Machine Learning

Naïve Bayes Classifier

Probability Basics

Consider Problem of tossing two coins. The sample space as

{HH, HT, TH, TT}

Some of the probabilities in this experiment would be as follows:

The probability of getting two heads

= 1/4

The probability of at least one tail

= 3/4

The prob. of the second coin being head given the first coin is tail

= 1/2

Bayes Theorem

- Let us apply Bayes theorem to our coin example. S={HH, HT, TH, TT}

  Let A be the event that the second coin is head, and B be the event that the first coin is tails.

- Probability of A given B:

  P(A|B)= [P(B|A) * P(A)] / P(B)

  =[P(First coin being tail given the second coin is the head) * P(Second coin being head)] / P(First coin being tail)
  = [(1/2) * (1/2)] / (1/2)
  = 1/2 = 0.5

- Bayes theorem calculates the conditional probability of the occurrence of an event based on prior knowledge of conditions that might be related to the event.

Bayes Theorem

The Bayes theorem gives us the conditional probability of event A, given that event B has occurred.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where:

$P(A|B)$ = Conditional Probability of A given B

$P(B|A)$ = Conditional Probability of A given B

$P(A)$ = Probability of event A

$P(B)$ = Probability of event A

Bayes Theorem

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

$$P(A \mid B) = \frac{\overbrace{P(A \cap B)}^{\text{Joint Probability}}}{P(B)}$$

Conditional Probability

$$\underbrace{P(A \mid B)}_{\text{Posterior Probability}} = \frac{\overbrace{P(B \mid A)}^{\text{Likelihood}}\, \overbrace{P(A)}^{\text{Class Prior}}}{\underbrace{P(B)}_{\text{Predictor Prior}}}$$

Bayes Theorem

Bayes Theorem

- In a factory there are 100 units of a certain product, 5 of which are defective. We pick three units from the 100 units at random. What is the probability that none of them are defective?

Bayes Theorem

- Let us define A_i, $i=1,2,3$ as the event that the ith chosen unit is not defective.

  We are interested in $P(A1 \cap A2 \cap A3)$.

  Note that $P(A1)=95/100$.

- Given that the first chosen item was good, the second item will be chosen from 94 good units and 5 defective units,

  thus $P(A2|A1)=94/99$.
- Given that the first and second chosen items were okay, the third item will be chosen from 93 good units and 5 defective units,

  thus $P(A3|A2,A1)=93/98$.

- Thus, we have

  $P(A1 \cap A2 \cap A3)=P(A1)P(A2|A1)P(A3|A2,A1)=95/100 * 94/99 * 93/98=0.8560$

Example 4.14. An urn has 5 blue balls and 8 red balls. Each ball that is selected is returned to the urn along with an additional ball of the same color. Suppose that 3 balls are drawn in this way.

(a) What is the probability that the three balls are blue?
Solution: In this case, we can define the sequence of events $B_1, B_2, B_3, \ldots$, where B_i is the event that *the ith ball drawn is blue.* Applying the multiplication rule yields

$$\mathbb{P}(B_1 \cap B_2 \cap B_3) = \mathbb{P}(B_1)\mathbb{P}(B_2 \mid B_1)\mathbb{P}(B_3 \mid B_1 \cap B_2) = \frac{5}{13}\frac{6}{14}\frac{7}{15}.$$

Chain rule for conditional probability:

$$
\begin{aligned}
\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) &= \mathbb{P}((A_1 \cap A_2 \cap A_3) \cap A_4) && [\text{treat } A_1 \cap A_2 \cap A_3 \text{ as one event}] \\
&= \mathbb{P}(A_1 \cap A_2 \cap A_3)\,\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) && [\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)] \\
&= \mathbb{P}((A_1 \cap A_2) \cap A_3)\,\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) && [\text{treat } A_1 \cap A_2 \text{ as one event}] \\
&= \mathbb{P}(A_1 \cap A_2)\,\mathbb{P}(A_3 \mid A_1 \cap A_2)\,\mathbb{P}(A_4 \mid A_1 \cap A_3 \cap A_3) && [\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)] \\
&= \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 \cap A_2)\,\mathbb{P}(A_4 \mid A_1 \cap A_3 \cap A_3) && [\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)]
\end{aligned}
$$

Assume that we have two classes:
c_1=male and c_2=female

We have a person whose gender, we
Do not know, say *"drew"* or d

Classifying *drew* as male or female is equivalent to asking is it more probable that *drew* is male or female, ie. Which is greater p(male|drew) or p(female|drew)

(Note: "Drew can be a male or female name")

Drew Barrymore

Drew Carey

What is the probability of being called "*drew*" given that you are a male?

