Naive Bayes Algorithm

20CP401T

Himanshu K. Gajera
Department of Computer Science & Engineering
Pandit Deendayal Energy University, Gandhinagar

Introduction

- This model is easy to build and is mostly used for large datasets.
- It is a probabilistic machine learning model that is used for classification problems.
- The core of the classifier depends on the Bayes theorem with an assumption of independence among predictors.
- That means changing the value of a feature doesn’t change the value of another feature.

Why is it called Naive?

- It is called Naive because of the assumption that 2 variables are independent when they may not be. In a real-world scenario, there is hardly any situation where the features are independent.

- Naive Bayes does seem to be a simple yet powerful algorithm.
- But why is it so popular?

- Since it is a probabilistic approach, the predictions can be made real quick. It can be used for both binary and multi-class classification problems.

- Need to understand what is “Conditional probability”, what is “Bayes’ theorem” and how conditional probability help’s us in Bayes’ theorem.

Conditional Probability for Naive Bayes

- Let’s start with examples.

- Suppose I ask you to pick a card from the deck and find the probability of getting a king given the card is clubs.

- Observe carefully that here I have mentioned a condition that the card is clubs.

- Now while calculating the probability my denominator will not be 52, instead, it will be 13 because the total number of cards in clubs is 13.

- Since we have only one king in clubs the probability of getting a KING given the card is clubs will be 1/13 = 0.077.

Conditional Probability for Naive Bayes

- Let’s take one more example,
- Consider a random experiment of tossing 2 coins. The sample space here will be: S = {HH, HT, TH, TT}
- If a person is asked to find the probability of getting a tail his answer would be 3/4 = 0.75

- Now suppose this same experiment is performed by another person but now we give him the condition that both the coins should have heads. This means if event A: ‘Both the coins should have heads’, has happened then the elementary outcomes {HT, TH, TT} could not have happened. Hence in this situation, the probability of getting heads on both the coins will be 1/4 = 0.25

- From the above examples, we observe that the probability may change if some additional information is given to us. This is exactly the case while building any machine learning model, we need to find the output given some features.

Conditional Probability for Naive Bayes

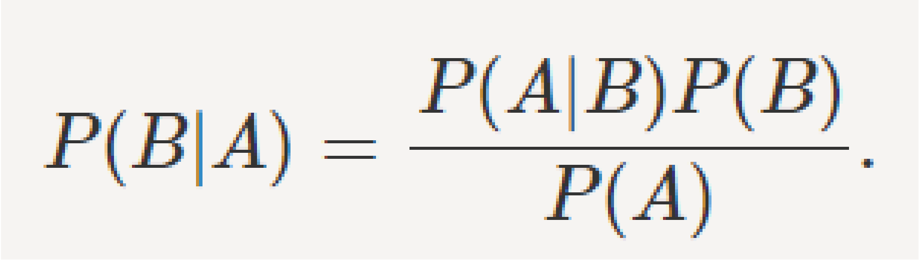
- Mathematically, the conditional probability of event A given event B has already happened is given by:

$$\underset{\text{Probability of } A \text{ given } B}{P(A\,|\,B)} = \frac{\overset{\text{Probability of } A \text{ and } B}{P(A \cap B)}}{\underset{\text{Probability of } B}{P(B)}}$$

