Bayes Theorem

Bayes Theorem

CONDITIONAL PROBABILITY

$$P(\text{Head}) = \frac{\text{\# Favourable Outcomes}}{\text{\# Total Outcomes}}$$

$$= \frac{50}{100} = 0.5$$

$$P(\text{Play}) = \frac{20}{30} \approx 66\%$$

$$P(\text{Play}|\text{Rain}) = \frac{1}{10} \approx 10\%$$

Bayes Theorem

	Play Golf	Not Playing Golf	Total
Rains	10	2	12
No Rain	50	38	88
	60	40	100

Bayes Theorem

JOINT PROBABILITY AND CONDITIONAL PROBABILITY

	A = Play	A = No Play	Total
B = Rain	10	2	12
B = No Rain	50	38	88
	60	40	100

Bayes Theorem

JOINT PROBABILITY AND CONDITIONAL PROBABILITY

	A = Play	A = No Play	Total
B = Rain	10	2	12
B = No Rain	50	38	88
	60	40	100

A = Play   B = Rain

Prior

$$P(A) = \frac{60}{100} \qquad P(B) = \frac{12}{100}$$

Bayes Theorem

Conditional probability

$$P(A|B) = \frac{10}{12} = \frac{10/100}{12/100} = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A|B).P(B)$$

$$= \frac{10}{12} \times \frac{12}{100} = \frac{10}{100}$$

	A = Play	A = No Play	Total
B = Rain	10	2	12
B = No Rain	50	38	88
	60	40	100

BAYES THEOREM

$$P(A \cap B) = \underset{\uparrow}{P(A|B)} \; . \; \underset{\uparrow}{P(B)} = \frac{10}{100}$$

$$\frac{10}{12} \times \frac{12}{100}$$

$$P(A \cap B) = \underset{\uparrow}{P(B|A)} \; . \; \underset{\uparrow}{P(A)} = \frac{10}{100}$$

$$\frac{10}{60} \times \frac{60}{100}$$

$$P(A|B) \; . \; P(B) = P(B|A) \; . \; P(A)$$

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

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

Naïve Bayes Theorem

Naïve Bayes Theorem

Naïve Bayes is a probabilistic classifier that returns the probability of a test point belonging to a class, using Bayes' theorem.

Byes Theorem

$$P(Ci / xj) = \frac{P(xj / Ci)\ P(Ci)}{P(xj)}$$

Naïve Bayes Theorem

Naïve Bayes is a probabilistic classifier that returns the probability of a test point belonging to a class, using Bayes' theorem. As you learned previously, Bayes' theorem is defined as —

$P(C_i/X) = P(X/C_i) \times P(C_i)/P(X)$, where C_i denotes the classes, and X denotes the features of the data point.

Probabilities are calculated simply by counting the number of instances/occurrences for categorical data.

The effect of the denominator P(x) is not incorporated while calculating probabilities as it is the same for both the classes and hence, can be ignored without affecting the final outcome.

The class assigned to the new test point is the class for which $P(C_i/X)$ is greater.

Naïve Bayes Theorem

S.No	Type of mushroom	Cap shape
1.	Poisonous	Convex
2.	Edible	Convex
3.	Poisonous	Convex
4.	Edible	Convex
5.	Edible	Convex
6.	Poisonous	Convex
7.	Edible	Bell
8.	Edible	Bell
9.	Edible	Convex
10.	Poisonous	Convex
11.	Edible	Flat
12.	Edible	Bell

Naïve Bayes Theorem

$$P(C = \text{edible} \,/\, x = \text{convex}) = \frac{P(x=\text{convex} \,/\, C = \text{edible}) \times P(C = \text{edible})}{P(x)}$$

$$P(C = \text{poisonous} \,/\, x = \text{convex}) = \frac{P(x=\text{convex} \,/\, C = \text{poisonous}) \times P(C = \text{poisonous})}{P(x)}$$

We can ignore the denominator as it is same for both the probabilities and we would compare the probabilities