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

What is the probability of being a male?

What is the probability of being named "*drew*"?
(actually irrelevant, since it is that same for all classes)

Officer Drew

This is Officer Drew (who arrested me in 1997). Is Officer Drew a Male or Female?

Luckily, we have a small database with names and gender

Name	Class
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

We can use it to apply Bayes rule…

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

$$p(\text{female} \mid drew) = \frac{p(drew \mid \text{female})\, p(\text{female})}{p(drew)}$$

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Officer Drew

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

$$= \frac{1/3 * 3/8}{3/8} \qquad = \frac{0.125}{3/8}$$

Name	Class
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\text{female} \mid drew) = \frac{p(drew \mid \text{female})\, p(\text{female})}{p(drew)}$$

$$= \frac{2/5 * 5/8}{3/8} \qquad = \frac{0.250}{3/8}$$

Officer Drew is more likely to be a Female.

Officer Drew

Officer Drew IS a female!

$$p(\text{male} \mid drew) = \frac{1/3 \;*\; 3/8}{3/8} = \frac{0.125}{3/8}$$

$$p(\text{female} \mid drew) = \frac{2/5 \;*\; 5/8}{3/8} = \frac{0.250}{3/8}$$

So far we have only considered Bayes Classification when we have one attribute (the "name"). But we may have many features.

How do we use all the features?

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Name	Over 170 CM	Eye	Hair Length	Class
Drew	No	Blue	Short	Male
Claudia	Yes	Brown	Long	Female
Drew	No	Blue	Long	Female
Drew	No	Blue	Long	Female
Alberto	Yes	Brown	Short	Male
Karin	No	Blue	Long	Female
Nina	Yes	Brown	Short	Female
Sergio	Yes	Blue	Long	Male

- To simplify the task, Bayesian classifier assume attributes have independent distributions, and thereby estimate

$$p(d|c_j) = p(d_1|c_j) * p(d_2|c_j) * \ldots.* p(d_n|c_j)$$

The probability of class c_j generating instance d, equals….

The probability of class c_j generating the observed value for feature 1, multiplied by..

The probability of class c_j generating the observed value for feature 2, multiplied by..

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

Naïve Bayes Assumption

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

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

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

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

Let us take an example to get some better intuition.

- **Consider the car theft problem with attributes Color, Type, Origin, and the target,**

  Stolen can be either Yes or No.

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Training

Consider the car theft problem with attributes Color, Type, Origin, and the target,

Stolen can be either Yes or No.

Color	Type	Origin	Stolen
Red	SUV	Domestic	?

Testing

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Frequency and Likelihood tables of 'Color'

		Stolen?	
		Yes	No
Color	Red	3	2
	Yellow	2	3

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Color	Red	3/5	2/5
	Yellow	2/5	3/5

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Frequency and Likelihood tables of ‘Type’

Frequency Table

		Stolen?	
		Yes	No
Type	Sports	4	2
	SUV	1	3

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Type	Sports	4/5	2/5
	SUV	1/5	3/5

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Frequency and Likelihood tables of 'Origin'

Frequency Table

		Stolen?	
		Yes	No
Origin	Domestic	2	3
	Imported	3	2

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Origin	Domestic	2/5	3/5
	Imported	3/5	2/5

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Color	Red	3/5	2/5
	Yellow	2/5	3/5

Testing part

Color	Type	Origin	Stolen
Red	SUV	Domestic	?

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Type	Sports	4/5	2/5
	SUV	1/5	3/5

As per the equations discussed above, we can calculate the posterior probability P(Yes | X) as :

P(Yes | X) = P(Red | Yes) * P(SUV | Yes) * P(Domestic | Yes) * P(Yes)

$= \frac{3}{5} * \frac{1}{5} * \frac{2}{5} * 1$
= 0.048

P(No | X):

Likelihood Table

		Stolen?	
		P(Yes)	P(No)
Origin	Domestic	2/5	3/5
	Imported	3/5	2/5

P(No | X) = P(Red | No) * P(SUV | No) * P(Domestic | No) * P(No)

$= \frac{2}{5} * \frac{3}{5} * \frac{3}{5} * 1$
= 0.144

Since 0.144 > 0.048,

Which means given the features RED SUV and Domestic, our example gets classified as 'NO' the car is not stolen.

