**Naïve Bayes**

Naïve Bayes classifiers, a family of classifiers that are known for creating simple yet well performing models, especially in the fields of document classification and disease prediction.

The probabilistic model of naïve Bayes classifiers is based on Bayes’ theorem, and the adjective naïve comes from the assumption that the features in a dataset are mutually independent. In practice, the independence assumption is often violated, but naïve Bayes classifiers still tend to perform very well under this unrealistic assumption.

However , strong violations of the independence assumptions and non-linear classification problems can lead to very poor performances of naïve Bayes classifiers. We have to keep in mind that the type of data and the type problem to be solved dictate which classification model we want to choose .In practice , it is always recommended to compare different classification models on the particular dataset and consider the prediction performances as well as computational efficiency

**Posterior Probabilities**

Bayes’ theorem :



It can be interpreted as:

1. what is the probability that a particular object belongs to class I given its observed feature values.
2. What is the probability that a person has diabetes given a certain value for a pre-breakfast blood glucose measurement and a certain value for a post-breakfast blood glucose measurement

Let

*  be the feature vector of sample i, ,
*  be the notation of class j, 
* And  be the probability of observing sample  given that is belongs to class .

The general notation of the posterior probability can be written as



The objective function in the naïve Bayes probability is to maximize the posterior probability given the the training data in order to formulate the decision rule.



**Class-conditional Probabilites**

One assumption that Bayes classifiers make is that the samples are i.i.d (independent and identically distributed).Independence means that the probability of one observation does not affect the probability of another observation.

An additional assumption of naïve Bayes classifiers is the conditional independence of features. Thus , given a d-dimensional feature vector x, the class conditional probability can be calculated as follows:



The individual likelihoods for every feature in the feature vector can be estimated via the maximum-likelihood estimate, which is simply a frequency in the case of categorical data:



* : Number of times feature xi appears in samples from class wj
* : Total count of all features in class wj.

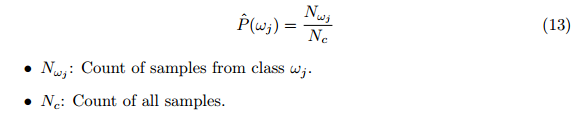
However, with respect to the *naive* assumption of conditional independence, we notice a problem here: The *naive* assumption is that a particular word does not influence the chance of encountering other words in the same document. In practice, the conditional independence assumption is indeed often violated, but naive Bayes classifiers are known to perform still well in those cases

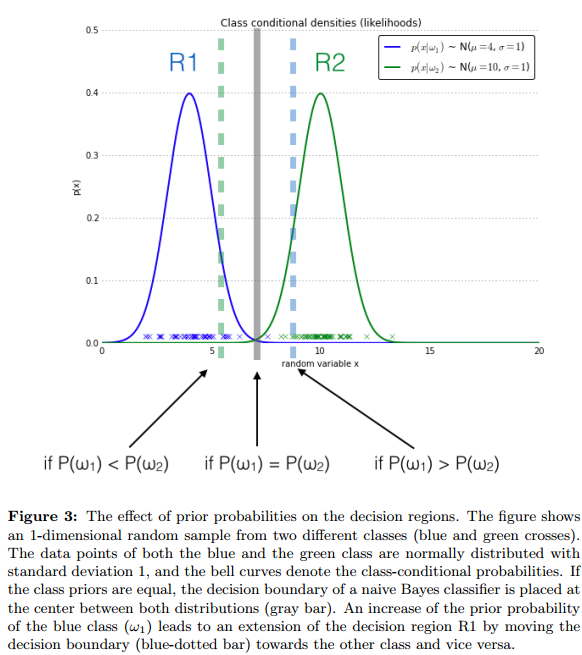
**Prior Probabilities**

In the context of pattern classification, the prior probabilities are also called class priors, which describe ‘the general probability of encountering a particular class’.

If the priors are following a uniform distribution , the posterior probabilities will be entirely determined by the class-conditional probabilities and the evidence term. And since the evidence term is a constant , the decision rule will entirely depend on the class-conditional probabilities.

Eventually, the *a priori* knowledge can be obtained, e.g., by consulting a domain expert or by estimation from the training data and a representative sample of the entire population. The  
maximum-likelihood estimate approach can be formulated as





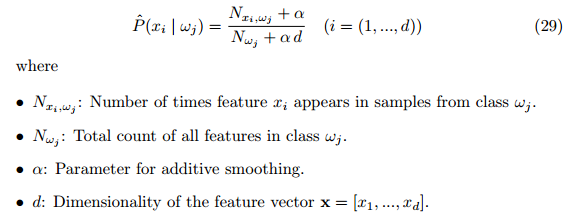
**Evidence**

The evidence P(x) can be understood as the probability of encountering a particular pattern x independent from the class label. The evidence can be calculated as followes ( stands for “complement” and basically translates to ‘not class wj.’)



**Additive Smoothing**

In order to avoid the problem of zero probabilities , an additional smoothing term can be added to the multinomial Bayes model. The most common variants of additive smoothing are the so-called Lidstone Smoothing ()and Laplace smoothing ()



**Variants of the Naïve Bayes Model**

1. Mutli-variate Bernoulli Navie Bayes

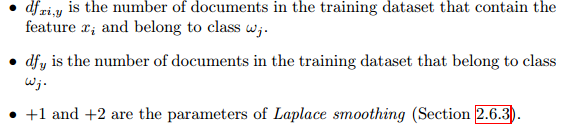
Base on binary data:



Let  be the maximum-likelihood estimate that a particular word (or token) xi occurs in class wj



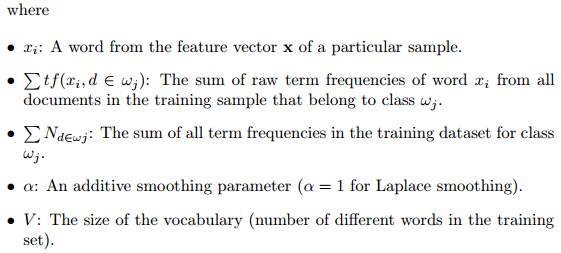
Where



1. Multinomial Naïve Bayes

The term frequencies can then be used to compute the maximum-likelihood estimate based on the training data to estimate the class-condtional probabilities in the multinomial model:





The class-conditional probability of encountering the text x can be calculated as the product from the likelihoods of the individual words



Empirical comparisons provide evidence that the multinomial model tends to outperform the multi-variate Bernoulli model if the vocabulary size is relatively large. However, the performance of machine learning algorithms is highly dependent on the appropriate choice of features. In the case of naive Bayes classifiers and text classification, large differences in performance can be attributed to the choices of stop word removal, stemming, and token-length. In practice, it is recommended that the choice between a multi-variate Bernoulli or multinomial model for text classification should precede comparative studies including different combinations of feature extraction and selection steps.

3. Gaussian Naïve Bayes

Under the assumption that the probability distributions of the features follow a normal (Gaussian) distribution, the Gaussian naïve Bayes model can be written as follows:



The class-conditional probability can than be computed as :



Reference:

1. Naive Bayes and Text Classification I Introduction and Theory