Naïve Bayes Theorem

P (Mushroom is edible) = 7 /14 = 0.5

P (Mushroom is poisonous) = 7 /14 = 0.5

P(cap-shape = convex / edible = yes) = 4 /7

P(cap-shape = convex / poisonous = yes) = 3/7

Naïve Bayes Theorem

P(edible = yes / x= convex) = P(x=convex/edible = yes) X P(edible=yes)
= 4 / 7 x 1 / 2
= 4 / 14

P(poisonous = yes / x= convex) = P(x=convex/ poisonous = yes) X P(poisonous =yes)
= 3 / 7 x 1 / 2
= 3 / 14

So test point is classified as Edible

Naïve Bayes Theorem

S.No	Type of mushroom	Cap shape
1.	Poisonous	Convex
2.	Edible	Convex
3.	Poisonous	Convex
4.	Edible	Convex
5.	Edible	Convex
6.	Poisonous	Convex
7.	Edible	Bell
8.	Edible	Bell
9.	Edible	Convex
10.	Poisonous	Convex
11.	Edible	Flat
12.	Edible	Bell

Naïve Bayes Theorem

The probabilities of a CONVEX mushroom being edible and poisonous are both 50%. The probability of a mushroom being edible, P(C = edible | X = CONVEX) is

P(X = CONVEX | C = edible) . P(C = edible) / P(X = CONVEX)
= (4/8).(8/12) / (8/12)
= 50%

Similarly, the probability of the mushroom being poisonous, P(C = poisonous| X = CONVEX) is

= P(X = CONVEX | C = poisonous) . P(C = poisonous) / P(X = CONVEX)
= (4/4).(4/12) / (8/12)
= 50%

Note that the denominator is common in both calculations, i.e. P(X = CONVEX) = 8/12, and thus you do not need to calculate it. You can simply compare the numerators and conclude the classes based on that:

Edible: P(X = CONVEX | C = edible) . P(C = edible) = (4/8).(8/12) = 4/12 = 33.33%
Poisonous: P(X = CONVEX | C = poisonous) . P(C = poisonous) = (4/4).(4/12) = 4/12 = 33.33%

Naïve Bayes Theorem

Let's now break down the Bayes theorem. The 50% probability that the CONVEX mushroom is edible (or poisonous) is a result of three probabilities. P(edible | CONVEX) is:

1) Proportional to P(edible), which tells us how abundant edible mushrooms are; if P(edible) is high, then P(edible | CONVEX) will be high simply because edible mushrooms are abundant!

    P(edible) is 66.66% and P(poisonous) is 33.33 %

    This pushes the favor towards edible since they are in abundance

2) Proportional to P(CONVEX | edible), which explains how likely you are to find a CONVEX mushroom if you separately consider all the edible ones;

    P(CONVEX | edible) is 50% and P(CONVEX | poisonous) is 100%

    This pushes the favor towards poisonous since all poisonous mushrooms are CONVEX

3) Inversely proportional to P(CONVEX); this term cancels out while comparing the two classes

Naïve Bayes Theorem

S.No	**Type of Mushroom**	**Cap shape**	**Cap surface**
1.	**Poisonous**	Convex	Scaly
2.	Edible	Convex	Scaly
3.	**Poisonous**	Convex	Smooth
4.	Edible	Convex	Smooth
5.	Edible	Convex	Fibrous
6.	**Poisonous**	Convex	Scaly
7.	Edible	Bell	Scaly
8.	Edible	Bell	Scaly
9.	Edible	Convex	Scaly
10.	**Poisonous**	Convex	Scaly
11.	Edible	Flat	Scaly
12.	Edible	Bell	Smooth

Applying Bayes Theorem

Assumption – cap surface and cap shape are conditionally independent

The expression is:

P(edible = yes / x=(convex, smooth)) = P(x=convex/edible) X P(x=smooth/edible) X P(edible)

