 = 0.00063$$

4

$$P(C_{yes}|X) = \frac{P(X|C_{yes})P(C_{yes})}{P(X)} = \frac{0.00003}{0.00063} = 0.048$$

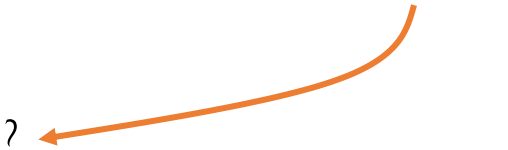
$$P(C_{no}|X) = \frac{P(X|C_{no})P(C_{no})}{P(X)} = \frac{0.0006}{0.00063} = 0.952$$

$$C = argmax\left(\left(P(C_{yes}|X)\right), \left(P(C_{no}|X)\right)\right) = 0.952$$

Problem pada Naïve Bayes

index	Outlook	Temperature	Humidity	Windy	Play
0	overcast	hot	high	false	yes
1	overcast	cool	normal	true	yes
2	overcast	mild	high	true	yes
3	overcast	hot	normal	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	rainy	mild	normal	false	yes
8	rainy	mild	high	true	no
9	sunny	hot	high	false	no
10	sunny	hot	high	true	no
11	sunny	mild	high	false	no
12	sunny	cool	normal	false	yes
13	sunny	mild	normal	true	yes

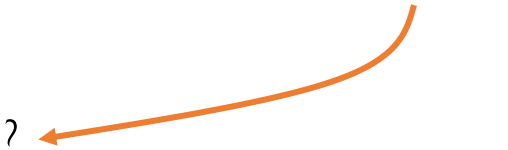
Outlook	Temperature	Humidity	Windy	Play
overcast	Cool	High	True	?



Problem pada Naïve Bayes

index	Outlook	Temperature	Humidity	Windy	Play
0	overcast	hot	high	false	yes
1	overcast	cool	normal	true	yes
2	overcast	mild	high	true	yes
3	overcast	hot	normal	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	rainy	mild	normal	false	yes
8	rainy	mild	high	true	no
9	sunny	hot	high	false	no
10	sunny	hot	high	true	no
11	sunny	mild	high	false	no
12	sunny	cool	normal	false	yes
13	sunny	mild	normal	true	yes

Outlook	Temperature	Humidity	Windy	Play
overcast	Cool	High	True	?



Tabel Likelihood

Outlook (o)		
	yes	***no***
sunny	2	3
overcast	4	0
rainy	3	2
sunny	2/9	3/5
overcast	4/9	0/5
rainy	3/9	2/5

Temperature (t)		
	yes	***no***
hot	2	2
mild	4	2
cool	3	1
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity (h)		
	yes	***no***
high	3	4
normal	6	1
high	3/9	4/5
normal	6/9	1/5

Windy (w)		
	yes	***no***
false	6	2
true	3	3
false	6/9	2/5
true	3/9	3/5

Play	
yes	***no***
9	5
9/14	5/14

pertanyaan/troubleshooting silahkan buat di channel Diskusi Teams