Naïve Bayes Algorithm

How the Naïve Bayes Classifier works in Machine Learning

Sara M. Steinel
March 22, 2018

Definition

- Naïve Bayes is a popular machine learning algorithm that uses statistical modeling to tackle classification problems
  - Classifies new observations based on observations in the past
- Machine learning is what enables a system to automatically learn and improve from experience without being explicitly programmed
- Classification is the problem of identifying to which category a new observation belongs based on past data

How does it work?

- Fundamental assumption of Bayes' Theorem: all predictors are independent from each other
  - knowing values of one attribute (predictor) does not tell us values of another
  - Not how things work in real life
- Algorithm studies training data
  - Does not store training data
  - Learns from training data and classification model adapt

Example

Step 1 - Raw Training Data

The Golf data set

play	outlook	temperature	humidity	windy
no	sunny	hot	high	false
no	sunny	hot	high	true
yes	overcast	hot	high	false
yes	rainy	mild	high	false
yes	rainy	cool	normal	false
no	rainy	cool	normal	true
yes	overcast	cool	normal	true
no	sunny	mild	high	false
yes	sunny	cool	normal	false
yes	rainy	mild	normal	false
yes	sunny	mild	normal	true
yes	overcast	mild	high	true
yes	overcast	hot	normal	false
no	rainy	mild	high	true

Example

Step 1 - Raw Training Data

Class

Predictors

play	outlook	temperature	humidity	windy
no	sunny	hot	high	false
no	sunny	hot	high	true
yes	overcast	hot	high	false
yes	rainy	mild	high	false
yes	rainy	cool	normal	false
no	rainy	cool	normal	true
yes	overcast	cool	normal	true
no	sunny	mild	high	false
yes	sunny	cool	normal	false
yes	rainy	mild	normal	false
yes	sunny	mild	normal	true
yes	overcast	mild	high	true
yes	overcast	hot	normal	false
no	rainy	mild	high	true

Example

Step 2 - Build a predictor-frequency table

- Construct a frequency table for each attribute against the target
- Purpose: record how many times a predictor occurs for each class

Frequency Tables

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

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

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

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

Example

Step 3 - Calculate predictor-class percentages

$P(x|c) = P(Sunny \mid Yes) = 2/9 = .22$

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

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

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

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

Theorem formula

- Goal: Calculate posterior probability
- Class with highest posterior probability is outcome of prediction

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

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

- 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.

Example

Step 3 - Calculate predictor-class percentages

$P(x|c) = P(Sunny \mid Yes) = 2/9 = .22$

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

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

$P(x) = P(Sunny)$

$= 5/14 = 0.36$

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

P(c|x) = P(Yes | Sunny) = .22 x .64 / .36 = .39

Let’s try it

Let’s say we have a new day which is:

- Sunny
- Cool
- High Humidity
- Windy

What is the likelihood for playing golf on this day?

Likelihood calculated by multiplying ratios which correspond to the combination of those weather attributes with “yes”

Let's try it

Likelihood of yes: 2/9 * 3/9 * 3/9 * 3/9 * 9/14 = 0.0053

	Outlook			Temperature			Humidity			Windy			Play	
Counts		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								
Ratios	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								

Likelihood of no: 3/5 * 1/5 * 4/5 * 3/5 * 5/14 = 0.0206

	Outlook			Temperature			Humidity			Windy			Play	
Counts		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								
Ratios	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								

Since 0.0053 < 0.0206, prediction is no

Converting likelihood to true probabilities

Likelihood of yes: 0.0053

Likelihood of no: 0.0206

Normalization:

Probability of yes: (0.0053 / (0.0053 + 0.0206)) * 100 = 21.5%

Probability of no: 0.0206 / ((0.0206 + 0.0053)) * 100 = 79.5%

Our research

- Our research involves big data
- We will use the Naïve Bayes Algorithm to classify new observations
- Recall: This algorithm is for classifying new observations into a category using past data
- We are trying to find possible links to mental health problems
- We look at many different predictors, including drug use, alcohol use, exercise, grades, etc.