

A.P. SHAH INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering
Data Science

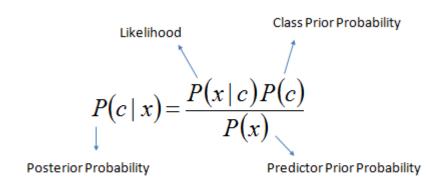


Semester: V Subject: DWM Academic Year: 2023 - 2024

Module 3

Algorithm

Bayes theorem provides a way of calculating the posterior probability, P(c/x), from P(c), P(x), and P(x/c). Naive Bayer classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.



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

- P(c/x) is the posterior probability of *class* (target) given *predictor* (attribute).
- P(c) is the prior probability of *class*.
- P(x/c) is the likelihood which is the probability of *predictor* given *class*.
- P(x) is the prior probability of *predictor*.

In ZeroR model there is no predictor, in OneR model we try to find the single best predictor, naive Bayesian includes all predictors using Bayes' rule and the independence assumptions between predictors.

Example 1:

We use the same simple Weather dataset here.

PARSHWANATH CHARITABLE TRUST'S



A.P. SHAH INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering Data Science



Semester: V Subject: DWM Academic Year: 2023 - 2024

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

The posterior probability can be calculated by first, constructing a frequency table for each attribute against the target. Then, transforming the frequency tables to likelihood tables and finally use the Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.



A.P. SHAH INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering Data Science

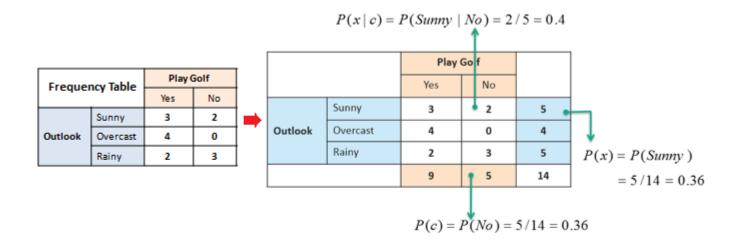


Semester: V Subject: DWM Academic Year: 2023 - 2024

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

							1						
					Likalih	ood Table	Play	Golf					
Frequency Table		Play Golf			Likelii	lood lable	Yes	No					
		Yes	No			Sunny	3/9	2/5	5/14				
	Sunny	3	2	 		7 77	0/5	4/14					
Outlook	Overcast	4	0								Rainy	2/9	3/5
	Rainy	2	3			Halliy			3/14	P(x) = P(Sunny)			
							9/14	5/14		= 5/14 = 0.36			
						P(c)	= P(Vos)) = 9/14 -	- 0 64				

Posterior Probability: $P(c \mid x) = P(Yes \mid Sunny) = 0.33 \times 0.64 \div 0.36 = 0.60$



Posterior Probability: $P(c \mid x) = P(No \mid Sunny) = 0.40 \times 0.36 \div 0.36 = 0.40$

The likelihood tables for all four predictors.



A.P. SHAH INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering Data Science



Semester: V Subject: DWM Academic Year: 2023 - 2024

Frequency Table

Likelihood Table

		,						_	
		Play Golf]				Play Golf	
		Yes	No]			Yes	No	
	Sunny	3	2	\Longrightarrow		Sunny	3/9	2/5	
Outlook	Overcast	4	0		Outlook	Overcast	4/9	0/5	
	Rainy	2	3			Rainy	2/9	3/5	
		Play	Golf				Play Golf		
		Yes	No				Yes	No	
н Н	High	3	4		Humidity	High	3/9	4/5	
Humidity	Normal	6	1			Normal	6/9	1/5	
				-					
		Play	Golf				Play	Golf	
		Yes	No				Yes	No	
Hot		2	2	\implies		Hot	2/9	2/5	
Temp.	Mild	4	2] ,	Temp.	Mild	4/9	2/5	
	Cool	3	1			Cool	3/9	1/5	
I				I	I				

		Play Golf					Play	Golf
		Yes	No				Yes	No
Windy False		6	2	$\qquad \qquad \longrightarrow \qquad \qquad \\$	Mindu	False	6/9	2/5
winay	True	3	3		Windy	True	3/9	3/5

Example 2:

In this example we have 4 inputs (predictors). The final posterior probabilities can be standardized between 0 and 1

Outlook 1	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

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

$$P(Yes \mid X) = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.00529$$

$$0.2 = \frac{0.00529}{0.02057 + 0.00529}$$

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

$$P(No \mid X) = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.02057$$

$$0.8 = \frac{0.02057}{0.02057 + 0.00529}$$

PARSHWANATH CHARITABLE TRUST'S



A.P. SHAH INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering Data Science



Semester: V Subject: DWM Academic Year: 2023 - 2024

The zero-frequency problem

Add 1 to the count for every attribute value-class combination (*Laplace estimator*) when an attribute value (*Outlook=Overcast*) doesn't occur with every class value (*Play Golf=no*).