Naïve Bayes (NB) Classifier Algorithm

- It is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
- It is mainly used in *text classification* that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.
- Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles**.
- Why is it called Naïve Bayes?
- **Naïve**: *It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.*

  Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- **Bayes**: it depends on the principle of Bayes' Theorem

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

Naïve Bayes Assumption

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

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

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

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

Additional Example

Whether	Play
Sunny	No
Sunny	No
Overcast	Yes
Rainy	Yes
Rainy	Yes
Rainy	No
Overcast	Yes
Sunny	No
Sunny	Yes
Rainy	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table

Whether	No	Yes
Overcast		4
Sunny	2	3
Rainy	3	2
Total	5	9

Likelihood Table 1

Whether	No	Yes		
Overcast		4	=4/14	0.29
Sunny	2	3	=5/14	0.36
Rainy	3	2	=5/14	0.36
Total	5	9		
	=5/14	=9/14		
	0.36	0.64		

Likelihood Table 2

Whether	No	Yes	Posterior Probability for No	Posterior Probability for Yes
Overcast		4	0/5=0	4/9=0.44
Sunny	2	3	2/5=0.4	3/9=0.33
Rainy	3	2	3/5=0.6	2/9=0.22
Total	5	9		

Additional Example

Now suppose you want to calculate the probability of playing when the weather is overcast.

Probability of playing:

P(Yes | Overcast) = P(Overcast | Yes) P(Yes) / P (Overcast)

Calculate Prior Probabilities:

P(Overcast) = 4/14 = 0.29

P(Yes)= 9/14 = 0.64

Calculate Likelihood Probabilities:

P(Overcast |Yes) = 4/9 = 0.44

P (Yes | Overcast) = 0.44 * 0.64 / 0.29 = 0.98(Higher)

Similarly, you can calculate the probability of not playing:

Probability of not playing:

P(No | Overcast) = P(Overcast | No) P(No) / P (Overcast)

Calculate Prior Probabilities:

P(Overcast) = 4/14 = 0.29

P(No)= 5/14 = 0.36

Calculate Likelihood Probabilities:

P(Overcast |No) = 0/9 = 0

P (No | Overcast) = 0 * 0.36 / 0.29 = 0

The probability of a 'Yes' class is higher. So you can determine here if the weather is overcast than players will play the sport.

Likelihood Table 1

Whether	No	Yes		
Overcast		4	=4/14	0.29
Sunny	2	3	=5/14	0.36
Rainy	3	2	=5/14	0.36
Total	5	9		
	=5/14	=9/14		
	0.36	0.64		

Likelihood Table 2

Whether	No	Yes	Posterior Probability for No	Posterior Probability for Yes
Overcast		4	0/5=0	4/9=0.44
Sunny	2	3	2/5=0.4	3/9=0.33
Rainy	3	2	3/5=0.6	2/9=0.22
Total	5	9		

Limitation

- **The zero-frequency problem**

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

An Additional Example

Whether	Temperature	Play
Sunny	Hot	No
Sunny	Hot	No
Overcast	Hot	Yes
Rainy	Mild	Yes
Rainy	Cool	Yes
Rainy	Cool	No
Overcast	Cool	Yes
Sunny	Mild	No
Sunny	Cool	Yes
Rainy	Mild	Yes
Sunny	Mild	Yes
Overcast	Mild	Yes
Overcast	Hot	Yes
Rainy	Mild	No

An Additional Example

Now suppose you want to calculate the probability of playing when the weather is overcast, and the temperature is mild.

Probability of playing:

*P(Play= Yes | Weather=Overcast, Temp=Mild) = P(Weather=Overcast, Temp=Mild | Play= Yes)*P(Play=Yes) ....(1)

P(Weather=Overcast, Temp=Mild | Play= Yes)= P(Overcast |Yes) P(Mild |Yes) ………..(2)

Calculate Prior Probabilities: P(Yes)= 9/14 = 0.64

Calculate likelihood Probabilities: P(Overcast |Yes) = 4/9 = 0.44 P(Mild |Yes) = 4/9 = 0.44

Put Likelihood probabilities in equation (2) P(Weather=Overcast, Temp=Mild | Play= Yes) = 0.44 * 0.44 = 0.1936(Higher)

Put Prior and likelihood probabilities in equation (1) P(Play= Yes | Weather=Overcast, Temp=Mild) = 0.1936*0.64 = 0.124

Similarly, you can calculate the probability of not playing:

Probability of not playing:

*P(Play= No | Weather=Overcast, Temp=Mild) = P(Weather=Overcast, Temp=Mild | Play= No)*P(Play=No) ....(3)

P(Weather=Overcast, Temp=Mild | Play= No)= P(Weather=Overcast |Play=No) P(Temp=Mild | Play=No) ………..(4)

Calculate Prior Probabilities: P(No)= 5/14 = 0.36

Calculate likelihood Probabilities: P(Weather=Overcast |Play=No) = 0/9 = 0 P(Temp=Mild | Play=No)=2/5=0.4

Put posterior probabilities in equation (4) P(Weather=Overcast, Temp=Mild | Play= No) = 0 * 0.4= 0

Put prior and posterior probabilities in equation (3) P(Play= No | Weather=Overcast, Temp=Mild) = 0*0.36=0

The probability of a 'Yes' class is higher. So you can say here that if the weather is overcast than players will play the sport.

