Naïve Bayes

Introduction

It is one of the easiest **not structured** supervised algorithms. It's based on conditional probability. It is used in classification problems. It can be used only in case of independent features.

What is the mathematics behind?

It is used to map input feature of x1, x22, x3, ….xn into labels y1, y2,….. We just choose the label with the highest probability.

For example, if we have bunch of mails, some of these mails for Sara, and the other mails for Chris.

How can we discriminate between Chris and Sara's mails? By features.

This features come from the words in both Sara and Chris mails. Let's assume features for both mails, we assume that Sara mails contain love with probability 0.5, deals with 0.2, and life with 0.3. But for Chris mails contain love with 0.1, deals with 0.8, and life with 0.1.

If we have a new mail, and we don't know if this mail from Sara or from Chris, what should we do? How could we discover if this mail was sent by Chris or Sara?

In this case, we will investigate the mail words,

- 1st case: If there are words for love and life, then it is more likely to be from Sara.
- 2nd case: If this mail contains a deal and life words, then it is more likely to be sent by Chris.

How did we decide in the above case?

We decided according to calculations:

In the 1st case, we calculate;

p(love and life | Chris) * p(Chris)= 0.1 * 0.1 * 0.5 = 0.0005

But for Sara, we calculate:

p(love and life | Sara) * p(Sara) = 0.5 * 0.3 * 0.5 = 0.075

In the 2nd case, we calculate:

p(Deal and life | Chris) * p(Chris) = 0.8 * 0.1 * 0.5 = 0.04

But for Sara, we calculated:

p(Deal and life | Sara) * p(Sara)= 0.2 * 0.3 * 0.5 = 0.03

Let's ask another question what the p(Chris | "life and deal")? Which means *P(Y=y|X=(x1, x2, …, xm))*

To answer this question, we will apply the Bayes equation directly, or which is called posterior probability,

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

P(Chris | "life and deal") = [($\frac{P(Life\ and\ deal\ |\ Chris)\ P(Chris)}{P(Life\ and\ deal\ in\ all\ mails)}$)] = [0.04 /(0.04+0.03)]

= 0.571

P(Sara | "life and deal") = [($\frac{P(Life\ and\ deal\ |\ Sara)\ P(Sara)}{P(Life\ and\ deal\ in\ all\ mails)}$)] = [0.03 /(0.04+0.03)]

= 0.42

From the above example, we can understand how to apply the byes theorem directly.

Why is it called Naïve?

Because it is used in intuitive application. Also, it is simple, since it is **not structured**. It is assumed to be used in **independent** features. It can't be used in case of dependent features. The independent feature happen rarely in normal life. So, it is used in limited number applications.

Independent features! What is the meaning of the independent features?

The independent features mean that each feature is independent from other features. In the above example, the love words feature is independent from deal words feature and from the life words feature. Each one can't depend on other feature.

Another example can illustrate this more, if we expect our main high way in our city,

Let's assume that every day except week ends, in the morning and the evening, it will have a heavy traffic. In the afternoon and the night, the way will have light traffic.

While in the week end, in the morning and the evening, it will have light traffic but it will have heavy traffic in the afternoon and in the night.

In the traffic state example above, the day (week end or not) feature is independent from the day time (morning, afternoon, evening, and night). The day time can't affect the day.

Now, what is the meaning of not structured algorithm? It means that the algorithm doesn't consider the order of words. For example, if we tried to search for the phrase ' Marwa Matar', the Naive Bayes algorithm will fail to find this phrase in the document. It can't return Marwa Matar as a full name with the same order, but rather it only consider Marwa only as a feature and Matar as other feature. The algorithm can't structure the Marwa as a 1st name and Matar as last name since Naïve Bayes doesn't consider the words order.

What the Naïve Bayes Advantages and disadvantages?

Advantages:

Naïve Bayes is used mainly in text classification such as spam or not spam mails, who is more likely to be the owner of this book according to the words in this book, who is the more likely to the owner of this mail.

Disadvantages:

Naïve Bayes can't be used in:

- In case of structured data or ordered words since the Naïve Bayes doesn't consider the word context or words order.
- In case of dependent words.