Naïve Bayes Classifier and its realization with R

Zihan Zhang*

March 10, 2019

Abstract

In this note, I would like to have a basic overview about the Naïve Bayes Classifier, which is a useful and fundamental probabilistic model in machine learning, as well as its principle in need and realization with R.

1 Foundation – Bayes Theory

The Naïve Bayes Classifier is based on the Bayes Theorem (alternatively Bayes rule), describing that the correlation between conditional probabilities of two events. It could be written in the mathematical language as follows.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{1}$$

Where $P(A|B)$ means that the incidence of A given event B is observed, while $P(A)$ means the probability of observing event A independently. For same reason, the explanation of $P(B|A)$ and $P(B)$ are similar. As a result, we would be able to predict $P(A|B)$ with the knowledge of$P(A)$, $P(B)$ and $P(B|A)$.

The proof of the Bayes theorem is quite straight forward. According to the definition of conditional probability,

$$P(A|B) \times P(B) = P(A \cap B) = P(B \cap A) = P(B|A) \times P(A) \tag{2}$$

Bayes theorem is the result of the equation above after transposing.

2 Naive Bayes Classifier

Bayes Classifier, based on Bayes Theorem, making it possible to make some predictions according some given information. Take an example to illustrate, peer-to-peer (P2P) companies make living through lending money to borrowers. However, they would never know whether the borrower they are facing to would defect when their contract

*2016 WISE IUEC, Student ID: 15220162202518

expires. In other word, the company don't know whether they can get the principal back in the future. So, making some predictions based on historical data is sensible. Naïve Bayes Classifier provides a method of making it into realization.

We define $\mathbf{x} = (x_1, x_2, ..., x_n)$, where each x_i manifesting dimensions of features of individuals in data. Besides, there's a prerequisite requiring the dimensions to be **independent** with each other. Therefore, what we are wondering could be represented as

$$P(C_i \mid \mathbf{x}_i) = \frac{P(C_i) \times P(\mathbf{x}_i \mid C_i)}{P(\mathbf{x}_i)} \tag{3}$$

where C_i is a discrete variable representing the categories of individual i. Within the P2P company case, it would be a binary variable measuring whether the borrower (the individual) is defected or not. Finally, having got the estimated conditional probability, the model would like to compare them and make the prediction. In other word, eventually, C_i would be given the value that makes $P(C_i|\mathbf{x}_i)$ to be the highest.

Next, I would like to introduce the mechanism of Naive Bayes Classifier. Actually, we only consider the numerator of (3) since C_i only appears there. We could represent it as:

$$P(C_i|\mathbf{x}) \propto P(\mathbf{x}|C_i)P(C_i) \tag{4}$$

According to the conditional probability formula and independence among x_i,

$$\begin{aligned} P(\mathbf{x}|C_i)P(C_i) &= P(\mathbf{x}, C_i) = P(x_1, x_2, ..., x_n, C_i) \\ &= P(x_1|x_2, x_3, ..., x_n, C_i)\ P(x_2, x_3, ..., x_n, C_i) \\ &= P(x_1|x_2, x_3, ..., x_n, C_i)\ P(x_2|x_3, x_4, ..., x_n, C_i)\ P(x_3, x_4, ..., x_n, C_i) \\ &= P(x_1|x_2, x_3, ..., x_n, C_i)\ P(x_2|x_3, x_4, ..., x_n, C_i) \times ... \times P(x_n|C_i)\ P(C_i) \\ &= P(x_1|C_i)P(x_2|C_i) \times ... \times P(x_n|C_i)P(C_i) \\ &= P(C_i) \prod_{k=1}^{n} P(x_k|C_i) \end{aligned} \tag{5}$$

With (5), we would be able to quantify the $P(C_i|\mathbf{x})$ by computing $P(x_k|C_i)$ for each dimension. Usually, we assume each x_k come from a Gaussian Distribution and we can settle the coefficients according to the training set. Thus, we would be able to calculate the conditional probability for different C_i and predict what is the most likely value of C_i. In addition, we have discussed the loss function in class, this idea of finding the highest probability also confirms the application of 0-1 loss function.

3 Realization in R

In this section, I constructed a model with the idea mentioned in previous sections and generated some data to test it.

Suppose There is a group of borrowers (test set) with two features observable (feature 1 and feature 2). The task for us is to find out who will defect in the future. The following diagram manifests the distribution of those individuals within the test set.

