5.7 Naïve Bayes Method

It is based on Bayes theorem given by Thomas Bayes in middle of the eighteenth century. It is amazing that despite being such an old theorem it has found its way into many modern fields such as AI and machine learning. Classification using Bayes theorem is different from the decision tree approach. It is based on a hypothesis that the given data belongs to a particular class. In this theorem probability is calculated for the hypothesis to be true.

Our understanding of this theorem begins with the fact that the Bayes Theorem in based on probabilities and so it is important to define some notations in the beginning.

$P(A)$ refers to the probability of occurrence of event A, while $P(A|B)$ refers to the conditional probability of event A given that event B has already occurred.

Bayes theorem is defined as follows …

$$P(A|B) = P(B|A)\, P(A)/P(B) \qquad \ldots(1)$$

Let us first prove this theorem.

We already know that

$$P(A|B) = P(A \,\&\, B)/P(B) \qquad \ldots(2)$$

[It is the probability that next event will be A when B has already happened]

Similarly,
$$P(B|A) = P\,(B \,\&\, A)/P(A) \qquad \ldots(3)$$

From equation (3):

Thus, $P (B \& A) = P(B|A) * P(A)$

By putting this value of $P (B \& A)$ in equation (2), we get ….

$$P(A|B) = P(B|A)\ P(A)/P(B)$$

This proves the Bayes theorem.

Now let us consider that we have to predict the class of X out of three given classes C_1, C_2 and C_3. Here, the hypothesis is that event X has already occurred. Thus, we have to calculate $P(C_i|X)$, conditional probability of class being C_i when X has already occurred.

According to Bayes theorem:

$$P(C_i\ |X) = P(X|C_i)\ P(C_i)/P(X)$$

In this equation:

$P(C_i|X)$ is the conditional probability of class being C_i when X has already occurred or it is probability that X belongs to class C_i.

$P(X|C_i)$ is the conditional probability of record being X when it is known that output class is C_i

$P\ (C_i)$ is probability that object belongs to class C_i.

$P(X)$ is the probability of occurrence of record X.

Here, we have already made the hypothesis that X has already occurred so $P(X)$ is 1 so we have to calculate $P(X|C_i)$ and $P(C_i)$ in order to find required value.

To further understand this concept let us consider the following database, where the class of customer is defined based on his/her marital status, gender, employment status and credit rating as shown below.

Owns Home	*Married*	*Gender*	*Employed*	*Credit rating*	*Class*
Yes	Yes	Male	Yes	A	II
No	No	Female	Yes	A	I
Yes	Yes	Female	Yes	B	III
Yes	No	Male	No	B	II
No	Yes	Female	Yes	B	III
No	No	Female	Yes	B	I
No	No	Male	No	B	II
Yes	No	Female	Yes	A	I
No	Yes	Female	Yes	A	III
Yes	Yes	Female	Yes	A	III

Here, we have to predict the class of occurrence of X and let us suppose X is as shown below.

Owns Home	*Married*	*Gender*	*Employed*	*Credit rating*	*Class*
Yes	No	Male	Yes	A	?

Then we have to calculate the probability for class C_i, when X has already occurred. Thus it is

$$P(C_i \mid Yes, No, Male, Yes, A) = P(Yes, No, Male, Yes, A \mid C_i) * P(C_i)$$

Let us first calculate probability of each class, i.e., $P(C_i)$. Here, we have three classes, i.e., *I, II* and *III*. There are total 10 instances in the given dataset and there are three instances for class *I*, three for class *II* and four for class *III*. Thus,

Probability of class *I*, i.e., $P(I) = 3/10 = 0.3$

Probability of class *II*, i.e., $P(II) = 3/10 = 0.3$

Probability of class *III*, i.e., $P(III) = 4/10 = 0.4$

Now, let us calculate $P(X|C_i)$, i.e., $P(Yes, No, Male, Yes, A \mid C_i)$

$P(Yes, No, Male, Yes, A \mid C_i) = P(Owns\ Home = Yes| C_i) * P(Married = No| C_i) * P(Gender = Male| C_i) * P(Employed = Yes| C_i) * P(Credit\ rating = A| C_i)$

Thus, we need to calculate

$P(Owns\ Home = Yes|C_i)$,

$P(Married = No| C_i)$,

$P(Gender = Male|C_i)$,

$P(Employed = Yes|C_i)$,

$P(Credit\ rating = A|C_i)$.

