Naïve Bayes

Chapter · January 2016

DOI: 10.1007/978-1-4899-7502-7_581-1

1 author:

Geoffrey I Webb
Monash University (Australia)
449 PUBLICATIONS **12,956** CITATIONS

SEE PROFILE

To be published in C Sammut and G I Webb (Eds) Encyclopedia of Machine Learning and Data Mining, Springer, 2017.

Naïve Bayes

Author: Geoffrey I Webb, Monash University

Synonyms

Idiot's Bayes, Simple Bayes

Definition

Naïve Bayes is a simple learning algorithm that utilizes Bayes' rule together with a strong assumption that the attributes are conditionally independent given the class. While this independence assumption is often violated in practice, naïve Bayes nonetheless often delivers competitive classification accuracy. Coupled with its computational efficiency and many other desirable features, this leads to naïve Bayes being widely applied in practice.

Motivation and Background

Naïve Bayes provides a mechanism for using the information in sample data to estimate the posterior probability $P(y \mid \mathbf{x})$ of each class y given an object $\mathbf{x}$. Once we have such estimates we can use them for classification or other decision support applications.

Naïve Bayes' features include the following.

- *Computational efficiency*: training time is linear with respect to both the number of training examples and the number of attributes and classification time is linear with respect to the number of attributes and unaffected by the number of training examples.
- *Low variance*: because naïve Bayes does not utilize search, it has low variance, albeit at the cost of high bias.
- *Incremental learning*: naïve Bayes operates from estimates of low order probabilities that are derived from the training data. These can readily be updated as new training data are acquired.
- *Direct prediction of posterior probabilities*.
- *Robustness in the face of noise*: naïve Bayes always uses all attributes for all predictions and hence is relatively insensitive to noise in the examples to be classified. Because it uses probabilities, it is also relatively insensitive to noise in the training data.
- *Robustness in the face of missing values*: because naïve Bayes always uses all attributes for all predictions, if one attribute value is missing, information from other attributes is still used, resulting in graceful degradation in performance. It is also relatively insensitive to missing attribute values in the training data due to its probabilistic framework.

Structure of Learning System

Naïve Bayes is a form of Bayesian Network Classifier based on Bayes' rule

$$P(y \mid \mathbf{x}) = P(y)P(\mathbf{x} \mid y)/P(\mathbf{x}) \quad (1)$$

together with an assumption that the attributes are conditionally independent given the class. For attribute-value data, this assumption entitles

$$\mathrm{P}(\mathbf{x} \mid y) = \prod_{i=1}^{n} \mathrm{P}(x_i \mid y) \tag{2}$$

where x_i is the value of the i^{th} attribute in $\mathbf{x}$, and n is the number of attributes.

$$\mathrm{P}(\mathbf{x}) = \prod_{i=1}^{k} \mathrm{P}(c_i)\mathrm{P}(\mathbf{x} \mid c_i) \tag{3}$$

where k is the number of classes and c_i is the i^{th} class. Thus, (1) can be calculated by normalizing the numerators of the right-hand-side of the equation.

The resulting classifier uses a linear model, equivalent to that used by logistic regression, differing only in the manner in which the parameters are chosen.

For categorical attributes, the required probabilities $\mathrm{P}(y)$ and $\mathrm{P}(x_i \mid y)$ are normally derived from frequency counts stored in arrays whose values are calculated by a single pass through the training data at training time. These arrays can be updated as new data are acquired, supporting incremental learning. Probability estimates are usually derived from the frequency counts using smoothing functions such as the Laplace estimate or an m-estimate.

For numeric attributes, either the data are discretized (see discretization), or probability density estimation is employed.

In text mining, two variants of naïve Bayes are often employed (McCallum & Nigam, 1998). The *multi-variate Bernoulli model* utilizes naïve Bayes as described above, with each document represented as a vector of binary variables, each representing the presence or absence of a specific word. However, only the words that are present in a document are considered when calculating the probabilities for that document.

In contrast, the *multinomial model* uses information about the number of times a word appears in a document. It treats each occurrence of a word in a document as a separate event. These events are assumed independent of each other. Hence the probability of a document given a class is the product of the probabilities of each word event given the class.

Cross References

Bayes Rule
Semi-naïve Bayesian learning
Bayesian methods
Bayesian networks
Bayesian network classifiers

Recommended Reading

Lewis, D. (1998) Naive Bayes at forty: The independence assumption in information retrieval. In Machine Learning: ECML-98, Proceedings of the 10th European Conference on Machine Learning, Chemnitz, Germany, Springer, Berlin, pp 4-15.

Andrew McCallum and Kamal Nigam (1998) A comparison of event models for Naive Bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization, AAAI Press, pp. 41-48.