# *12* Naïve-Bayes Analysis

Learning Objectives

* Understand the concept of Naïve-Bayes Analysis as a classification technique
* Learn how the NB model works
* Use the NB technique in different situations
* Know the many advantages and disadvantages of Naïve-Bayes analysis

### INTRODUCTION

Naïve-Bayes (NB) technique is a supervised learning technique that uses prob- ability-theory-based analysis. It is a machine-learning technique that computes the probabilities of an instance belonging to each one of many target classes, given the prior probabilities of classification using individual factors. Naïve- Bayes technique is used often in classifying text documents into one of the multiple predefined categories.

#### Caselet: Fraud Detection in Government Contracts

*TCU, the Brazilian Court of Audit, wanted to use data mining for more efficient and effective use of resources for fraud detection in government transactions. They wanted to find ways to identify projects that had a high probability of fraud. An NB classifier was designed to assign a risk factor to each case.*

*The audit universe was the set of purchases, contracts, assets, programs, public employees, or contractors. The audit universe was the set of pairs of public and private parties of all public contracts signed by the Brazilian federal administration between 1997 and 2011, totalizing almost 800,000 pairs, allowing and assisting audit agencies in transitioning from reactive to proactive fraud detection model.*

*Second, the risk factors of an audit universe were identified that could relate to the entire entity set. The risk factors included the project related elements as well as external factors.*

Competitors *less competition = higher risk*

Value *more money = higher risk*

Bid Protests *more protests = higher risk*

Bid Winner Profile *no qualified winner = higher risk*

Bid Winner Previous Sanctions *sanctioned winner = higher risk*

*Factors associated with different concerns than the risk itself were also added.* Political/Economic Impact *related to policy goals = higher economic impact* Social Impact *public health spend = higher social impact*

Timing and Effectiveness of *older purchases = less auditing effectiveness*

*Naïve-Bayes algorithm analysis was used to compute the probability distribution of entities belonging to risk classes. They have only two classes (high and low risk). So two values for each entity were obtained – the probability of high risk and probability of low risk. These probabilities were used to compute the conditional probability of fraud for the set of 800,000 pairs of public and private parties. The probability of high risk was used to sort the auditable units by total risk score. The result was that almost 2500 pairs had a high-risk probability score of higher than 99 percent. This result included parties with obvious links, like private not-for-profit foundations for which the Brazilian public procurement law gives special treatment. Moreover, several high-risk pairs were known by the auditors from previous investigations where corruption was effectively found. The auditors found the derivation of the high-risk list as very reasonable. The ranked audit universe was then used to develop an audit plan.*

1. *What are the advantages to the government and people of Brazil from this analysis? How could this analysis be improved?*
2. *Identify another environment where a similar NB analysis could be beneficial?*

### PROBABILITY

Probability is defined as the chance of something happening. The probability values thus range from zero to one; with a value of zero representing no chance, and one representing total certainty. Using past event records, the probability of something happening in the future can be reliably assessed. For example, one can assess the probability of dying from an airline accident, by dividing the total number of airline accident-related deaths in a period by the total number of people flying during that period. These probabilities can then be compared to come to the conclusions, such as the safety levels of various event types. For example, past data may show that the probability of dying from an airline accident is less than that of dying from being hit by lightning.

The Naïve-Bayes algorithm is special in that it takes into consideration the prior probability of an instance belonging to a class, in addition to the recent track record of the instance belonging to that class.

* The word Bayes refers to Bayesian analysis (based on the work of mathematician Thomas Bayes) which computes the probability of a new occurrence not only on the recent record but also based on prior experience.
* The word Naïve represents the strong assumption that all parameters/features of an instance are independent variables with little or no correlation. Thus if people are identified by their height, weight, age, and gender, all these variables are assumed to be independent of each other.

NB algorithm is easy to understand and works fast. It also performs well in multiclass prediction, such as when the target class has multiple options beyond binary yes/no classification. NB can perform well even in the case of categorical input variables compared to the numerical variable(s).