Owns Home	*Married*	*Gender*	*Employed*	*Credit rating*	*Class*
No	No	Female	Yes	A	I
No	No	Female	Yes	B	I
Yes	No	Female	Yes	A	I
Yes	**No**	**Male**	**Yes**	**A**	**? [To be found]**
1/3	1(3/3)	0(0/3)	1(3/3)	2/3	Probability of having{Yes, No, Male, Yes, A} Attribute value given the risk Class I
Yes	Yes	Male	Yes	A	II
Yes	No	Male	No	B	II
No	No	Male	No	B	II
Yes	**No**	**Male**	**Yes**	**A**	**? [To be found]**

Contd.

Owns Home	*Married*	*Gender*	*Employed*	*Credit rating*	*Class*
2/3	2/3	1(3/3)	1/3	1/3	Probability of having{Yes, No, Male, Yes, A} Attribute value given the risk Class II
Yes	Yes	Female	Yes	B	III
No	Yes	Female	Yes	B	III
No	Yes	Female	Yes	A	III
Yes	Yes	Female	Yes	A	III
Yes	**No**	**Male**	**Yes**	**A**	**? [To be found]**
2/4	0(0/4)	0(0/4)	1(4/4)	2/4	Probability of having{Yes, No, Male, Yes, A} Attribute value given the risk Class III

Thus, $P(X|I) = 1/3 * 1 * 0 * 1 * 2/3 = 0$

$P(X|II) = 2/3 * 2/3 * 1 * 1/3 * 1/3 = 4/81$

$\mathrm{P(X|III)} = 2 * 0 * 0 * 1 * 2 = 0$

$P(I \mid Yes, No, Male, Yes, A) = P(Yes, No, Male, Yes, A \mid I) * P(I) = 0*0.3 = 0$

$P(II \mid Yes, No, Male, Yes, A) = P(Yes, No, Male, Yes, A \mid II) * P(II) = 4/81*0.3 = 0.0148$

$P(III \mid Yes, No, Male, Yes, A) = P(Yes, No, Male, Yes, A \mid III) * P(III) = 0*0.4 = 0$

Therefore, *X* is assigned to Class *II*. It is very unlikely in real life datasets that the probability of class comes out to be 0 as in this example.

5.7.1 Applying Naïve Bayes classifier to the 'Whether Play' dataset

Let us consider another example of Naïve Bayes classifier for the dataset shown in Figure 5.36. It has 4 input attributes outlook, temperature, humidity and windy. Play is the class or output attribute. These 14 records contain the information if play has been held or not on the basis of any given day's weather conditions.

Instance Number	*Outlook*	*Temperature*	*Humidity*	*Windy*	*Play*
1	sunny	hot	high	false	No
2	sunny	hot	high	true	No
3	overcast	hot	high	false	Yes
4	rainy	mild	high	false	Yes
5	rainy	cool	normal	false	Yes
6	rainy	cool	normal	true	No
7	overcast	cool	normal	true	Yes
8	sunny	mild	high	false	No
9	sunny	cool	normal	false	Yes
10	rainy	mild	normal	false	Yes
11	sunny	mild	normal	true	Yes
12	overcast	mild	high	true	Yes
13	overcast	hot	normal	false	Yes
14	rainy	mild	high	true	No

Here, we have to predict the output class for some *X*. Let us suppose *X* is as given below.

Outlook	*Temperature*	*Humidity*	*Windy*	*Play*
sunny	cool	high	true	?

Attribute values and counts

Outlook	Temperature	Humidity	Windy	Play
sunny = 5	hot = 4	high = 7	true = 6	yes = 9

overcast = 4	mild = 6	normal = 7	false = 8	no = 5
rainy = 5	cool = 4			

Figure 5.36 Dataset for play prediction based on a given day's weather conditions

From this dataset, we can observe that when the outlook is sunny, play is not held on 3 out of the 5 days as shown in Figure 5.37.

Instance Number	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes

Figure 5.37 Probability of whether play will be held or not on a Sunny day

Therefore, it can be concluded that there are 60% chances that Play will not be held when it is a sunny day.

The Naïve Bayes theorem is based on probabilities, so we need to calculate probabilities for each occurrence in the instances. So, let us calculate the count for the each occurrence as shown in Figure 5.38.