Example

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.

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

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	

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

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

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

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

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

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

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

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

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

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

		Play Golf Yes	Play Golf No
Humidity	High	3	4
	Normal	6	1

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

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

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

		Play Golf Yes	Play Golf No
Windy	False	6	2
	True	3	3

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

Outlook	Temp	Humidity	Windy	Play
Rainy	Cool	High	True	?

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

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

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

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

More example

- Let’s consider a simple dataset comprised of 10 data samples:

Example

Type of family structure	Age group	Income status	Will they buy a car?
Nuclear	Young	Low	Yes
Extended	Old	Low	No
Childless	Middle-aged	Low	No
Childless	Young	Medium	Yes
Single Parent	Middle-aged	Medium	Yes
Childless	Young	Low	No
Nuclear	Old	High	Yes
Nuclear	Middle-aged	Medium	Yes
Extended	Middle-aged	High	Yes
Single Parent	Old	Low	No

Given three inputs–for example, Single Parent, Young, and Low–we want to compute the probability of these people buying a car. Let's use Naive Bayes.

Firstly, let's compute the probability of the output labels (P(Y)) given the data.

P(No) = 4/10

P(Yes) = 6/10

Now let's calculate the probability of the likelihood of the evidence. Given the inputs Childless, Young, and Low, we'll calculate the probability with respect to both class labels as follows:

P(Single Parent|Yes) = 1/6
P(Single Parent|No) = 1/4
P(Young|Yes) = 2/6
P(Young|No) = 1/4
P(Low|Yes) = 1/6
P(Low|No) = 4/4

Since P(X1) * P(X2) * ... * P(Xn) remains the same when calculating the probability for both Yes and No output labels, we can eliminate that value.

Thus, the posterior probability is computed as follows (note that *X* is the test data):

P(Yes|X) = P(Single Parent|Yes) * P(Young|Yes) * P(Low|Yes) = 1/6 * 2/6 * 1/6 = 0.0063

P(No|X) = P(Single Parent|No) * P(Young|No) * P(Low|No) = 1/4 * 1/4 * 4/4 = 0.0625

The final probabilities are:

P(Yes|X) = 0.0063/(0.0063 + 0.0625) = 0.09
P(No|X) = 0.0625/(0.0063 + 0.0625) = 0.91

Thus, the results clearly show that the car probably will not be purchased.

We previously mentioned that the "naiveness" of the algorithm is that it assumes each feature is independent of the others. We calculated the probabilities with respect to the output label with this assumption, so that each feature has an equal contribution and is independent of all the other features.

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

Now if we send our test data, suppose **test = (Cow, Medium, Black)**

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%

Now if we send our test data, suppose **test = (Cow, Medium, Black)**

Probability of petting an animal :

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

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

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

Numerical Predictors

Numerical variables need to be transformed to their categorical counterparts (binning) before constructing their frequency tables.

The other option we have is using the distribution of the numerical variable to have a good guess of the frequency. For example, one common practice is to assume normal distributions for numerical variables. The probability density function for the normal distribution is defined by two parameters (mean and standard deviation).

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

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

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

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

		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

Numerical Value Example

- Problem: classify whether a given person is a male or a female based on the measured features.
- The features include height, weight, and foot size.
- Although with NB classifier we treat them as independent, they are not in reality.

Person	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
Female	5.75 (5'9")	150	9

The classifier created from the training set using a Gaussian distribution assumption

Person	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033×10^{-2}	176.25	1.2292×10^{2}	11.25	9.1667×10^{-1}
female	5.4175	9.7225×10^{-2}	132.5	5.5833×10^{2}	7.5	1.6667

Below is a sample to be classified as male or female.