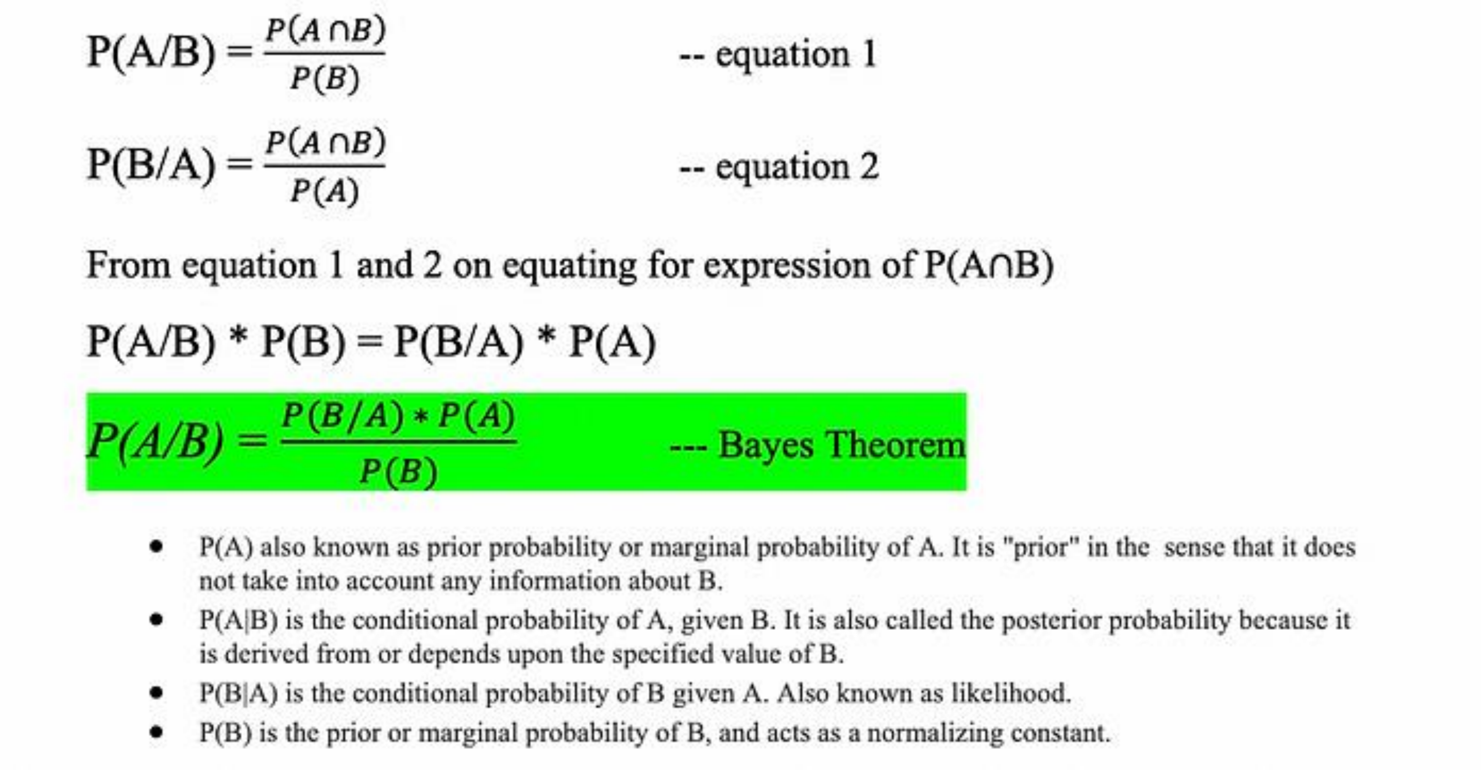
Bayes’ Rule

- Bayes’ theorem which was given by Thomas Bayes, a British Mathematician, in 1763 provides a means for calculating the probability of an event given some information.

- Mathematically Bayes’ theorem can be stated as:



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



$$P(A/B) = \frac{P(A \cap B)}{P(B)} \qquad \text{-- equation 1}$$

$$P(B/A) = \frac{P(A \cap B)}{P(A)} \qquad \text{-- equation 2}$$

From equation 1 and 2 on equating for expression of P(A∩B)

$$P(A/B) * P(B) = P(B/A) * P(A)$$

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)} \qquad \text{--- Bayes Theorem}$$

- P(A) also known as prior probability or marginal probability of A. It is "prior" in the sense that it does not take into account any information about B.
- P(A|B) is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.
- P(B|A) is the conditional probability of B given A. Also known as likelihood.
- P(B) is the prior or marginal probability of B, and acts as a normalizing constant.

Bayes’ Rule

- Trying to find the probability of event A, given event B is true.
- Here P(B) is called prior probability which means it is the probability of an event before the evidence
- P(B|A) is called the posterior probability i.e., Probability of an event after the evidence is seen.
- With regards to dataset, this formula can be re-written as:

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

Y: class of the variable
X: dependent feature vector (of size n)

What is Naive Bayes?