Naïve Bayes Theorem

S.No	**Type of Mushroom**	**Cap shape**	**Cap surface**
1.	**Poisonous**	Convex	Scaly
2.	Edible	Convex	Scaly
3.	**Poisonous**	Convex	Smooth
4.	Edible	Convex	Smooth
5.	Edible	Convex	Fibrous
6.	**Poisonous**	Convex	Scaly
7.	Edible	Bell	Scaly
8.	Edible	Bell	Scaly
9.	Edible	Convex	Scaly
10.	**Poisonous**	Convex	Scaly
11.	Edible	Flat	Scaly
12.	Edible	Bell	Smooth

P(cap-surface = smooth / edible = yes) = 3/7

P(cap-surface = smooth / poisonous = yes) = 2/7

Naïve Bayes Theorem

P(edible / x = (convex, smooth))

= P(x=convex, smooth/edible) X P(edible)

= P(x=convex/edible) X P(x=smooth/edible) X P(edible)

= 4 / 7 x 3 / 7 x 1 / 2 = 12/98

P(poisonous / x = (convex, smooth))

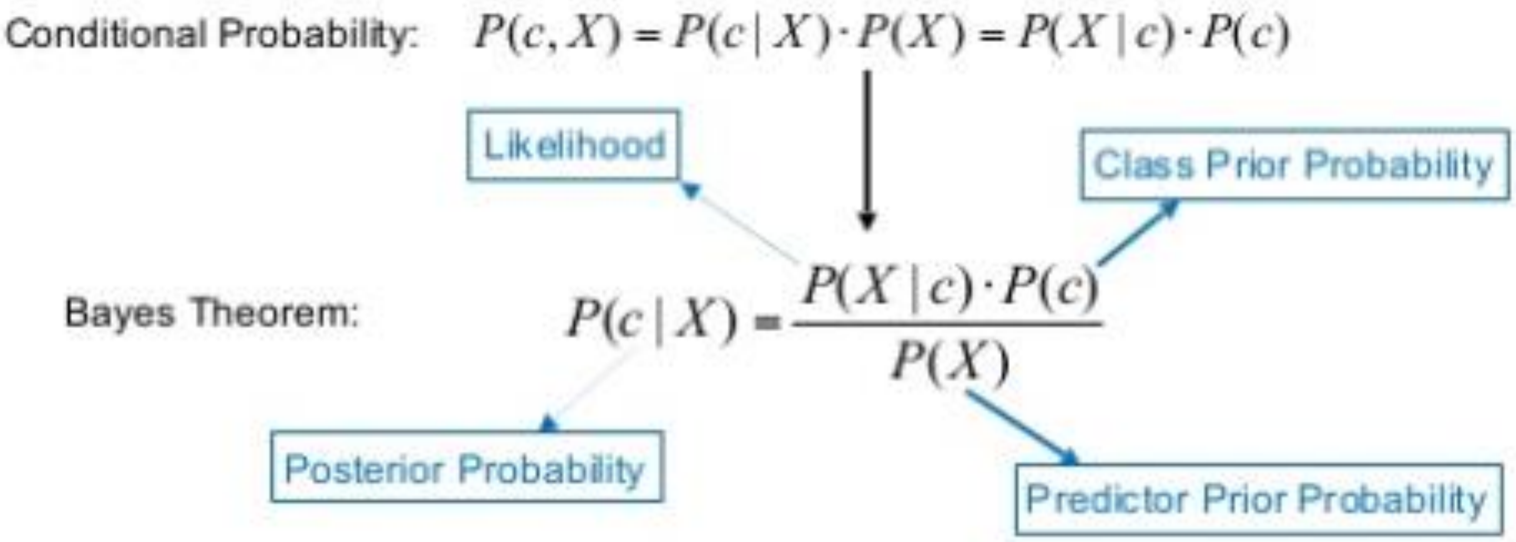
= P(x=convex, smooth/ poisonous) X P(poisonous)

= P(x=convex/ poisonous) X P(x=smooth/ poisonous) X P(poisonous)