	Play	
Outlook	yes	no
sunny	2	3
overcast	4	0
rainy	3	2
TOTAL	9	5

Instance Number	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	high	false	No
2	sunny	hot	high	true	No
3	overcast	hot	high	false	Yes
4	rainy	mild	high	false	Yes
5	rainy	cool	normal	false	Yes
6	rainy	cool	normal	true	No
7	overcast	cool	normal	true	Yes
8	sunny	mild	high	false	No
9	sunny	cool	normal	false	Yes
10	rainy	mild	normal	false	Yes
11	sunny	mild	normal	true	Yes
12	overcast	mild	high	true	Yes
13	overcast	hot	normal	false	Yes
14	rainy	mild	high	true	No

	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	2	3	hot	2	2	high	3	4	false	6	2	yes	9
overcast	4	0	mild	4	2	normal	6	1	true	3	3	no	5
rainy	3	2	cool	3	1								
TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	14

Figure 5.38 Summarization of count calculations of all input attributes

Now, we can calculate the probability of play being held or not for each value of input attribute as shown in Figure 5.39. For example, we have 2 instances for play not being held when outlook is rainy and there are in total 5 instances for the outlook attribute where play is not held.

Probability for play no given outlook rainy

$$= \frac{\text{occurrences of play no. given outlook = rainy}}{\text{total no. of occurrences for play no. for attribute outlook}} = 2/5 = 0.40$$

	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.22	0.60	hot	0.22	0.40	high	0.33	0.80	false	0.67	0.40	yes	0.64
overcast	0.44	0.00	mild	0.44	0.40	normal	0.67	0.20	true	0.33	0.60	no	0.36
rainy	0.33	0.40	cool	0.33	0.20								

Figure 5.39 Probability of play held or not for each value of attribute

Here, we have to predict the output class for some X. Let us suppose X is as given below.

Outlook	*Temperature*	*Humidity*	*Windy*	*Play*
sunny	cool	high	true	?

Then, we have to calculate probability of class C_i, when X has already occurred. Thus, it is

$$P(C_i \mid sunny, cool, high, true) = P(sunny, cool, high, true \mid C_i) * P(C_i)$$

Let us first calculate probability of each class, i.e., $P(C_i)$. Here, we have two classes, i.e., Yes and No. There are total 14 instances in the given dataset from which 9 instances are for class Yes and 5 for class No. Thus,

Probability of class Yes, i.e., P(Yes) = 9/14=0.64

Probability of class No, i.e., P(No) = 5/14=0.36

Now, let us calculate $P(X|C_i)$, i.e., $P(sunny, cool, high, true \mid C_i)$

$$P(sunny, cool, high, true|C_i) = P(outlook = sunny|C_i)*P(Temperature$$
$$= cool|C_i)* P(humidity$$
$$= high|C_i) * P(windy = true|C_i)$$

Therefore, probability of play to be held under these weather conditions is as shown in Figure 5.40.

$P(outlook = sunny|Yes)$ = 0.22

$P(Temperature = cool|Yes)$ = 0.33

$P(humidity = high|Yes)$ = 0.33

$P(windy = true|Yes)$ = 0.33

$P(Yes)$ = 0.64

	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.22	0.60	hot	0.22	0.40	high	0.33	0.80	false	0.67	0.40	yes	0.64
overcast	0.44	0.00	mild	0.44	0.40	normal	0.67	0.20	true	0.33	0.60	no	0.36
rainy	0.33	0.40	cool	0.33	0.20								

Figure 5.40 Probability for play 'Yes' for an unknown instance

Likelihood of play being held = $P(X|Yes) * P(Yes)$ = 0.22 * 0.33 * 0.33 * 0.33 * 0.64 = 0.0053

Similarly, probability of play not being held under these weather conditions is as shown in Figure 5.41.

$P(outlook = sunny|No)$ = 0.60

$P(Temperature = cool|No)$ = 0.20

$P(humidity = high|No)$ = 0.80

$P(windy = true|No)$ = 0.60

$P(No)$ = 0.36

	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.22	0.60	hot	0.22	0.40	high	0.33	0.80	false	0.67	0.40	yes	0.64
overcast	0.44	0.00	mild	0.44	0.40	normal	0.67	0.20	true	0.33	0.60	no	0.36
rainy	0.33	0.40	cool	0.33	0.20								

Figure 5.41 Probability for play 'No' for an unknown instance

Likelihood of play being held = *P(X|No)* * *P(No)* = 0.60 * 0.20 * 0.80 * 0.60 * 0.36 = 0.0206

Now, above calculated likelihoods are converted to probabilities to decide whether the play is held or not under these weather conditions.

Probability of play Yes = Likelihood for play Yes/(Likelihood for play Yes + Likelihood for play No) = 0.0053/(0.0053 + 0.0206) = 20.5%

Probability of play No = Likelihood for play No/(Likelihood for play Yes + Likelihood for play No) = 0.0206/(0.0053 + 0.0206) = 79.5%

From this, we can conclude that there are approximately 80% chances that play will not be held and 20% chances that play will be held. This makes a good prediction and Naïve Bayes has performed very well in this case.

