Teorema Bayes

Pengantar Statistik

Dosen: Moch. Zen Samsono Hadi, ST. MSc. Ph.D.

Algoritma Naïve Bayes

Likelihood

Class Prior Probability

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

Posterior Probability

Predictor Prior Probability

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

$$P(y|x_1, ..., x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

P(y|X) = P(y)P(x1|y)P(x2|y)…P(xn|y)

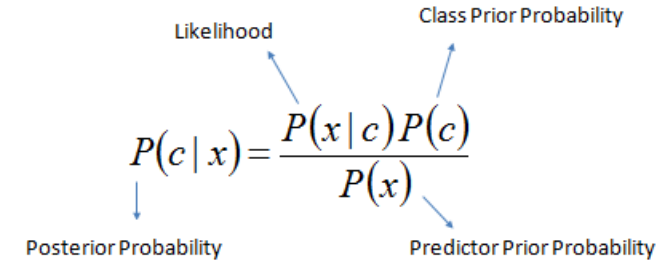
y: class variable
x: parameters/features

$$X = (x_1, x_2, x_3, ..., x_n)$$

Classifier

$$y = argmax_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

Contoh perhitungan (categorical)



Weather Dataset

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

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

$P(x \mid c) = P(Sunny \mid Yes) = 3/9 = 0.33$

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

Likelihood Table		Play Golf		
		Yes	No	
Outlook	Sunny	3/9	2/5	5/14
	Overcast	4/9	0/5	4/14
	Rainy	2/9	3/5	5/14
		9/14	5/14	

$P(x) = P(Sunny) = 5/14 = 0.36$

$P(c) = P(Yes) = 9/14 = 0.64$

$P(x \mid c) = P(Sunny \mid No) = 2/5 = 0.4$

Frequency Table		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
		9	5	14

$P(x) = P(Sunny) = 5/14 = 0.36$

$P(c) = P(No) = 5/14 = 0.36$

Tabel Likelihood (kemungkinan) untuk 4 predictor

Frequency Table

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3

Likelihood Table

		Play Golf	
		Yes	No
Outlook	Sunny	3/9	2/5
	Overcast	4/9	0/5
	Rainy	2/9	3/5

		Play Golf	
		Yes	No
Humidity	High	3/9	4/5
	Normal	6/9	1/5

		Play Golf	
		Yes	No
Temp.	Hot	2/9	2/5
	Mild	4/9	2/5
	Cool	3/9	1/5

		Play Golf	
		Yes	No
Windy	False	6/9	2/5
	True	3/9	3/5

Contoh dengan 4 predictor (sensor)

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

$$P(Yes \mid X) = P(Rainy \mid Yes) \times P(Cool \mid Yes) \times P(High \mid Yes) \times P(True \mid Yes) \times P(Yes)$$

$$P(Yes \mid X) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529 \rightarrow 0.2 = \frac{0.00529}{0.02057 + 0.00529}$$

$$P(No \mid X) = P(Rainy \mid No) \times P(Cool \mid No) \times P(High \mid No) \times P(True \mid No) \times P(No)$$

$$P(No \mid X) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.02057 \rightarrow 0.8 = \frac{0.02057}{0.02057 + 0.00529}$$

Kesimpulan: karena probability No lebih besar daripada Yes, maka disarankan tidak bermain golf.

Data kontinyu (numerical)

Menggunakan probability density function.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$ Mean

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5}$$ Standard deviation

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ Normal distribution

Contoh data kontinyu

		Humidity									Mean	StDev
Play Golf	yes	86	96	80	65	70	80	70	90	75	79.1	10.2
	no	85	90	70	95	91					86.2	9.7

$$P(\text{humidity} = 74 \mid \text{play} = \text{yes}) = \frac{1}{\sqrt{2\pi}(10.2)} e^{-\frac{(74-79.1)^2}{2(10.2)^2}} = 0.0344$$

$$P(\text{humidity} = 74 \mid \text{play} = \text{no}) = \frac{1}{\sqrt{2\pi}(9.7)} e^{-\frac{(74-86.2)^2}{2(9.7)^2}} = 0.0187$$

Kesimpulan:

Probability untuk bermain lebih besar daripada tidak, sehingga disarankan bermain.

Tugas

1. Data categorical

Handphone	Baterai	Kamera	Harga	Layak
H1	Kuat	Tinggi	Sangat murah	Ya
H2	Kuat	Tinggi	Sangat mahal	Ya
H3	Kuat	Sedang	Mahal	Ya
H4	Kuat	Rendah	Mahal	Tidak
H5	Cukup	Tinggi	Sangat murah	Ya
H6	Cukup	Sedang	Mahal	Ya
H7	Cukup	Sedang	Sangat mahal	Ya
H8	Cukup	Tinggi	Murah	Ya
H9	Cukup	Rendah	Mahal	Tidak
H10	Lemah	Tinggi	Sangat murah	Ya
H11	Lemah	Tinggi	Sangat mahal	Tidak
H12	Lemah	Sedang	Mahal	Tidak
H13	Lemah	Sedang	Murah	Tidak
H14	Lemah	Rendah	Sangat mahal	Tidak

Tentukan apakah layak direkomendasikan bila:
Baterai “Kuat”, kamera “Rendah”, Harga “Sangat Murah”?

2. Data kontinyu (numerical)

Handphone	Baterai (jam)	Kamera (Mega pixel)	Harga (juta)	Layak
H1	26	8	1.2	Ya
H2	27	13	15	Ya
H3	28	5	6	Ya
H4	25	2	5	Tidak
H5	23	10	1	Ya
H6	20	7	3.5	Ya
H7	22	7	10	Ya
H8	24	8	2	Ya
H9	21	3	4	Tidak
H10	16	13	0.8	Ya
H11	12	10	12	Tidak
H12	14	5	5	Tidak
H13	18	5	3	Tidak
H14	15	3	14	Tidak

Tentukan apakah layak direkomendasikan bila:
Baterai 28 jam, resolusi kamera 4M, harga 2 juta?

Referensi

1. https://en.wikipedia.org/wiki/Naive_Bayes_classifier
2. https://www.saedsayad.com/naive_bayesian.htm
3. https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html
4. https://informatikalogi.com/algoritma-naive-bayes/