- Bayes’ rule provides us with the formula for the probability of Y given some feature X.
- In real-world problems, we hardly find any case where there is only one feature.

- When the features are independent, we can extend Bayes’ rule to what is called Naive Bayes which assumes that the features are independent that means changing the value of one feature doesn’t influence the values of other variables and this is why we call this algorithm “NAIVE”

- Naive Bayes can be used for various things like face recognition, weather prediction, Medical Diagnosis, News classification, Sentiment Analysis, and a lot more.

What is Naive Bayes?

- When there are multiple X variables, we simplify it by assuming that X’s are independent, so

$$P(Y=k|X) = \frac{P(X|Y=k) * P(Y=k)}{P(X)}$$

- For n number of X, the formula becomes Naive Bayes:

$$P(Y=k|X1, X2....Xn) = \frac{P(X1|Y=k) * P\big(X2|Y=k\big)...... * P\big(Xn|Y=k\big) * P(Y=k)}{P(X1) * P(X2).... * P(Xn)}$$

- Which can be expressed as:

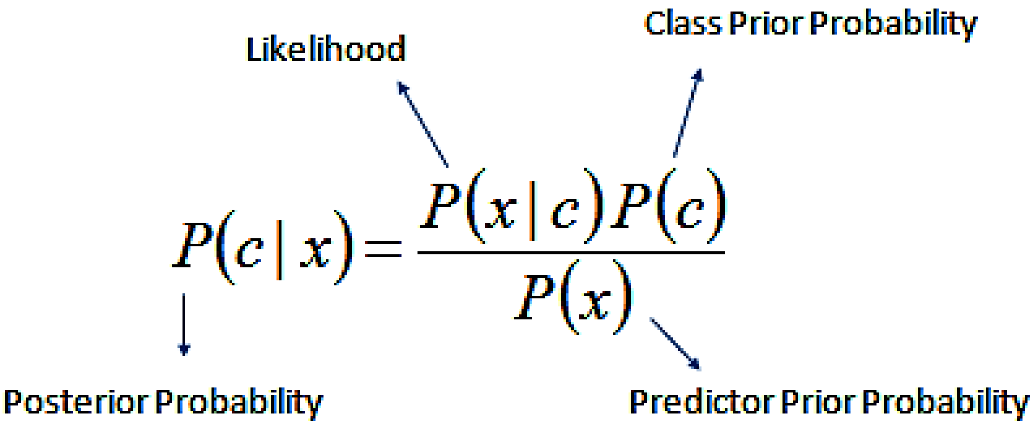
$$P\big(Y=k|X1, X2....Xn\big) = \frac{\mathrm{P(Y)} \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{P(Xi|Y)}}{P(X1) * P(X2).... * P(Xn)}$$

What is Naive Bayes?

- Since the denominator is constant here so we can remove it. It’s purely your choice if you want to remove it or not. Removing the denominator will help you save time and calculations.

$$P(Y = k|X1, X2....Xn) \propto \mathrm{P(Y)} \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P(Xi|Y)}$$

- This formula can also be understood as:



Likelihood

Class Prior Probability

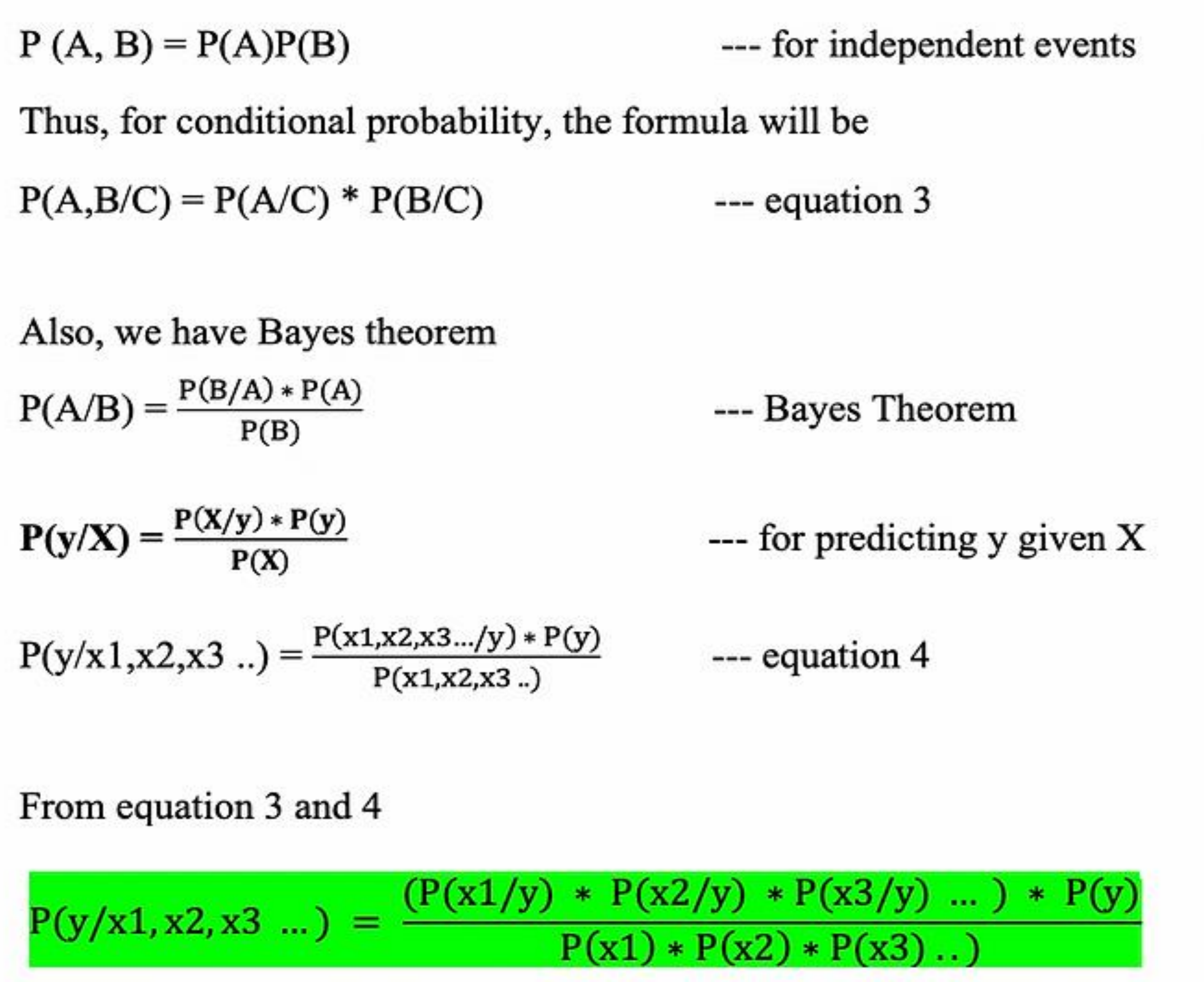
$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