Person	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

In order to classify the sample, one has to determine which posterior is greater, male or female. For the classification as male the posterior is given by

In order to classify the sample, one has to determine which posterior is greater, male or female. For the classification as male the posterior is given by

$$\text{posterior (male)} = \frac{P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male})}{evidence}$$

For the classification as female the posterior is given by

$$\text{posterior (female)} = \frac{P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})}{evidence}$$

$$P(\text{male}) = 0.5$$

$$p(\text{height} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789,$$

where $\mu = 5.855$ and $\sigma^2 = 3.5033 \cdot 10^{-2}$ are the parameters of normal distribution which have been previously determined from the training set.

$$p(\text{weight} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(130-\mu)^2}{2\sigma^2}\right) = 5.9881 \cdot 10^{-6}$$

$$p(\text{foot size} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(8-\mu)^2}{2\sigma^2}\right) = 1.3112 \cdot 10^{-3}$$

$$\text{posterior numerator (male)} = \text{their product} = 6.1984 \cdot 10^{-9}$$

$$P(\text{female}) = 0.5$$
$$p(\text{height} \mid \text{female}) = 2.23 \cdot 10^{-1}$$
$$p(\text{weight} \mid \text{female}) = 1.6789 \cdot 10^{-2}$$
$$p(\text{foot size} \mid \text{female}) = 2.8669 \cdot 10^{-1}$$
$$\text{posterior numerator (female)} = \text{their product} = 5.3778 \cdot 10^{-4}$$

Since posterior numerator is greater in the female case, the prediction is that the sample is female.

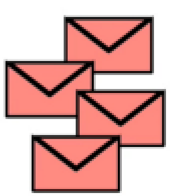
E-mail Filtering Problem

Example — Let us understand this with an example of email classification as spam or ham (i.e. no spam).
We simply count the number of words in both classes of email and then find the probability of each word's probability given the class prior probability of that email as spam or ham.
And then using the Naive Bayes, assuming that the occurrence of each word is independent of each other, we calculate the probability of a new email containing the words 'friend', 'rich', 'beach', 'money'.

HAM

Total emails: 72

SPAM



Total emails : 28

Prob. Ham
=72/100
= 0.72

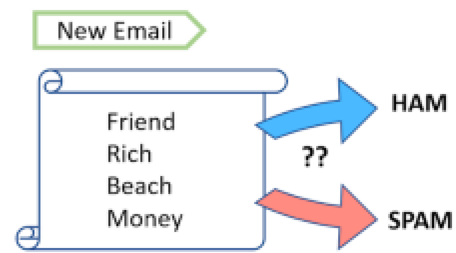
Prob. Spam
=28/100
= 0.28

Word	Count	Prob of Word if mail is Ham
Friend	86	0.238227
Rich	41	0.113573
Money	79	0.218837
Beach	80	0.221607
Office	75	0.207756

Ham: Total words=361

Word	Count	Prob of Word if mail is Spam
Friend	63	0.184751
Rich	36	0.105572
Money	97	0.284457
Beach	53	0.155425
Office	92	0.269795

Spam: Total words=341



Probability new email is Ham

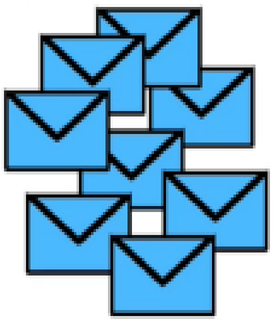
P(H) x P(Friend|H) x P(Rich|H) x P(Beach|H) x P(Money|H)

Probability new email is Spam

P(S) x P(Friend|S) x P(Rich|S) x P(Beach|S) x P(Money|S)

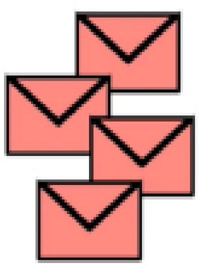
The Zero frequency problem:

HAM



Total emails: 72

SPAM



Total emails : 28

Prob. Ham
=72/100
= 0.72

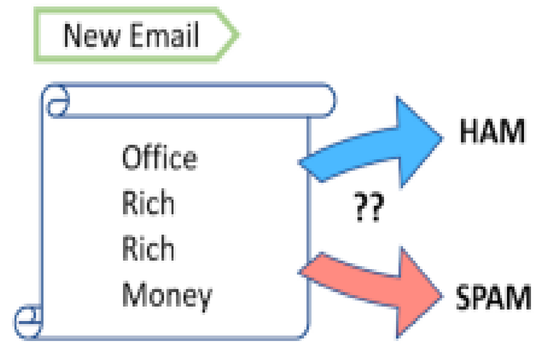
Prob. Spam
=28/100
= 0.28

HAM — Total Words **263**

Word	Count	Prob of Word if mail is Ham
Friend	86	0.326996
Rich	3	0.011407
Money	9	0.034221
Beach	90	0.342205
Office	75	0.285171