= 3/7 x 2/7 x 1/2 = 6 / 98

So test point is classified as Edible

Naïve Bayes Theorem



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

Naïve Bayes Assumption

1. Variables are independent given class
2. Variables are conditionally independent
3. Decreases computational time of the algorithm

Naïve Bayes Theorem

Test Document	Class
Machines occupied our world longback	Machines
Moon is occupied by machines now	Machines
Machines already killing the animals now in the world	Machines
Humans are scared in this world	Mammals
animals already longback escaped one by one	Mammals

Naïve Bayes Theorem

DICTIONARY/VOCABULARY

Dictionary before Stop Word removal

Machines
occupied
our
world
longback
Moon
is
by
now
already
the
killing
animals
humans
are
scared
in
this
animals
were
escaped
one

Stop Words

our
is
by
the
in
this
are
were

Dictionary after Stop Word removal

Machines
occupied
world
longback
Moon
now
killing
humans
scared
Animals
Escaped
one

Naïve Bayes Theorem

Given the test documents you saw what is the probability of the document belonging to Mammals class?

Naïve Bayes Theorem

Given the test documents you saw what is the probability of the document belonging to Mammals class?

Ans: 0.4 i.e. 2/5

Naïve Bayes Theorem

You will now look at how to use the vocabulary/dictionary to represent the given test documents by counting the occurences of various words. This is the way we represent the documents in **Multinomial Naive Bayes**

Naïve Bayes Theorem

	humans	already	machines	occupied	one	world	now	killing	escaped	sacred	animals	longback
0	0	0	1	0	0	1	0	1	0	0	0	1
1	0	1	1	1	0	0	1	0	0	0	0	0
2	0	0	1	1	0	1	1	0	0	0	1	0
3	1	0	0	0	0	1	0	0	0	1	0	0
4	0	1	0	0	2	0	0	0	1	0	1	0

Naïve Bayes Theorem

BAG OF WORDS REPRESENTATION (ARRAY)

$$D = \begin{bmatrix} 0,0,1,0,0,1,0,1,0,0,0,1 \\ 0,1,1,1,0,0,1,0,0,0,0,0 \\ 0,0,1,1,0,1,1,0,0,0,1,0 \\ 1,0,0,0,0,1,0,0,0,1,0,0 \\ 0,1,0,0,2,0,0,0,1,0,1,0 \end{bmatrix}$$

Naïve Bayes Theorem

[humans . . . . longback]

$$D^{Machines} = \begin{bmatrix} 0,0,1,0,0,1,0,1,0,0,0,1 \\ 0,1,1,1,0,0,1,0,0,0,0,0 \\ 0,0,1,1,0,1,1,0,0,0,1,0 \end{bmatrix}$$

$$D^{Mammals} \begin{bmatrix} 1,0,0,0,0,1,0,0,0,1,0,0 \\ 0,1,0,0,2,0,0,0,1,0,1,0 \end{bmatrix}$$

Naïve Bayes Theorem

[humans . . . . longback]

$$D^{\text{Machines}} = \begin{bmatrix} 0,0,1,0,0,1,0,1,0,0,0,1 \\ 0,1,1,1,0,0,1,0,0,0,0,0 \\ 0,0,1,1,0,1,1,0,0,0,1,0 \end{bmatrix} \quad 13$$

$$D^{\text{Mammals}} \quad \begin{bmatrix} 1,0,0,0,0,1,0,0,0,1,0,0 \\ 0,1,0,0,2,0,0,0,1,0,1,0 \end{bmatrix} \quad 8$$

Naïve Bayes Theorem

[humans . . . . .longback]

$$D^{Machines} = \begin{bmatrix} 0,0,1,0,0,1,0,1,0,0,0,1 \\ 0,1,1,1,0,0,1,0,0,0,0,0 \\ 0,0,1,1,0,1,1,0,0,0,1,0 \end{bmatrix} \quad 13$$

$$0,1,3,2,0,2,2,1,0,0,1,1$$

$$D^{Mammals} \begin{bmatrix} 1,0,0,0,0,1,0,0,0,1,0,0 \\ 0,1,0,0,2,0,0,0,1,0,1,0 \end{bmatrix} \quad 8$$

$$1,1,0,0,2,1,0,0,1,1,1,0$$

Prior

$P(\text{machines}) = 3/5$

$P(\text{mammals}) = 2/5$

$$\left.\begin{array}{l} P(\text{machines}|w_1,w_2,\ldots,w_n) \\ P(\text{mammals}|w_1,w_2,\ldots,w_n) \end{array}\right\} \text{Posterior}$$

