What are Naïve Bayes classifiers?

Naïve Bayes is part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category. Unlike discriminative classifiers, like logistic regression, it does not learn which features are most important to differentiate between classes.

Why it is Called Naive Bayes?

The "Naive" part of the name indicates the simplifying assumption made by the Naïve Bayes classifier. The classifier assumes that the features used to describe an observation are conditionally independent, given the class label. The "Bayes" part of the name refers to Reverend Thomas Bayes, an 18th-century statistician and theologian who formulated Bayes' theorem.

Bayes' Theorem

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

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

where A and B are events and P(B) ≠ 0

- Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as **evidence**.
- P(A) is the **priori** of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).
- P(B) is Marginal Probability: Probability of Evidence.
- P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.
- P(B|A) is Likelihood probability i.e the likelihood that a hypothesis will come true based on the evidence.

Now, with regards to our dataset, we can apply Bayes' theorem in following way:

$P(y|X)=P(X|y)P(y)P(X)P(y|X)=P(X)P(X|y)P(y)$

where, y is class variable and X is a dependent feature vector (of size n) where:

$X=(x1,x2,x3,.....,xn)X=(x1,x2,x3,.....,xn)$

Just to clear, an example of a feature vector and corresponding class variable can be: (refer 1st row of dataset)

X = (Rainy, Hot, High, False)
y = No

So basically, $P(y|X)$*P*(*y*|*X*)here means, the probability of "Not playing golf" given that the weather conditions are "Rainy outlook", "Temperature is hot", "high humidity" and "no wind".

With relation to our dataset, this concept can be understood as:

- We assume that no pair of features are dependent. For example, the temperature being 'Hot' has nothing to do with the humidity or the outlook being 'Rainy' has no effect on the winds. Hence, the features are assumed to be **independent**.
- Secondly, each feature is given the same weight(or importance). For example, knowing only temperature and humidity alone can't predict the outcome accurately. None of the attributes is irrelevant and assumed to be contributing **equally** to the outcome.

Now, its time to put a naive assumption to the Bayes' theorem, which is, **independence** among the features. So now, we split **evidence** into the independent parts.

Now, if any two events A and B are independent, then,

P(A,B) = P(A)P(B)

Hence, we reach to the result:

$P(y|x1,\dots,xn)$=$P(x1|y)P(x2|y)\dots P(xn|y)P(y)P(x1)P(x2)\dots P(xn)$*P*(*y*|*x*1,…,*xn*)=*P*(*x*1)*P*(*x*2)…*P*(*xn*)*P*(*x*1|*y*)*P*(*x*2|*y*)…*P*(*xn*|*y*)*P*(*y*)

which can be expressed as:

$P(y|x1,\dots,xn)$=$P(y)\prod i=1nP(xi|y)P(x1)P(x2)\dots P(xn)$*P*(*y*|*x*1,…,*xn*)=*P*(*x*1)*P*(*x*2)…*P*(*xn*)*P*(*y*)∏*i*=1*nP*(*xi*|*y*)

Now, as the denominator remains constant for a given input, we can remove that term:

$P(y|x1,\dots,xn)\propto P(y)\prod i=1nP(xi|y)$*P*(*y*|*x*1,…,*xn*)∝*P*(*y*)∏*i*=1*nP*(*xi*|*y*)

Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible values of the class variable *y* and pick up the output with maximum probability. This can be expressed mathematically as:

$y=argmaxyP(y)\prod i=1nP(xi|y)$*y*=*argmaxyP*(*y*)∏*i*=1*nP*(*xi*|*y*)

So, finally, we are left with the task of calculating $P(y)$*P*(*y*)and $P(xi|y)$*P*(*xi*|*y*).

Please note that $P(y)$*P*(*y*) is also called class probability and $P(xi|y)$*P*(*xi*|*y*) is called conditional probability.

The different naive Bayes classifiers differ mainly by the assumptions they make regarding the distribution of $P(xi|y)$.*P*(*xi*|*y*).