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

Bayes’ Rule



$P\ (A, B) = P(A)P(B)$ --- for independent events

Thus, for conditional probability, the formula will be

$P(A,B/C) = P(A/C) * P(B/C)$ --- equation 3

Also, we have Bayes theorem

$P(A/B) = \frac{P(B/A) * P(A)}{P(B)}$ --- Bayes Theorem

$\mathbf{P(y/X) = \frac{P(X/y) * P(y)}{P(X)}}$ --- for predicting y given X

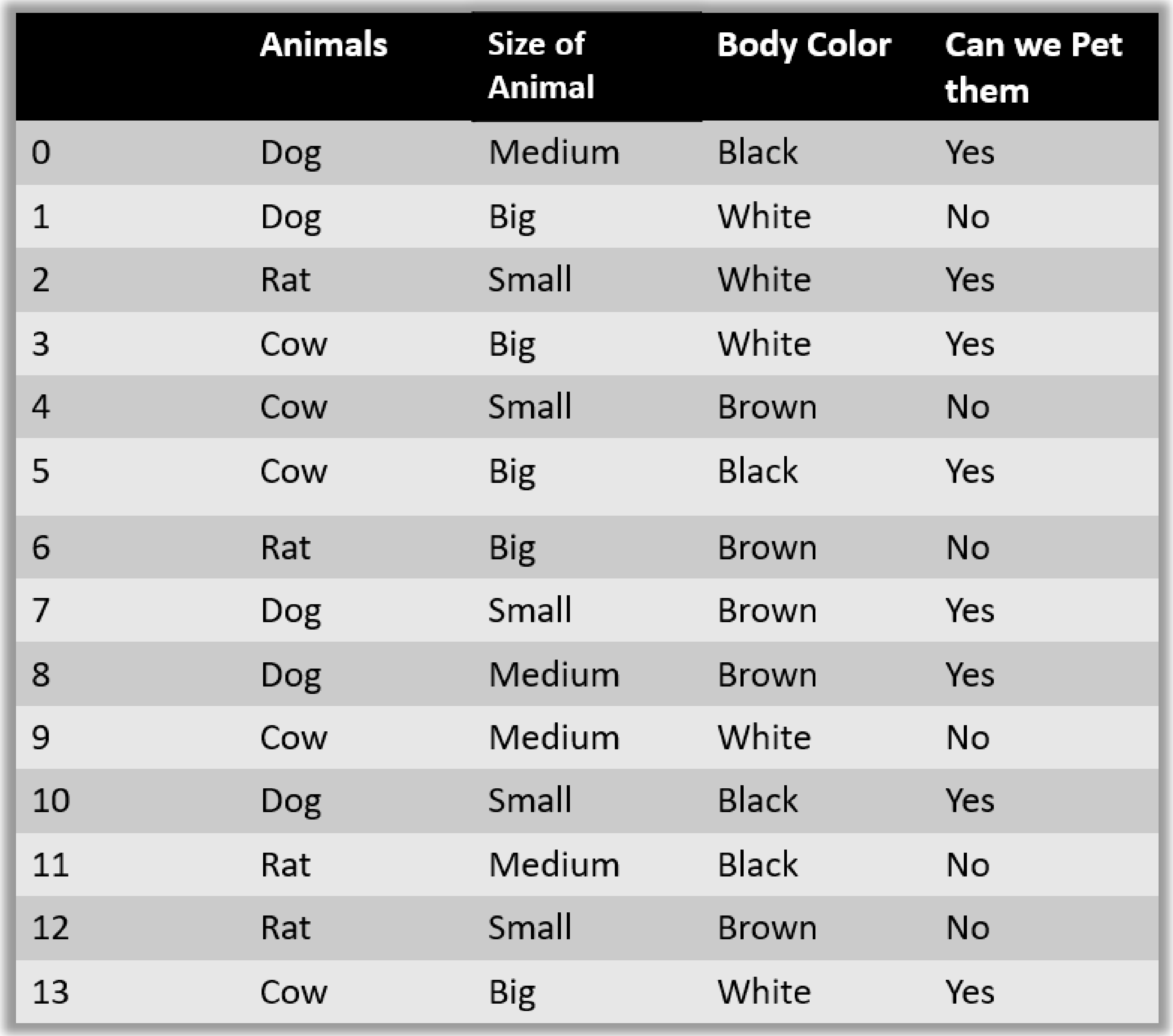
$P(y/x1,x2,x3\ ..) = \frac{P(x1,x2,x3.../y) * P(y)}{P(x1,x2,x3\ ..)}$ --- equation 4

From equation 3 and 4

$$P(y/x1, x2, x3\ ...) = \frac{(P(x1/y)\ *\ P(x2/y)\ * P(x3/y)\ ...\ )\ *\ P(y)}{P(x1) * P(x2) * P(x3)\ ..)}$$

Naive Bayes Example

Let’s take a dataset to predict whether we can pet an animal or not.