But Naïve Bayes has a drawback to understand which, let us consider another example from this dataset under same weather conditions except that now the outlook is overcast instead of rainy. The problem here is that there is no instance in the dataset when outlook is overcast and play is No which indicates that play is always held on overcast day as shown in Figure 5.38. This result about likelihood of play not being held on overcast days is 0 which causes all the other attributes to be rejected. Simply put, we can say that play is always being held when outlook is overcast, regardless of what the other attributes are. The calculations for this are shown in Figure 5.42.

Therefore, likelihood of play not being held under these weather conditions = 0.00 * 0.20 * 0.80 * 0.60 * 0.36 = 0

	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.22	0.60	hot	0.22	0.40	high	0.33	0.80	false	0.67	0.40	yes	0.64
overcast	0.44	0.00	mild	0.44	0.40	normal	0.67	0.20	true	0.33	0.60	no	0.36
rainy	0.33	0.40	cool	0.33	0.20								

Figure 5.42 Probability of play not being held when outlook is overcast

However, we may have reached this result because we may not have enough instances in the dataset especially of cases where play was not held when the outlook was overcast. To solve this issue, a Laplace estimator is used which simply adds a value of 1 to each count of attribute values.

5.7.2 Working of Naïve Bayes classifier using the Laplace Estimator

After adding 1 to each count of attribute values, the modified values are shown in Figure 5.43.

	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	2	3	hot	2	2	high	3	4	false	6	2	yes	9
overcast	4	0	mild	4	2	normal	6	1	true	3	3	no	5
rainy	3	[illegible]	cool	3	1								
TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	14

Laplace estimator:
Add 1 to each count

Add 1 to count (Laplace estimator)

	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	3	[illegible]	hot	3	3	high	4	5	false	7	3	yes	9
overcast	5	1	mild	5	3	normal	7	2	true	4	4	no	5
rainy	4	3	cool	4	2								
TOTAL	12	8	TOTAL	12	8	TOTAL	11	7	TOTAL	11	7	TOTAL	14

Figure 5.43 Values of attributes after adding Laplace estimator

Now, probability of play being held or not for each modified value of input attribute is recomputed as shown in Figure 5.44.

	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.25	0.50	hot	0.25	0.38	high	0.36	0.71	false	0.64	0.43	yes	0.64
overcast	0.42	0.13	mild	0.42	0.38	normal	0.64	0.29	true	0.36	0.57	no	0.36
rainy	0.33	0.38	cool	0.33	0.25								

Figure 5.44 Probability of play held or not for each modified value of attribute

Now, probability of play whether held or not for the given weather conditions such as Outlook =overcast, Temperature = cool, Humidity = high, Windy = true can be calculated from the Figure 5.45.

	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.25	0.50	hot	0.25	0.38	high	0.36	0.71	false	0.64	0.43	yes	0.64
overcast	0.42	0.13	mild	0.42	0.38	normal	0.64	0.29	true	0.36	0.57	no	0.36
rainy	0.33	0.38	cool	0.33	0.25								

Figure 5.45 Attribute values for given example instance

Likelihood for play Yes = 0.42 * 0.33 * 0.36 * 0.36* 0.64 = 0.0118

Likelihood for play No = 0.13 * 0.25 * 0.71 * 0.57 * 0.36 =

0.0046Probability of play Yes = 0.0118/(0.0118+0.0046) = 72%

Probability of play No = 0.0046/(0.0118+0.0046) =28%

From this, it can be concluded that instead of 100% chance of play being held whenever outlook is overcast, we now calculate the probability of 72% for this combination of weather conditions, which is more realistic.

Thus, the Laplace estimator has been used to handle the case of zero probabilities which commonly occurs due to insufficient training data.

Advantages of Naïve Bayes Classifier:

- Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
- It can be used for Binary as well as Multi-class Classifications.
- It performs well in Multi-class predictions as compared to the other Algorithms.
- It is the most popular choice for **text classification problems**.

Disadvantages of Naïve Bayes Classifier:

- Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

Applications of Naïve Bayes Classifier:

- It is used for **Credit Scoring**.
- It is used in **medical data classification**.
- It can be used in **real-time predictions** because Naïve Bayes Classifier is an eager learner.
- It is used in Text classification such as **Spam filtering** and **Sentiment analysis**.

Types of Naïve Bayes Model:

There are three types of Naive Bayes Model, which are given below:

- **Gaussian**: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.
- **Multinomial**: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc.
  The classifier uses the frequency of words for the predictors.
- **Bernoulli**: The Bernoulli classifier works similar to the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.