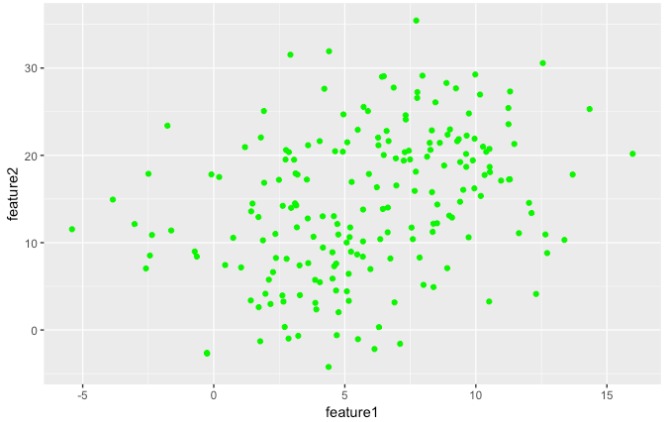

Figure 1: Test set

Also, we have some past data shown in the following diagram (training set) to settle the coefficients in the model (the μ and σ of Gaussian distribution). Sky-blue point represents the individuals who is not defected, while the dark-blue points shows defected borrowers.

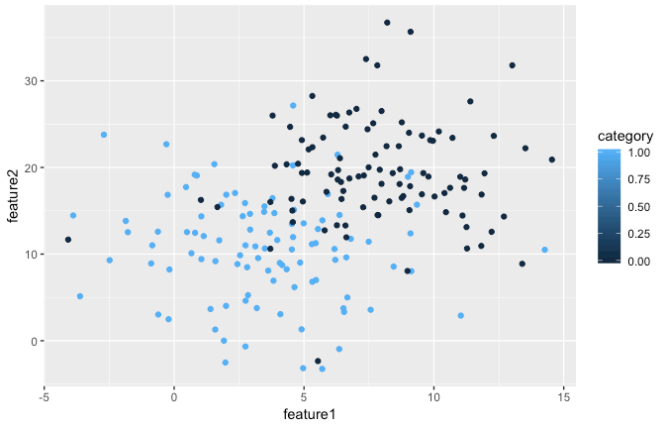

Figure 2: Training set

Before making prediction, we firstly check the correlation between the feature 1 and 2, making sure they are independent to each other.

From the correlation plot (Figure 3), we can see the fitted straight line in both down-left and up-right diagram are almost flat, meaning the correlation between them

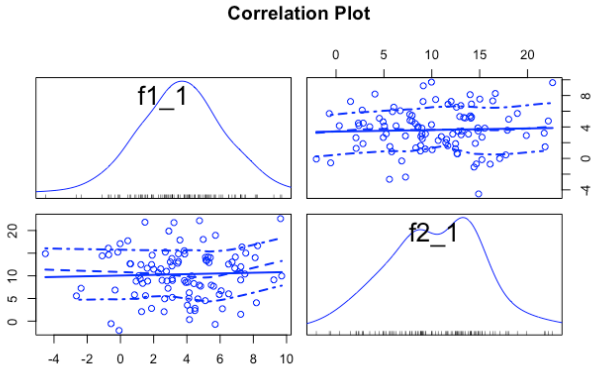

Figure 3: Correlation plot

are extremely insignificant. This result is reasonable because the data is generated randomly with specific distributions and there are no links between two features when generating.

Therefore, after training, the prediction result is as follows.

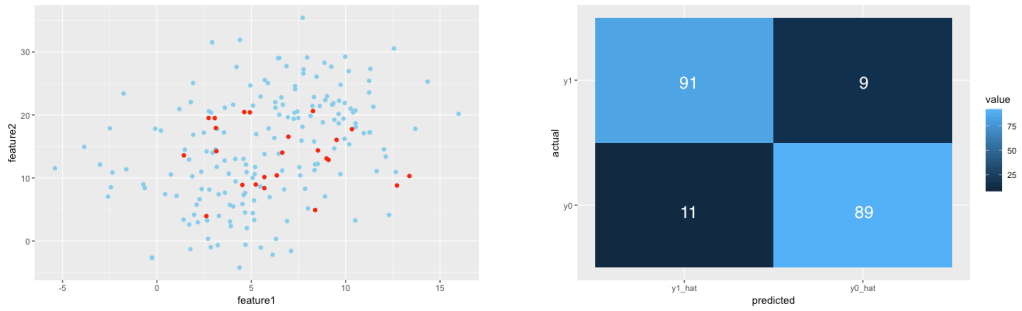

Figure 4: Prediction result

The blue points in the left figure show the individuals with a correct prediction by models while the red points shows converse ones. The heat map on the right indicates the magnitude of each genres. The brighter the box is, the more data is here. In general, we could see the prediction is precise to a large extent.

References

[1] Bayes Theorem, https://en.wikipedia.org/wiki/Bayes%27_theorem.

[2] Naive Bayes Classifier, https://en.wikipedia.org/wiki/Naive_Bayes_classifier.

[3] Jiaming Mao, Foundations of Statistical Learning,
https://jiamingmao.github.io/data-analysis/assets/Lectures/Foundations_of_Statistical_Learning.pd