C = class

$$P(\text{mammals}|w_1,w_2,\ldots,w_n) = \frac{P(w_1,w_2,\ldots,w_n|\text{class})\,P(\text{class})}{P(w_1,w_2,\ldots,w_n)}$$

$$= \boxed{P(w_1|c)}\,P(w_2|c)\ldots\ldots P(w_n|c) \times P(c)$$

Naïve Bayes Theorem

Dictionary/Vocabulary

	n (machines)	p(w / c = machines)	n (mammals)	p(w / c = mammals)
w1 = humans	0	0	1	1/8
w2 = already	1	1/13	1	1/8
w3 = machines	3	3/13	0	0/8
w4 = occupied	2	2/13	0	0/8
w5 = one	0	0/13	2	2/8
w6 = world	2	2/13	1	1/8
w7 = now	2	2/13	0	0/8
w8 = killing	1	1/13	0	0/8
w9 = escaped	0	0/13	1	1/8
w10 = scared	0	0/13	1	1/8
w11 = animals	1	1/13	1	1/8
w12 = longback	1	1/13	0	0/8

Naïve Bayes Theorem

Test Document : **world animals**

1 1

Machines

	n (machines)	p(w / c = machines)	n (mammals)	p(w / c = mammals)
w1 = humans	0	0	1	1/8
w2 = already	1	1/13	1	1/8
w3 = machines	3	3/13	0	0/8
w4 = occupied	2	2/13	0	0/8
w5 = one	0	0/13	2	2/8
w6 = world	2	2/13	1	1/8
w7 = now	2	2/13	0	0/8
w8 = killing	1	1/13	0	0/8
w9 = escaped	0	0/13	1	1/8
w10 = scared	0	0/13	1	1/8
w11 = animals	1	1/13	1	1/8
w12 = longback	1	1/13	0	0/8

Naïve Bayes Theorem

Test Document : world animals

1 1

Machines

P(Machines | "World Animals")
= P("World" | Machines)
X P("Animals" | Machines)
X P("Machines")

$$\frac{2}{13} \times \frac{1}{13} \times \frac{3}{5}$$
= 0.007

	n (machines)	p(w / c = machines)	n (mammals)	p(w / c = mammals)
w1 = humans	0	0	1	1/8
w2 = already	1	1/13	1	1/8
w3 = machines	3	3/13	0	0/8
w4 = occupied	2	2/13	0	0/8
w5 = one	0	0/13	2	2/8
w6 = world	2	2/13	1	1/8
w7 = now	2	2/13	0	0/8
w8 = killing	1	1/13	0	0/8
w9 = escaped	0	0/13	1	1/8
w10 = scared	0	0/13	1	1/8
w11 = animals	1	1/13	1	1/8
w12 = longback	1	1/13	0	0/8

Naïve Bayes Theorem

world animals

Machines

P(Mammals | “World Animals”)
= P(“World” | Mammals)
X P(“Animals” | Mammals)
X P(“Mammals”)

$$= \frac{1}{8} \times \frac{1}{8} \times \frac{2}{5}$$

$$= 0.006$$

P(Machines | “World Animals”) > P(Mammals | “World Animals”)
Hence the document can be Machines category

LAPLACE SMOOTHING

Naïve Bayes Theorem

WHY WE NEED LAPLACE SMOOTHING?