SPAM — Total Words **243**

Word	Count	Prob of Word if mail is Spam
Friend	59	0.242798
Rich	63	0.259259
Money	97	0.399177
Beach	24	0.098765
Office	**0**	**0**



Probability new email is Ham

P(H) x P(Office|H) x P(Rich|H) x P(Rich|H) x P(Money|H)

0.72 x 0.28 x 0.01 x 0.01 x 0.03 = **0.00000091**

Probability new email is Spam

P(S) x P(Office |S) x P(Rich|S) x P(Rich|S) x P(Money|S)

0.28 x **0** x 0.24 x 0.24 x 0.37 = **0.0**

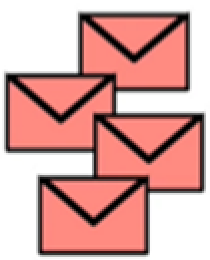
HAM??

The Zero frequency problem:

HAM

Total emails: 72

SPAM



Total emails : 28

Prob. Ham
=72/100
= 0.72

Prob. Spam
=28/100
= 0.28

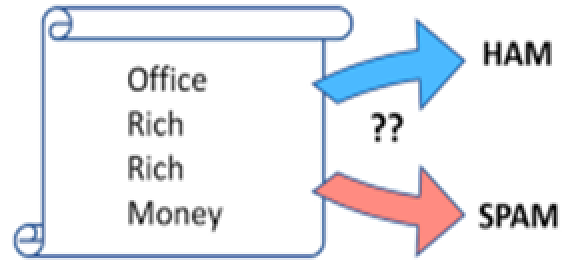
HAM — Total Words **268**

Word	Count	Count New	Prob of Word if mail is Ham
Friend	86 **+1**	87	0.324627
Rich	3 **+1**	4	0.014925
Money	9 **+1**	10	0.037313
Beach	90 **+1**	91	0.339552
Office	75 **+1**	76	0.283582

SPAM — Total Words **248**

Word	Count	Count New	Prob of Word if mail is Spam
Friend	59 **+1**	60	0.241935
Rich	63 **+1**	64	0.258065
Money	97 **+1**	98	0.395161
Beach	24 **+1**	25	0.100806
Office	0 **+1**	1	0.004032

New Email



Probability new email is Ham

P(H) x P(Office|H) x P(Rich|H) x P(Rich|H) x P(Money|H)

0.72 x 0.28 x 0.01 x 0.01 x 0.04 = **0.0000017**

Probability new email is Spam

P(S) x P(Office |S) x P(Rich|S) x P(Rich|S) x P(Money|S)

0.28 x 0.004 x 0.25 x 0.25 x 0.39 = **0.000029**

SPAM!!!

Suppose we are building a classifier that says whether a text is about sports or not. Our training data has 5 sentences:

Text	**Tag**
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

Now, which tag does the sentence ***A very close game*** belong to?

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match'].

$$P(sports|a\,very\,close\,game) = \frac{P(a\,very\,close\,game|sports) \times P(sports)}{P(a\,very\,close\,game)}$$

$$P(a\,very\,close\,game|Sports) = P(a|Sports) \times P(very|Sports) \times P(close|Sports) \times P(game|Sports)$$

$$P(game|Sports) = \frac{2}{11}$$

$$P(a|Sports) \times P(very|Sports) \times 0 \times P(game|Sports)$$

$$P(game|sports) = \frac{2+1}{11+14}$$

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

Word	P (word | Sports)	P (word | Not Sports)
a	(2 + 1) ÷ (11 + 14)	(1 + 1) ÷ (9 + 14)
very	(1 + 1) ÷ (11 + 14)	(0 + 1) ÷ (9 + 14)
close	(0 + 1) ÷ (11 + 14)	(1 + 1) ÷ (9 + 14)
game	(2 + 1) ÷ (11 + 14)	(0 + 1) ÷ (9 + 14)

Now we just multiply all the probabilities, and see who is bigger:

$$P(a|Sports) \times P(very|Sports) \times P(close|Sports) \times P(game|Sports) \times P(Sports)$$
$$= 2.76 \times 10^{-5}$$
$$= 0.0000276$$

$$P(a\,|Not\ Sports) \times P(very|Not\ Sports) \times P(close|Not\ Sports) \times P(game|Not\ Sports) \times P(Not\ Sports)$$
$$= 0.572 \times 10^{-5}$$
$$= 0.00000572$$

Excellent! Our classifier gives “A very close game” the **Sports** tag.