### NAÏVE-BAYES MODEL

In the abstract, Naïve-Bayes is a conditional probability model for classification purposes.

The goal is to find a way to predict the class variable using a vector of independent variables , i.e., finding the function . In probability terms, the goal is to find , i.e., the probability of belonging to a certain class . is generally assumed to be a categorical variable with two or more discrete values.

Given an instance to be classified, represented by a vector representing ‘*n*’ features (independent variables), the Naïve-Bayes model assigns, to an instance, probabilities of belonging to any of the classes. The class *K* with the highest posterior probability is the label assigned to the instance.

The posterior probability (of belonging to a Class *K*) is calculated as a function of prior probabilities and current likelihood value, as shown in the equation below

is the posterior probability of class , given predictor .

*k*

is the prior probability of class *K*.

is the prior probability of the predictor.

is the current likelihood of the predictor given class.

*k*

### SIMPLE CLASSIFICATION EXAMPLE

Suppose a salon needs to predict the service required by the incoming customer. If there are only two services offered – Hair cut (*R*) and Manicure-Pedicure (*M*),

then the value to be predicted is whether the next customer will be for *R* or *M*. The number of classes (*K*) is 2.

The first step is to compute the prior probability. Suppose the data gathered for the last one year showed that during that period there were 2500 customers for *R* and 1500 customers for *M*. Thus, the default (or prior) probability for the next customer to be for *R* is 2500/4000 or 5/8. Similarly, the default probability for the next customer to be for *M* is 1500/4000 or 3/8. Based on this information alone, the next customer would likely be for *R*.

Another way to predict the service requirement by the next customer is to look at the most recent data. One can look at the last few (choose a number) customers, to predict the next customer. Suppose the last five customers were for the services … *R*, *M*, *R*, *M*, *M* order. Thus, the data shows the recent probability of *R* is 2/5 and that of *M* is 3/5. Based on just this information, the next customer will likely be for *M*.

Thomas Bayes suggested that the prior probability should be informed by the more recent data. Naïve-Bayes posterior probability for a class is computed by multiplying the prior probability and the recent probability.

Thus in this example, the NB posterior probability *P*(*R*) is (5/8 x 2/5) = 10/40. Similarly, the NB probability *P*(*M*) is (3/8 x 3/5) = 9/40. Since *P*(*R*) is greater than *P*(*M*), it follows that there is a greater probability of the next customer to be for *R*. Thus the expected class label assigned to the next customer would be *R*.

Suppose, however, the next customer coming in was for *M* service. The last five customer sequence now becomes *M*, *R*, *M*, *M*, *M*. Thus, the recent data shows the probability for *R* to be 1/5 and that of *M* to be 4/5.

Now the NB probability for *R* is (5/8 x 1/5) = 5/40. Similarly, the NB probability for *M* is (3/8 x 4/5) = 12/40. Since *P*(*M*) is greater than *P*(*R*), it follows that there is a greater probability of the next customer to be for *M*. Thus the expected class label assigned to the next customer is *M*.

The NB predictor thus dynamically changes its prediction value based on the recent data.

### TEXT CLASSIFICATION EXAMPLE

The probability of the document ‘*d*’ being in class ‘*c*’ is computed as follows

Where, is the conditional probability of term occurring in a document of class *c*.

Dataset 12.1 shows the text classification training and test data. The goal is to classify the test data into the right class as *h* or ~*h* (read as not *h*).