Test Document : Machines occupied one planet

ignored (not present in dictionary)

	n (machines)	p(w / c = machines)	n (mammals)	p(w / c = mammals)
w1 = humans	0	0	1	1/8
w2 = already	1	1/13	1	1/8
w3 = machines	3	3/13	0	0/8
w4 = occupied	2	2/13	0	0/8
w5 = one	0	0/13	2	2/8
w6 = world	2	2/13	1	1/8
w7 = now	2	2/13	0	0/8
w8 = killing	1	1/13	0	0/8
w9 = escaped	0	0/13	1	1/8
w10 = scared	0	0/13	1	1/8
w11 = animals	1	1/13	1	1/8
w12 = longback	1	1/13	0	0/8

Naïve Bayes Theorem

P(Machines | "Machines occupied one")
= P("Machines" | Machines)
X P("occupied" | Machines)
X P("one" | Machines)
X P("Machines")

= 3 /13 X 2 / 13 X 0 X 3/5
= 0

WHY WE NEED LAPLACE SMOOTHING?

Test Document : Machines occupied one planet

ignored (not present in dictionary)

	n (machines)	p(w / c = machines)	n (mammals)	p(w / c = mammals)
w1 = humans	0	0	1	1/8
w2 = already	1	1/13	1	1/8
w3 = machines	3	3/13	0	0/8
w4 = occupied	2	2/13	0	0/8
w5 = one	0	0/13	2	2/8
w6 = world	2	2/13	1	1/8
w7 = now	2	2/13	0	0/8
w8 = killing	1	1/13	0	0/8
w9 = escaped	0	0/13	1	1/8
w10 = scared	0	0/13	1	1/8
w11 = animals	1	1/13	1	1/8
w12 = longback	1	1/13	0	0/8

Naïve Bayes Theorem

P(Mammals | “Machines
occupied one”)
= P(“Machines” |
Mammals)
X P(“occupied” |
Mammals)
X P(“one” | Mammals)
X P(“Mammals”)

= 0 X 0 X 2 / 8 X 2/5
= 0

Naïve Bayes Theorem

LAPLACE SMOOTHING

Laplace Smoothing

	n (machines)	p(w / c = machines)	n (mammals)	p(w / c = mammals)
w1 = humans	0+1	0	1+1	1/8
w2 = already	1+1	1/13	1+1	1/8
w3 = machines	3+1	3/13	0+1	0/8
w4 = occupied	2+1	2/13	0+1	0/8
w5 = one	0+1	0/13	2+1	2/8
w6 = world	2+1	2/13	1+1	1/8
w7 = now	2+1	2/13	0+1	0/8
w8 = killing	1+1	1/13	0+1	0/8
w9 = escaped	0+1	0/13	1+1	1/8
w10 = scared	0+1	0/13	1+1	1/8
w11 = animals	1+1	1/13	1+1	1/8
w12 = longback	1+1	1/13	0+1	0/8
	13+12 = 25			

Naïve Bayes Theorem

LAPLACE SMOOTHING

Laplace Smoothing

	n (machines)	p(w / c = machines)	n (mammals)	p(w / c = mammals)
w1 = humans	0+1	0	1+1	1/8
w2 = already	1+1	1/13	1+1	1/8
w3 = machines	3+1	3/13	0+1	0/8
w4 = occupied	2+1	2/13	0+1	0/8
w5 = one	0+1	0/13	2+1	2/8
w6 = world	2+1	2/13	1+1	1/8
w7 = now	2+1	2/13	0+1	0/8
w8 = killing	1+1	1/13	0+1	0/8
w9 = escaped	0+1	0/13	1+1	1/8
w10 = scared	0+1	0/13	1+1	1/8
w11 = animals	1+1	1/13	1+1	1/8
w12 = longback	1+1	1/13	0+1	0/8
	13+12 = 25		8+12 = 20	