Build a NB classifier for predicting a class category (either + or -)

Table 1: The training and testing datasets for Question 5.

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no fun

Naive Bayes

Advantages of Naive Bayes

- This algorithm works **very fast and can easily predict the class** of a test dataset.
- You can use it to solve multi-class prediction problems as it’s quite useful with them.
- Naive Bayes classifier performs better than other models with less training data if the assumption of independence of features holds.
- If you have categorical input variables, the Naive Bayes algorithm performs exceptionally well in comparison to numerical variables.
- It can be used for Binary and Multi-class Classifications.
- It effectively works in Multi-class predictions.

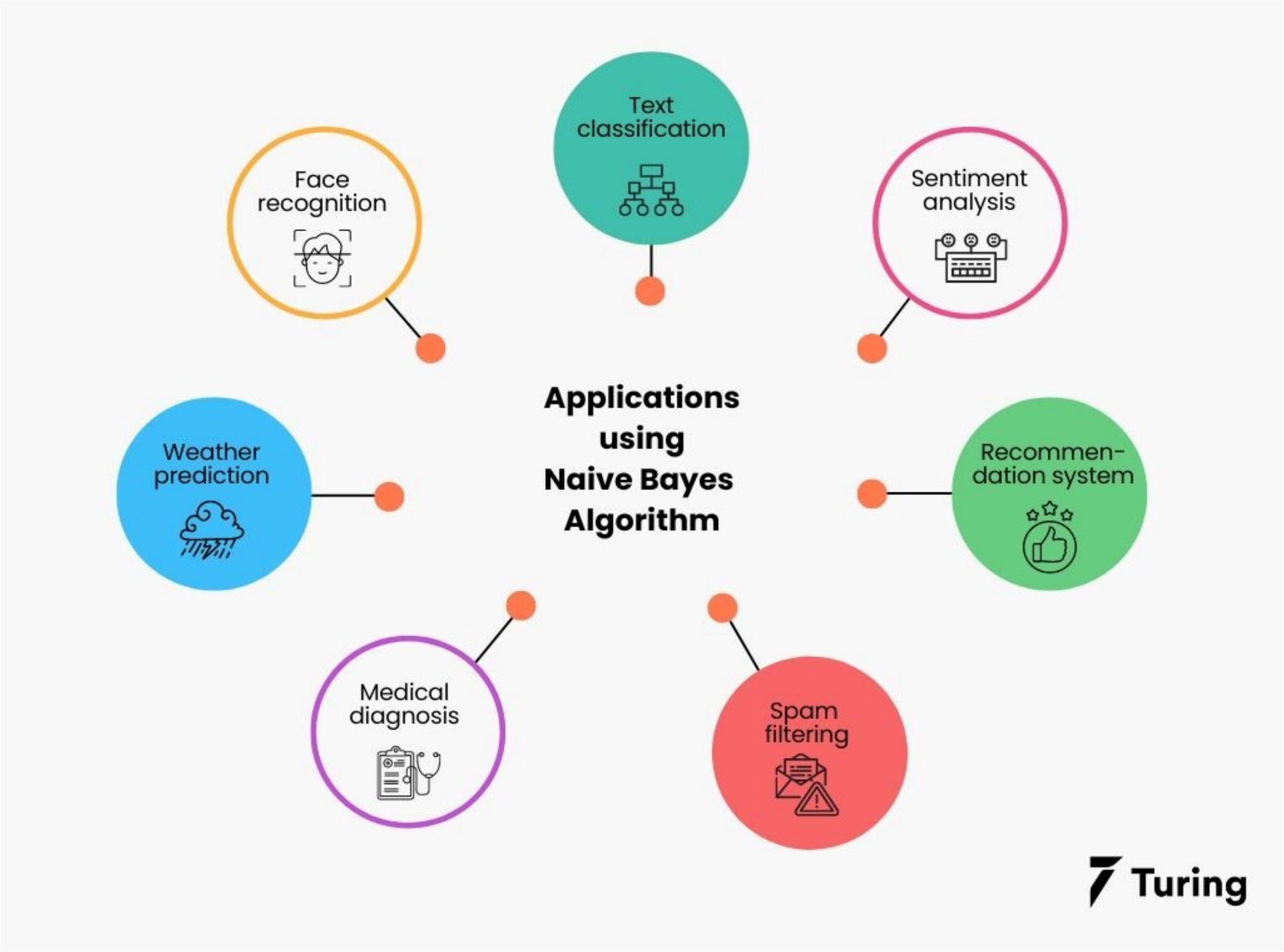
Disadvantages of Naive Bayes

- If your test data set has a categorical variable of a category that wasn’t present in the training data set, the Naive Bayes model will assign it **zero probability** and won’t be able to make any predictions in this regard. This phenomenon is called ‘Zero Frequency,’ and you’ll have to use a smoothing technique to solve this problem.
- It assumes that all the features are independent. While it might sound great in theory, in real life, you’ll hardly find a set of independent features.

