Bayes Classifier

Penghuan Huang*

Oct 27,2019

Abstract

Since Bayes classifier plays an important role in many areas of data science like learning $p(y|x)$, in this short article I try to give a brief introduction to Bayes classifier concerning its four categories, theory and its correction in real study. Though in all the four classifiers the Naive Bayes will be the main focus.

Definition

The Bayes clssifier is:

$$C^{Bayes}(x) = \underset{r\in\{1,2,...,n\}}{\operatorname{argmax}} P(Y = r|X = x)$$

r refers to the possible values of Y, ranging from 1 to K. It means that there are N classes. This function gives out the best value of Y, namely r, which maximizes the conditional probability of Y equal to r condition on the observed x, thus classifies the point x to the class C(x). In other words, Bayes classifier gives out the most likely value of r knowing x.

Fundamental Knowledge

The thoughts and calculation of Bayes classifier is based on Bayesian theorem. The theorem can be stated in the following equation:

$$P(r|x) = \frac{P(r)P(x|r)}{P(x)} = \frac{P(x,r)}{P(x)}$$

*Department of Economy, Xiamen University. Email: 1131936012@qq.com

P(x) can be calculated as following:

$$P(x) = P(r_1)P(x|r_1) + \cdots + P(r_n)P(x|r_n)$$

Categories

Bayes classifier has four main categories—Naive Bayes, TAN, BAN, and GBN.

Naive Bayes[3]: In machine learning, Npaive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independence assumptions between the features. This strong assumption is the reason why we call it 'naive', which can be expressed in the mathematical form:

$$P(r|x) = \frac{P(r)P(x|r)}{P(x)} = \frac{P(r)}{P(x)} \prod_{i=1}^{n} P(x_1|r)$$

Though this classification method is widely used, there actually exists a problem in such a strong assumption, as we know features are often related with each other in real life. For example, when learning people's income, the two features—sex and height, are usually highly related.

TAN: Its full name is Tree Augmented Naive-Bayes, which is an extension of Naive Bayes since it allows dependency between features. But there is one restriction that each feature can only be a determinant of at most one other feature. TAN schematic diagram is as follows, looking like a tree:

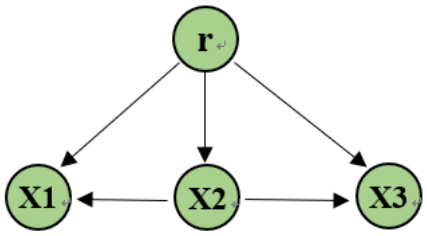

Figure 1: TAN model[1]

Being pointed to by an arrow indicates being an determinant of the respective feature. "r" is a class variable and "x_i" are the features.

[1]The idea of such diagrams was generated from Deng and Fu's article[1] (2008).

BAN: Its full name is BN Augmented Naive-Bayes. BAN further relax the strong assumption of independence, with no restriction on the mutual effect of features. It can be understood as follows:

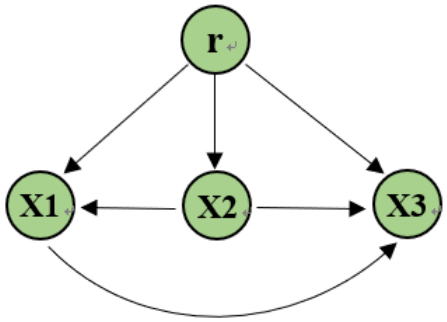

Figure 2: BAN model

GBN: GBN is the most free classifier out of the four. Its full name is General Bayesian Network. The "r" point is treated the same as other "x" points, which means no restriction on dependency of all the points. The following schematic diagram can help to understand:

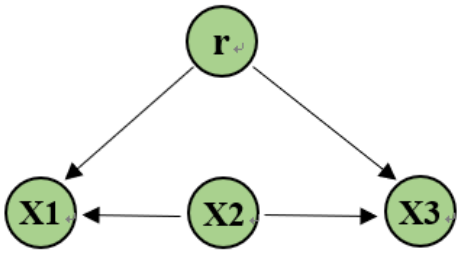

Figure 3: GBN model

From experiment results from the reference[1], we can conclude that Naive Bayes still performs well for small data sets even though its assumption seems unrealistic. Moreover, it is easy in calculation. Therefore, the Naive Bayes classifier is useful and widely adopted by researchers.

Laplacian correction

There is another problem when using Naive Bayes that needs correction. When the data set is small and we use $\hat{P}(x_i|r) = \frac{|D_{r,x_i}|}{|D_r|}$ to calculate $\hat{P}(x_i|r)$, the extreme situation

would likely appear:

$$\exists i \in n, \hat{P}(x_i|r) = 0$$

D is the data set. $|D_{r,x_i}|$ represents the number of samples that r and x_i appear. $|D_r|$ is the number of samples exhibiting the character r.

From the above calculation formula of Naive Bayes, $P(r|x)$ will always equal to zero if such an extreme situation appears, no matter how likely the other x_i suggests r to appear. It is just like a saying, "Tar with the same brush."

To prevent such an situation, Laplacian correction is needed. The correction can be concluded in the following two equations:

$$\begin{cases} \hat{P}(x_i|r) = \frac{|D_{r,x_i}+1|}{|D_r+N_i|} \\ \hat{P}(r) = \frac{|D_r+1|}{|D+N|} \end{cases}$$

N_i is the number of possible classification "r", while N is the number of possible values of the ith feature. For example, the possible values of sex is "male" and "female", thus N_i should be counted as 2.

References

[1] 邓甦,付长贺.四种贝叶斯分类器及其比较[J]. 沈阳师范大学学报(自然科学版),2008(01):31-33.

[2] 朴素贝叶斯分类器(Naive Bayesian Classifier)[EB/OL]. 网址：https://blog.csdn.net/qq_32690999/article/details/78737393

[3] Wikipedia: https://en.wikipedia.org/wiki/Bayes_classifier

[4] Wikipedia: https://en.wikipedia.org/wiki/Naive_Bayes_classifier