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset 12.1 |  | | |
| Training Set | Document ID | Keywords in the Document | Class = *h*  (Healthy) |
|  | 1 | Love Happy Joy Joy Love | Yes |
|  | 2 | Happy Love Kick Joy Happy | Yes |
|  | 3 | Love Move Joy Good | Yes |
|  | 4 | Love Happy Joy Pain Love | Yes |
|  | 5 | Joy Love Pain Kick Pain | No |
|  | 6 | Pain Pain Love Kick | No |
| Test Data | 7 | Love Pain Joy Love Kick | ? |

The prior probabilities of a document being classified using the six documents are

*P*(*h*) = 4/6 = 2/3

*P*(~*h*) = 2/6 = 1/3

i.e., there is a 2/3 prior probability that a document will be classified as *h* and 1/3 probability of not *h*.

The conditional probability for each term is the relative frequency of the term occurring in each class of the documents ‘*h* class’ and ‘not *h* class’.

Conditional Probabilities

Class *h* Class ~*h*

*P*(Love | *h*) = 5/19 *P*(Love | ~*h*) = 2/9 *P*(Pain | *h*) = 1/19 *P*(Pain | ~*h*) = 4/9 *P*(Joy | *h*) = 5/19 *P*(Joy | ~*h*) = 1/9 *P*(Kick| *h*) = 1/19 *P*(Kick| ~*h*) = 2/9

The probability of the test instance belonging to class *h* can be computed by multiplying the prior probability of the instance belonging to class *h*, with the conditional probabilities for each of the terms in the document for class *h*. Thus,

*P*(*h* | *d*7) = *P*(*h*) \* (*P*(Love | *h*))^2 \* *P*(Pain | *h*) \* *P*(Joy | *h*) \* *P*(Kick| *h*)

= (2/3) \* (5/19) \* (5/19) \* (1/19) \* (5/19) \* (1/19) = ~0.0000067

Similarly, the probability of the test instance being (not *h*) can be computed using conditional probabilities for not *h*.

*P*(~*h* | *d*7) = *P*(~*h*) \* *P*(Love | ~*h*) \* *P*(Love | ~*h*) \* *P*(Pain | ~*h*) \* *P*(Joy | ~*h*) \*

*P*(Kick| ~*h*)

= (1/3) \* (2/9) \* (2/9) \* (4/9) \* (1/9) \* (2/9) = 0.00018

The NB probability of the test instance being ‘not *h*’ is much higher than its be- ing *h*. Thus the test document will be classified as ‘not *h*’.

### ADVANTAGES AND DISADVANTAGES OF NAÏVE-BAYES

* + The NB logic is simple and so is the NB posterior probability computation.
  + Conditional probabilities can be computed for discrete data and probabilistic distributions. When there are some variables in the vector *X*, then the problem can be modeled using probability functions to simulate the incoming values. A variety of methods exist for modeling the conditional distributions of the *X* variables, including normal, lognormal, gamma, and Poisson.
  + Naïve-Bayes assumes that all the features are independent for most in- stances that work fine. However, it can be a limitation. If there are no joint occurrences at all of a class label with a certain attribute, then the frequency-based conditional probability will be zero. When all the probabilities are multiplied, it will make the entire posterior probability estimate to be zero. This can be rectified by adding 1 to all the numerators and adding *n*, the number of variables in *X*, to all the denominators. This will make those probabilities very small but not zero.
  + A limitation of NB is that the posterior probability computations are good for comparison and classification of the instances. However, the probability values themselves are not good estimates of the event happening.

## Conclusion

Naïve-Bayes is a probability-based machine learning technique used for classifications. It is a mathematically simple way to include the contributions of many factors in predicting the class of the next data instance. It is often used to classify texts.

## Questions

1. What is Naïve-Bayes technique? What do Naïve and Bayes stand for?
2. In what ways is NB better than other classification techniques? Compare with decision trees.
3. What are the most popular applications of the NB technique?

## True/False

1. Naïve-Bayes is a statistical learning technique.
2. NB is popular for classifying text documents.
3. NB uses a probability-based analysis to predict future occurrences.
4. NB uses prior and recent probabilities to compute the posterior probability.
5. NB assumes that the input variables are independent of each other.
6. NB requires discrete data values and not continuous distributions.