Applications

- **Real-time prediction:** Naive Bayes is an eager learning classifier and is quite fast in its execution. Thus, it could be used for making predictions in real-time.
- **Multi-class prediction:** The Naive Bayes algorithm is also well-known for multi-class prediction, or classifying instances into one of several different classes.
- **Text classification/spam filtering/sentiment analysis:** When used to classify text, a Naive Bayes classifier often achieves a higher success rate than other algorithms due to its ability to perform well on multi-class problems while assuming independence. As a result, it is widely used in spam filtering (identifying spam email) and sentiment analysis (e.g. in social media, to identify positive and negative customer sentiments).
- **Recommendation Systems:** A Naive Bayes Classifier can be used together with Collaborative Filtering to build a Recommendation System which could filter through new information and predict whether a user would like a given resource or not.

Applications



Types of Naive Bayes classifier?

The main types of Naive Bayes classifier are mentioned below:

- **Multinomial Naive Bayes** — These types of classifiers are usually used for the problems of document classification. It checks whether the document belongs to a particular category like sports or technology or political etc and then classifies them accordingly. The predictors used for classification in this technique are the frequency of words present in the document.
- **Bernoulli Naive Bayes** — This classifier is also analogous to multinomial naive bayes but instead of words, the predictors are Boolean values. The parameters used to predict the class variable accepts only yes or no values, for example, if a word occurs in the text or not.
- **Gaussian Naive Bayes** — In a Gaussian Naive Bayes, the predictors take a continuous value assuming that it has been sampled from a Gaussian Distribution. It is also called a Normal Distribution.

How to Find Accuracy of Classifier?

Example of Classifier

Total are 12= 8 (with cancer) + 4 (without cancer)

Encoding: cancer (Positive- class 1), cancer-free (negetive- class 0)

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

Individual Number	1	2	3	4	5	6	7	8	9	10	11	12
Actual Classification	1	1	1	1	1	1	1	1	0	0	0	0
Predicted Classification	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

Confusion Matrix

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**True positive (TP)**	**False negative (FN)**
	Negative (N)	**False positive (FP)**	**True negative (TN)**

		Predicted condition	
	Total 8 + 4 = 12	**Cancer** 7	**Non-cancer** 5
Actual condition	**Cancer** 8	6	2
	Non-cancer 4	1	3

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

"Out of all the positive predictions we made, how many were true?"

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$

"Out of all the data points that should be predicted as true, how many did we correctly predict as true?"

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

F1 can therefore be used to measure how effectively our models make that trade-off among Precision and Recall

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

		Predicted condition	
	Total 8 + 4 = 12	Cancer 7	Non-cancer 5
Actual condition	Cancer 8	6	2
	Non-cancer 4	1	3

Example-1

Problem: Email spam filtering
Total sample: 100

Accuracy =70%

Precision=TP/(TP+FP)=10/30=33%

Confusion Matrix

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	TP=10	FN=10
	Negative (N)	FP=20	TN=60

Example-2

Problem: Tumor as cancerous or noncancerous
Total sample: 2000

Accuracy =70%

Recall=TP/(TP+FN)=200/600=33%

Confusion Matrix

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	**TP=200**	**FN=400**
	Negative (N)	**FP=200**	**TN=1200**

Example-3

Problem: Tumor as cancerous or noncancerous
Total sample: 2000

Accuracy =35%

Recall=TP/(TP+FN)=200/250=80%

Confusion Matrix

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	TP=200	FN=50
	Negative (N)	FP=1250	TN=500

Summary of Strength and Weakness of Metrics

Metrics	Strengths	Weaknesses
Accuracy	Easy to understand and compute; provides a general performance measure.	It can be misleading in imbalanced datasets and does not differentiate between types of errors.
Precision	Useful when the cost of false positives is high; measures the accuracy of positive predictions.	It does not account for false negatives and can be less informative if not considered with recall.
Recall	Crucial when the cost of false negatives is high; it measures the ability to identify positive instances.	It does not account for false positives and can be less informative if not considered with precision.
F1 Score	Balances precision and recall; useful in imbalanced datasets.	It does not account for true negative rates