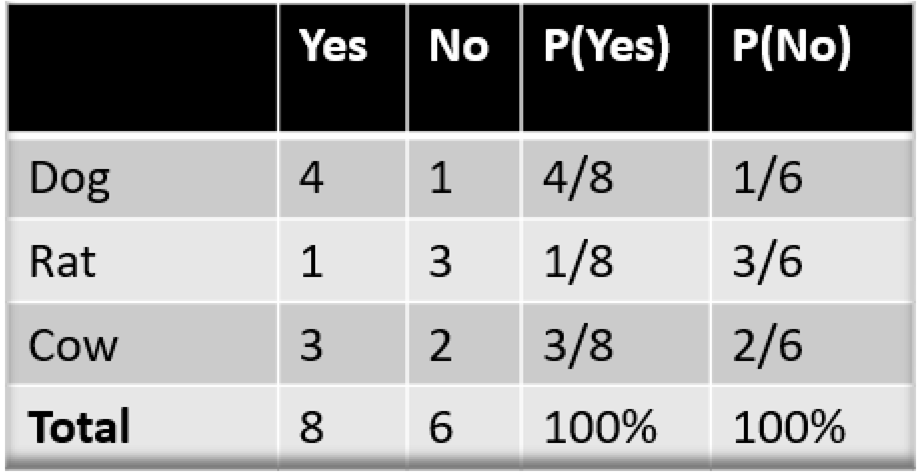


	Animals	Size of Animal	Body Color	Can we Pet them
0	Dog	Medium	Black	Yes
1	Dog	Big	White	No
2	Rat	Small	White	Yes
3	Cow	Big	White	Yes
4	Cow	Small	Brown	No
5	Cow	Big	Black	Yes
6	Rat	Big	Brown	No
7	Dog	Small	Brown	Yes
8	Dog	Medium	Brown	Yes
9	Cow	Medium	White	No
10	Dog	Small	Black	Yes
11	Rat	Medium	Black	No
12	Rat	Small	Brown	No
13	Cow	Big	White	Yes

Bayes’ Rule

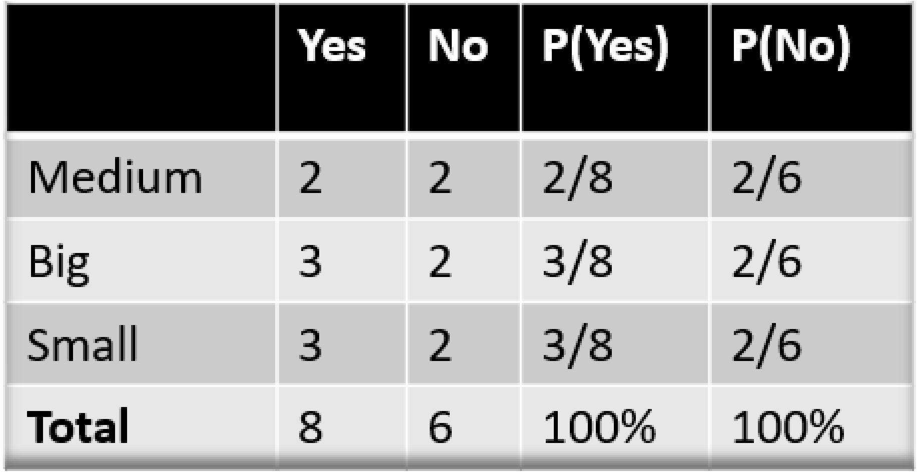
We need to find P(xi|yj) for each xi in X and each yj in Y. All these calculations have been demonstrated below:

Animals



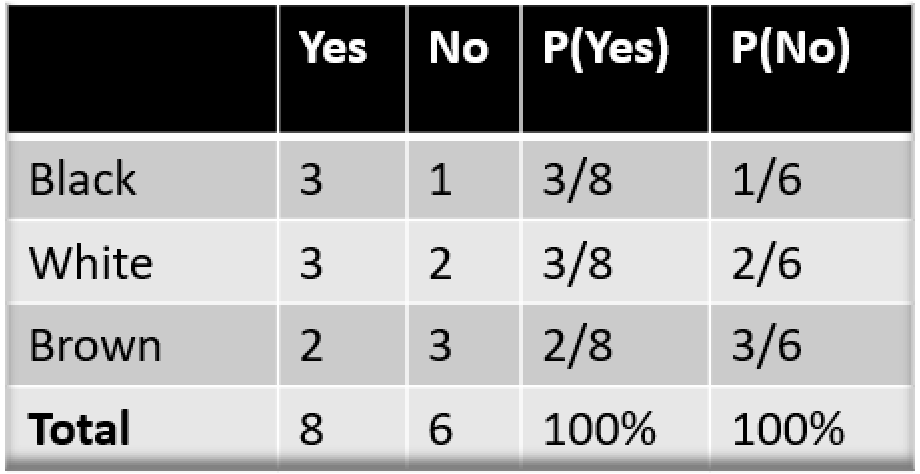
	Yes	No	P(Yes)	P(No)
Dog	4	1	4/8	1/6
Rat	1	3	1/8	3/6
Cow	3	2	3/8	2/6
Total	8	6	100%	100%

