Naive Bayes

It can be explained that if a essence of an event is unknown, people can rely on the occurrence of events related to the particular essence of the event to determine the probability of its essential attribute, which can be translated into mathematical explanation: The more events that support a certain attribute occur, the more likely the attribute is true.

you don't know the exact essence of anything, you can rely on the occurrence of occurrences connected to that thing's specific essence to determine the probability of its fundamental quality. It is defined as the occurrence of an event that supports a specific attribute in mathematical terms.

The Bayes technique, unlike other statistical methods, is dependent on subjective judgment. Here we first explain the Naïve Bayes from a mathematical point of view, and we translated the function into a way that easily to understand:

Class Prior Probability

likelihood

文本, 信件

描述已自动生成

Posterior Probability

Predictor Prior Probability

As part of the Inferential Statistics of the project we applied Naive Bayes to help predict whether a person has heart disease, relying on several factors, such as age, chest pain type, fasting blood sugar, maximum heart rate achieved, etc. Here the dataset has 303 observations and 14 variables shown as follows:

图片包含 文本

描述已自动生成

We start by setting random numbers which is used to ensure that the results remain unchanged during each iteration. Secondly, we divided the dataset into two parts: training and testing. We choose 80% of dataset for training and the other 20% for testing and:

图形用户界面, 应用程序

中度可信度描述已自动生成

We observed the details of our model, and we find categorical data which is different from numeric data. The former is still the traditional probability, the first field of the latter is the average value, and the second field is the standard deviation.

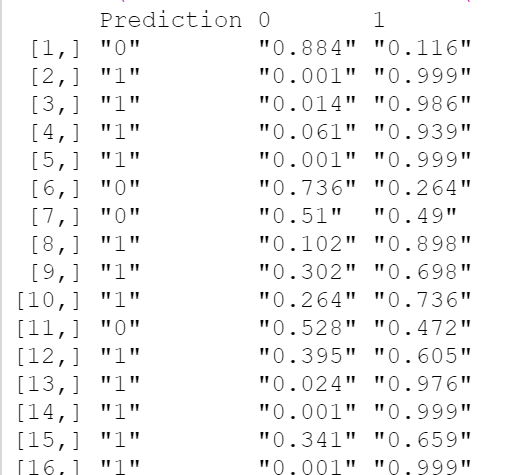


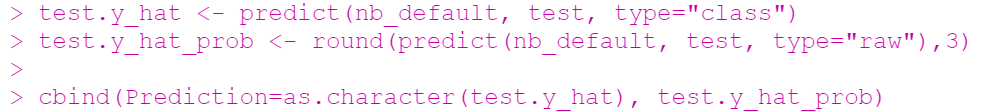
手机屏幕截图

描述已自动生成文本

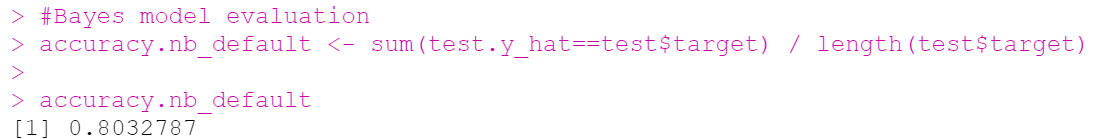
描述已自动生成

Then we used R code as follows to make Bayes' prediction for validation and reserved three decimal and we could see from the following examples that the probability has been normalized.





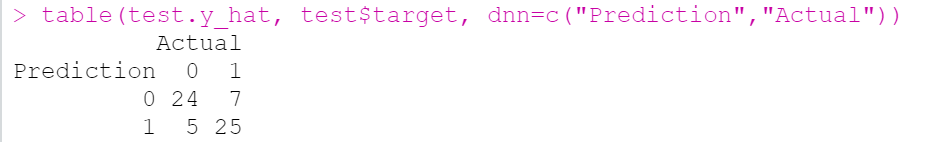
We then evaluated the accuracy of the Bayes model, and the value could as high as 80%.



Let us make a comparison with the real situation.

表格

描述已自动生成

We finally built a cross table to compare the actual situation and the Bayesian model prediction, as Naive Bayes is essentially a binary classification problem, the result can only be shown in 0 as well 1.

From the result, we found the accuracy of the model is very high at 80% percent; and for testing part, it can also be found that the established Naive Bayes model has high predictive performance. The target is actual to be 0, 24 predicted to be 0, and only 5 are predicted to be 1; and the target are actual to be 1, there are 25 predicted to be 1, and 7 predicted to be 0.

Here, we choose Naive Bayes to make prediction due to the small size of the sample and the characters are independent to each other. And the Pros Naive Bayes model are:

Firstly, it is simple and convince to forecast the test data set and it also good at multi-class prediction which was used in this project.

Secondly, a Naive Bayes model outperforms alternative models such as logistic regression and decision tree as well as the less training data was needed when independence hypothesis holds.

Thirdly, in comparison to numerical input variables, it performs well with categorical input variables (s). A normal distribution is assumed for numerical variables (bell curve, which is a strong assumption).

However, the important limitation of [Naive Bayes](https://courses.analyticsvidhya.com/courses/naive-bayes?utm_source=blog&utm_medium=naive-bayes-explained) is the assumption of independent predictors. In actual life, getting a collection of predictors that are totally independent is sometimes impossible.