Naïve Bayes Theorem

LAPLACE SMOOTHING

Laplace Smoothing

	n (machines)	p(w / c = machines)	n (mammals)	p(w / c = mammals)
w1 = humans	0+1	1/(12+13)=1/25	1+1=2	2/(12+8)=2/20
w2 = already	1+1	2/(12+13)=2/25	1+1=2	2/(12+8)=2/20
w3 = machines	3+1	4/(12+13)=4/25	0+1=1	1/(12+8)=1/20
w4 = occupied	2+1	3/(12+13)=3/25	0+1=1	1/(12+8)=1/20
w5 = one	0+1	1/(12+13)=1/25	2+1=3	3/(12+8)=3/20
w6 = world	2+1	3/(12+13)=3/25	1+1=2	2/(12+8)=2/20
w7 = now	2+1	3/(12+13)=3/25	0+1=1	1/(12+8)=1/20
w8 = killing	1+1	2/(12+13)=2/25	0+1=1	1/(12+8)=1/20
w9 = escaped	0+1	1/12+13)=1/25	1+1=2	2/(12+8)=2/20
w10 = scared	0+1	1/(12+13)=1/25	1+1=2	2/(12+8)=2/20
w11 = animals	1+1	2/(12+13)=2/25	1+1=2	2/(12+8)=2/20
w12 = longback	1+1	2/(12+13)=2/25	0+1=1	1/(12+8)=1/20
	13+12 = 25		8+12 = 20	

Naïve Bayes Theorem

Now calculate the probabilities

P(Machines | “Machines occupied one”)
= P(“Machines” | Machines)
X P(“occupied” | Machines)
X P(“one” | Machines)
X P(“Machines”)

= 4 /25 X 3 / 25 X 0 X 3/5
= 0.00046

P(Mammals | “Machines occupied one”)
= P(“Machines” | Mammals)
X P(“occupied” | Mammals)
X P(“one” | Mammals)
X P(“Mammals”)

= 1 /20 X 1 / 20 X 3/20 X 2/5
= 0.000075

BERNOULLI THEOREM

Naïve Bayes Theorem

BERNOULLI THEOREM : INTRODUCTION

$$D = \begin{bmatrix} 0,0,1,0,0,1,0,1,0,0,0,1 \\ 0,1,1,1,0,0,1,0,0,0,0,0 \\ 0,0,1,1,0,1,1,0,0,0,1,0 \\ 1,0,0,0,0,1,0,0,0,1,0,0 \\ 0,1,0,0,2,0,0,0,1,0,1,0 \end{bmatrix}$$

Naïve Bayes Theorem

BERNOULLI THEOREM : INTRODUCTION

$$D = \begin{bmatrix} 0,0,1,0,0,1,0,1,0,0,0,1 \\ 0,1,1,1,0,0,1,0,0,0,0,0 \\ 0,0,1,1,0,1,1,0,0,0,1,0 \\ 1,0,0,0,0,1,0,0,0,1,0,0 \\ 0,1,0,0,2,0,0,0,1,0,1,0 \end{bmatrix}$$

$$\uparrow$$

$$0,1,0,0,1,0,0,0,1,0,1,0$$

Naïve Bayes Theorem

In Bernoulli's Naive Bayes classification initial steps are the same. We first need to construct our dictionary. If the training documents remain same dictionary will be same .

Bernoulli's feature vectors are a bit different. It generates a Boolean indicator about each term of the vocabulary and equals to 1 if the term belongs to the examining document and 0 if it does not. Let us see the feature vector of our document for better understanding.

Formula for calculating individual probabilities in Bernouli is :

P(Machines | “Occupied”) = no of times the word ‘Occupied’ appearing in class / Total no of documents

Naïve Bayes Theorem

Similarly for calculating the overall probability we use the below formula:

Note : *d here means either 0 or 1 (Whether that word exists in the class or not)*

$$\textbf{P(Machines | d) = P(Machines) X} \prod_{i=1}^{12} [\ d_i P(w_i \mid C = Machines + (1 - d_i)(1 - P(w_i \mid C = Machines\ ))]$$

$$\textbf{P(Mammals | d) = P(Mammals) X} \prod_{i=1}^{12} [\ d_i P(w_i \mid C = Mammals\ + (1 - d_i)(1 - P(w_i \mid C = Mammals\ ))]$$