nb: Naive Bayes Model in R

Renato Rodrigues Silva[1]

1 Federal University of Goias

Summary

One of the most important tasks in statistics is to classify experimental units and or individuals and one of the simplest algorithm to make classification is the naive Bayes (NB) (Stuart and Norvig 2003).

The algorithm assumes each instances belongs to only one of the categories, which are finites and mutually exclusives. Moreover, it is assumed all features are independents from each other, which is unrealistic in most of the times but provides the efficiency and simplicity to the algorithm. In a overview, the algorithm classifies the instances into respectively groups, using Bayes' theorem (Koller and Friedman 2009; Al-Aidaroos, Bakar, and Othman 2010).

Naive Bayes classifier has been widely used in several areas of knowledge. (El Kourdi, Bensaid, and Rachidi 2004) used the Naive Bayes to classify non-vocalized Arabic web documents into pre-defined categories. They found an average accuracy over all categories of 68.78 %. where the maximum accuracy found were equal to 92.8%. (Gong and Yu 2010) applied the naive Bayes to implement automatic classification of the web chinese text. Experimental results showed that the classification system were efficient and accurate. In the diagnosis medical area, (K.M. Al-Aidaroos and Othman 2012) compared the performance of naive Bayes with others fifteen famous classifier used in the literature, they concluded that naive Bayes had high performance in most of the examined medical problems.

The naive Bayes has been implemented in many programming languages. Nowadays, it is very easy to find a tutorial about how to performance naive Bayes using python (Rossum 1995) or R (Team 2018).

R is a programming language and software for data analysis supported R Foundation for Statistical Computing (Team 2018). One of an advantages to use R is the large number of packages available by users and scientists around the world. Up to our knowledge, there some R packages where naive Bayes is implemented. One of the options is the Bnlearn package (Scutari 2010), a R package for graphical models in general. Among the various functions defined in the library, they implemented bernoulli and multinomial naive Bayes, where feature and categorical variables follows multinomial or bernoulli distribution. It is possible the users run the Gaussian naive bayes, but in that case, the users must provide the graphical model by theirselves.

Gaussian, bernoulli and multinomial naive Bayes are implemented in e1071 (Meyer et al. 2017) and mlr (Bischl et al. 2016) also. However, none of them allows modelling the qualitative/discrete and real valued features at the same time.

Here, it is being purposed a new R package for naive Bayes models. Standards variants of naive Bayes models: Gaussian, bernoulli and multinomial naive Bayes were implemented. However, in `nb` package is possbile to mixed quantitative and qualitative features. It is

possible because this implementation takes advantages of the fact that features are independent each other. Hence, The package uses the normal distribution for modelling real valued features, and multinomial / bernoulii distribution for modelling discrete / qualitative features. The library contains four functions: `naive.bayes`, `post.prob`, `classify` and `confusion.matrix`.

The `naive.bayes` function estimates parameters of naive Bayes model via maximum likelihood estimation, The input of this function is a data.frame `x` which represents the dataset to be analyzed and `ct`, which is a string of the name of categorical variable that classifies the instances into groups. If `x` contain only real valued features the `naive.bayes` estimates the parameters of a Gaussian naive Bayes model. On the other hand, if there are only qualitative / discrete features the parameters are estimated from Multinomial naive Bayes model, otherwise, the function uses the gaussian distribution for real valued features and multinomial / bernoulli distribution for qualitative / discrete features because the features are assumed to be independent.

The `post.prob` computes the probability posterior of a new instances belong to a class given the a set of observed features. The output is a vector with posterior probability for each category. In turn, the function `classify` computes which the category has higher posterior probability.

$$\arg\max_{j \in (1,...,k)} p(\mathbf{z}) \prod p(x_i | \mathbf{Z} = \mathbf{z})$$

where $\mathbf{Z}$ the random vector that represents the categorical variables and x_i the features.

Furthermore, the `nb` packages has another advantage in comparison to other packages. It is already implemented a `confusion matrix` function for measuring the predictive performance of model. The confusion matrix is defined via 5 or 10 fold cross validation. To create the folds, the function createFolds of caret package (Kuhn 2008) is invoked. Currently, `nb` package is hosted at [github](#).

References

Al-Aidaroos, K. M., A. A. Bakar, and Z. Othman. 2010. "Na?Ve Bayes Variants in Classification Learning." In *2010 International Conference on Information Retrieval Knowledge Management (Camp)*, 276–81. https://doi.org/10.1109/INFRKM.2010.5466902.

Bischl, Bernd, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones. 2016. "Mlr: Machine Learning in R." *Journal of Machine Learning Research* 17 (170):1–5. http://jmlr.org/papers/v17/15-066.html.

El Kourdi, Mohamed, Amine Bensaid, and Tajje-eddine Rachidi. 2004. "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm." In *Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages*, 51–58. Semitic '04. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1621804.1621819.

Gong, Z., and T. Yu. 2010. "Chinese Web Text Classification System Model Based on Naive Bayes." In *2010 International Conference on E-Product E-Service and E-Entertainment*, 1–4. https://doi.org/10.1109/ICEEE.2010.5660869.

K.M. Al-Aidaroos, A.A. Bakar, and Z. Othman. 2012. "Medical Data Classification with Naive Bayes Approach." *Information Technology Journal* 2 (11):1166–74. https://doi.org/10.3923/itj.2012.1166.1174.

Koller, Daphne, and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge: MIT Press.

Kuhn, Max. 2008. "Building Predictive Models in R Using the Caret Package." *Journal of Statistical Software* 28 (5):1–26.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and Friedrich Leisch. 2017. *E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)*. TU Wien. https://CRAN.R-project.org/package=e1071.

Rossum, Guido. 1995. "Python Reference Manual." Amsterdam, The Netherlands, The Netherlands: Centre for Mathematics; Computer Science; CWI (Centre for Mathematics; Computer Science).

Scutari, Marco. 2010. "Learning Bayesian Networks with the Bnlearn R Package." *Journal of Statistical Software, Articles* 35 (3):1–22. https://doi.org/10.18637/jss.v035.i03.

Stuart, Russell, and Peter Norvig. 2003. *Artificial Intelligence: A Modern Approach*. 2nd ed. New Jersey: Prentice Hall.

Team, R Core. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.