Naive Bayes

Author: Zhang Xiaozheng Date: 2019/10/31

When we do classification, we want to find the class of a given data. From the perspective of probability theory, we want to find the distribution function *P(c|x)*. So we can choose the class which makes it's distribution function max.

Some algorithm like decision tree or logistic regression just want to find *P(c|x)* directly. But what Bayes Classifier do is using Bayes formula.

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)}$$

So our task turns into find *P(c)* and *P(x|c)*.

To find *P(c)* and *P(x|c)*, we need to use MLE(Maximum Likelihood Estimation) to estimate *P(x|c)*.

But if we directly estimate *P(x|c)*, there will be so many possible combinations so it will takes a long time to calculate. So what Naive Bayes do is to suppose all the features in ***X*** is independent. So we can rewrite our equation.

$$P(x|c) = \prod_{i=1}^{K} P(x_i|c)$$

In this way, we just need to estimate these *K* distribution function separately.

MLE

After an observation, we have a data set

$$X = \{x_1, x_2, \dots, x_n\}$$

Suppose we already know the form of its probability density function which is:

$$f(x; \theta_1, \theta_2, \dots, \theta_k)$$

So we define Likelihood Function:

$$L(\theta_1, \theta_2, \dots, \theta_k) = \prod_{i=i}^{k} f(x_i; \theta_1, \theta_2, \dots, \theta_k)$$

As this data has been observed by us, so we can say that the probability of this kind of sample is quite large. What MLE do is to find a group of *θ* which can make the value of Likelihood Function max.

So after use MLE, for a data set *D*,we can have:

$$P(y = c_k) = \frac{|D_k|}{|D|} \ \ where\ |D_k|\ is\ the\ number\ of\ y\ whose\ label\ is\ c_k$$

$$P(x_i|c_k) = \frac{|D_{ik}|}{|D_k|} \ \ where\ |D_{ik}|\ is\ the\ number\ of\ x\ whose\ label\ is\ c_k\ and\ whose\ feature\ is\ x_i$$

Because *P(x)* has nothing to do with *c*, so we just need to maximize *P(c) P(x|c)* so that we can maximize *P(c|x)* too.

Problem

What if there is no such *x* whose feature is *xi* and whose label is *ck*? In this case, the *P(x|c)=0*, no matter what other features are, the *P(c|x)* will always equal to 0 which not make sense. So we use "Laplace smooth" for our formula:

$$P(y = c_k) = \frac{|D_k| + 1}{|D| + N}$$
$$P(x_i | c_k) = \frac{|D_{ik}| + 1}{|D_k| + N_i}$$

where *N* is the number of all the possible classes of *y*, *Ni* is the number of all the possible classes of feature *x*.

In this way, it is impossible that *P(x|c)=0*.