Size of Animal



	Yes	No	P(Yes)	P(No)
Medium	2	2	2/8	2/6
Big	3	2	3/8	2/6
Small	3	2	3/8	2/6
Total	8	6	100%	100%

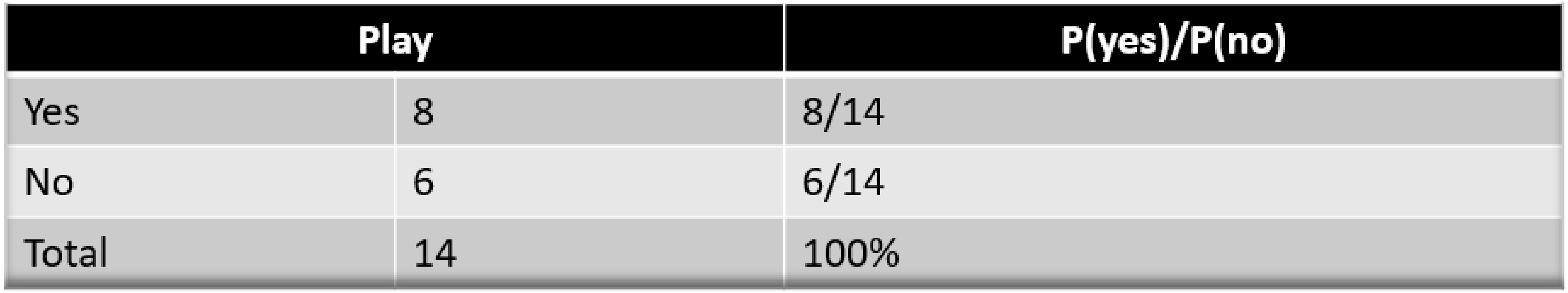
Body Color



	Yes	No	P(Yes)	P(No)
Black	3	1	3/8	1/6
White	3	2	3/8	2/6
Brown	2	3	2/8	3/6
Total	8	6	100%	100%

Bayes’ Rule

We also need the probabilities (P(y)), which are calculated in the table below. For example, P(Pet Animal = NO) = 6/14.



Play		P(yes)/P(no)
Yes	8	8/14
No	6	6/14
Total	14	100%

Now if we send our test data, suppose **test = (Cow, Medium, Black)**
Probability of petting an animal :

$$P(Yes|Test) = \frac{P(Animal = Cow|Yes) * P(Size = Medium|Yes) * P(Color = Black|Yes) * P(Yes)}{P(Test)}$$

$$P(Yes|Test) = \frac{3}{8} * \frac{2}{8} * \frac{3}{8} * \frac{8}{14} = 0.0200$$

Bayes' Rule

And the probability of not petting an animal:

$$P(No|Test) = \frac{P(Animal = Cow|No) * P(Size = Medium|No) * P(Color = Black|No) * P(No)}{P(Test)}$$

$$P(No|Test) = \frac{2}{6} * \frac{2}{6} * \frac{1}{6} * \frac{6}{14} = 0.0079$$

We know P(Yes|Test)+P(No|test) = 1, So, normalize the result:

$$\mathrm{P(Yes|Test)} = \frac{0.0200}{0.0200 + 0.0079} = 0.7168$$

$$\mathrm{P(No|Test)} = \frac{0.0079}{0.0079 + 0.0200} = 0.2831$$

We see here that P(Yes|Test) > P(No|Test),
so the prediction that we can pet this animal is **“Yes”**.

Gaussian Naive Bayes

- Gaussian Naïve Bayes is used when we assume all the continuous variables associated with each feature to be distributed according to Gaussian Distribution. Gaussian Distribution is also called Normal distribution.

- The conditional probability changes here since we have different values now. Also, the (PDF) probability density function of a normal distribution is given by:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

- We can use this formula to compute the probability of likelihoods if our data is continuous.