**NAÏVE BAYES ALGORITHM:**

* Naïve Bayes is the probabilistic classifier algorithm that makes strong (naïve) assumptions on each input variable.
* Naïve Bayes algorithm supports categorical dataset.

**Example**

|  |  |
| --- | --- |
| **WEATHER** | **PLAY** |
| Sunny | No |
| Overcast | Yes |
| Rainy | Yes |
| Sunny | Yes |
| Sunny | Yes |
| Overcast | Yes |
| Rainy | No |
| Rainy | No |
| Sunny | Yes |
| Rainy | Yes |
| Sunny | No |
| Overcast | Yes |
| Overcast | Yes |
| Rainy | No |

* The above dataset has two variables ‘Weather’ and ‘play’. The Naïve Bayes predicts the play variable depending upon the assumption made on the weather variable.

**Class probability:**

* The dataset has two class values say ‘yes’ or ‘no’ OR ‘1’ or ‘0’.
* From the above example we can easily classify two classes from ‘play’ variable having ‘yes’ (1) or ‘no’(0).
* The probabilities for class 1 (yes) or class 0 (no) can be calculated as below:
* **P(class=1) = count(class=1) / (count(class=0) + count(class=1))**
* **P(class=0) = count(class=0) / (count(class=0) + count(class=1))**
* For our example, the probability for class ’yes’ can be calculated as follows:
* P (class=yes) = 9 / (5+9) = 9 / 14 = 0.6, the total ‘yes’ count from the above dataset example is 9 and the total ‘no’ count is 5.
* Similarly the probability of ‘no’ count can be calculated as below:
* P (class = no) = 5 / (5 + 9) = 5 / 14 = 0.3

**Conditional Probability:**

* The conditional probabilities calculate the probability of each input variable to the given class values from the given dataset.
* The conditional probability for our example can be calculates as follows:
* **Sunny Input Variable:**

P (weather = sunny | play = yes) = count (weather == sunny and play == yes) / count (play==yes)

* + - P ( weather == sunny | play == yes) = 3/9 = 0.3

P (weather = sunny | play = no) = count (weather == sunny and play == no) / count (play==no)

* + - P ( weather == sunny | play == no) = 2/5 = 0.4
* **Rainy Input Variable:**

P (weather = Rainy | play = yes) = count (weather == rainy and play == yes) / count (play==yes)

* + - P ( weather == rainy | play == yes) = 2/9 = 0.2

P (weather = rainy | play = no) = count (weather == rainy and play == no) / count (play==no)

* + - P ( weather == rainy | play == no) = 3/5 = 0.6
* **Overcast Input Variable:**

P (weather = overcast | play = yes) = count (weather == overcast and play == yes) / count (play==yes)

* + - P ( weather == overcast | play == yes) = 4/9 = 0.4

P (weather = overcast | play = no) = count (weather == overcast and play == no) / count (play==no)

* + - P ( weather == overcast | play == no) = 0/5 = 0

**Naïve Bayes Theorem:**

The naïve bayes theorem is used to predict the outcome for given the value of the class variable from the dataset. The theorem is as follows:

**P (h|d)= (P(d|h) \* P(h)) / P(d),** where

* P(h|d) = Posterior probability. The probability of hypothesis h being true, given the data d, where P(h|d)= P(d1| h)\* P(d2| h)\*....\*P(dn| h)\* P(d)
* P(d|h) = Likelihood. The probability of data d given that the hypothesis h was true.
* P(h) = Class prior probability. The probability of hypothesis h being true (irrespective of the data)
* P(d) = Predictor prior probability. Probability of the data (irrespective of the hypothesis)

For our dataset example, to predict the whether the outcome is ‘yes’ or ‘no’ for the weather input variable ‘sunny’ we have apply our dataset to Naïve Bayes theorem and choose the probability with higher range.

Let us consider our example here, where we have to calculate P (sunny | yes) and P (Sunny | No) and ‘yes or ‘no’ to ‘weather = sunny’ based on the higher probability value. It is depicted as follows:

P (yes | sunny) = (P (sunny | yes) \* P (yes)) / P (sunny)

= (3/9 \* 9/14) / (5/14) = 0.60

P (no | sunny) = (P (sunny | no) \* P (no)) / P (sunny)

  = (2/5 \* 5/14) / (5/14) = 0.40

Thus, if the weather =’sunny’, the outcome is play= ‘yes’, since it has higher probability. From the above calculation we can understand that, the count of weather = ‘sunny’ and play = ‘yes’ count is 3, the count of ‘yes’ is 9, the count of ‘sunny’ is 9 and total number of dataset count is ‘14’. Similarly we can obtain the probability of outcome for ‘Rainy’ and ‘overcast’.

P (yes | rainy) = (P (rainy | yes) \* P (yes)) / P (rainy)

= (2/9 \* 9/14) / (5/14) = 0.4

P (no | rainy) = (P (rainy | no) \* P (no)) / P (rainy)

  = (3/5 \* 5/14) / (5/14) = 0.6

Thus, if the weather =’rainy’, the outcome is play= ‘no’, since it has higher probability

P (yes | overcast) = (P (overcast | yes) \* P (yes)) / P (overcast)

= (4/9 \* 9/14) / (4/14) = 1

P (no | overcast) = (P (overcast | no) \* P (no)) / P (overcast)

  = (0/5 \* 5/14) / (4/14) = 0

Thus, if the weather =’overcast’, the outcome is play= ‘yes’, since the probability for ‘no’ is 0.

**Advantages of Naïve Bayes:**

* Easy implementation
* Requires less training data for prediction
* Results are in good form

**Drawbacks of Naïve Bayes:**

* Class conditional independence results in loss of accuracy
* By assumption we can derive independence among variable, but practically dependencies exist among variables. Example: hospitals: patients: profile: age, family etc.
* Hence dependent variables cannot be modeled by Naive Bayes Classifier.