Machine Learning

4. week

- Probability and Conditional Probability
- Bayes Theorem
- Naïve Bayes Classifier

Probability

The term shows the occurring likelihood of each situation in a random process A and is represented by P(A) notation.

This simple concept can be defined as prior or unconditional or marginal probability of the random process A.

Example 1

When a coin toss process is repeated twice, the probability of both results being heads can be computed as below.

$$P(A \cap B) = P(A)P(B) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

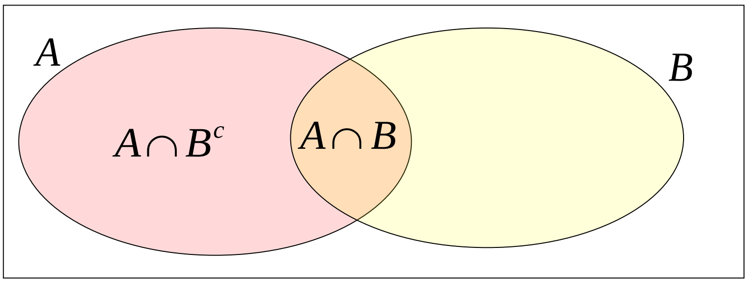
Because two process are independent to each other, intersection of them means product of them.

Conditional Probability

If a random process A depends on another random process B, then the using only prior probability is not enough. Therefore conditional or posterior probability term represented by P(A|B) is used, and is computed as follow.

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

Conditional Probability



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Example 2

We have 3 blue and 4 yellow balls at the first jar, and 5 blue and 2 yellow balls at second jar. What is the probability of picking a blue ball from a random selected jar?

A: Selection of first jar

B: Selection of second jar

C: Selection of blue ball

$$P(C) = P(A)P(C \mid A) + P(B)P(C \mid B) = \frac{1}{2}\frac{3}{7} + \frac{1}{2}\frac{5}{7} = \frac{4}{7}$$

Bayes Theorem

According to this theorem, any conditional probability can be computed by using known conditional and marginal probabilities.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Here, by changing places of A and B variables:

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

Bayes Teoremi

Bayes theorem is proposed by Thomas Bayes (1702-1761).

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

$$\Downarrow$$

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

Example 3

Suppose that the probability of being cancer of any person in a city 0.8%. Also think that we have a test to determine cancer. Among people whose test result is positive (+), the probability of seeing cancer is 0.98, and the probability of being healthy is 0.97 is for ones having negative (-) result.

$$P(cancer) = 0.008 \qquad P(\neg cancer) = 0.992$$
$$P(+ \mid cancer) = 0.98 \qquad P(+ \mid \neg cancer) = 0.03$$
$$P(- \mid cancer) = 0.02 \qquad P(- \mid \neg cancer) = 0.97$$

Example 3

Thus, what is the probability of being cancer of any person whose test result is positive?

$$P(cancer \mid +) = \frac{0.98 \times 0.008}{P(+) = 0.0376} = 0.2085$$

$$P(\neg cancer \mid +) = \frac{0.03 \times 0.992}{P(+) = 0.0376} = 0.7915$$

$$P(cancer \mid +) < P(\neg cancer \mid +)$$

We should decide as healthy (not cancer).

Naïve Bayes Classifier

Thus, Naïve Bayes classifier chosen the maximum one uses formula below.

$$C = \arg\max_{c_j \in C} P(c_j \mid f)$$

Naïve Bayes Classifier

If dataset contains many independent feature vectors, the posterior probability of a record is the product of conditional probabilities of its all feature vectors.

$$P(f_1, f_2, \ldots, f_n \mid c_j) = \prod_{i=1}^{n} P(f_i \mid c_j)$$

Naïve Bayes Classifier

When we use Bayes Theorem in classification, denominators for all classes are the same. Therefore we can disregard them.

$$P(cancer \mid +) = \frac{0.98 \times 0.008}{\cancel{P(+) = 0.0376}}$$

$$P(\neg cancer \mid +) = \frac{0.03 \times 0.992}{\cancel{P(+) = 0.0376}}$$

Naïve Bayes Classifier

Then the formula of Naïve Bayes transforms to following one.

$$C = \arg\max_{c_j \in C} \left(P(c_j) \prod_{i=1}^{n} P(f_i \mid c_j) \right)$$

Example 4

f_1	f_2	f_3	f_4	*Class*
Yes	No	No	Yes	**B**
Yes	No	No	No	**B**
No	Yes	Yes	No	M
No	No	Yes	Yes	M
Yes	No	No	Yes	**B**
Yes	No	No	No	M
Yes	Yes	Yes	No	M
Yes	Yes	No	Yes	M
No	No	No	Yes	**B**
No	No	Yes	No	M

For the dataset given at left, find the class of record <Yes, No, Yes, Yes>

Example 4

f_1	f_2	f_3	f_4	*Class*
Yes	No	Yes	Yes	B
Yes	No	No	No	B
Yes	No	No	Yes	B
No	No	No	Yes	B
No	Yes	Yes	No	M
No	No	Yes	Yes	M
Yes	No	No	No	M
Yes	Yes	Yes	No	M
Yes	Yes	No	Yes	M
No	No	Yes	No	M

P(Class | Yes, No, Yes, Yes) = ?

$$P(Class = B \mid f_1, f_2, f_3, f_4)$$
$$= \frac{2}{5}\left(\frac{3}{4} 1 \frac{1}{4} \frac{3}{4}\right) = \frac{9}{160}$$
$$P(Class = M \mid f_1, f_2, f_3, f_4)$$
$$= \frac{3}{5}\left(\frac{1}{2} \frac{1}{2} \frac{2}{3} \frac{1}{3}\right) = \frac{1}{30}$$

So, the result must be **B**.

Numerical Data

Bayes Theorem is designed only for categorical feature vectors. To be able to implement it to numerical features, it is thought that samples in numerical feature vector have Normal (Gaussian) distribution. Thus computation only of mean (μ) and standard deviation (σ) values is enough to use following known formula.

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Example 5

V_1	V_2	V_3	V_4	Age	Gender
Yes	No	Yes	No	38	F
Yes	Yes	Yes	No	40	F
Yes	Yes	Yes	No	41	F
No	No	No	No	55	F
No	Yes	No	No	27	M
Yes	Yes	Yes	Yes	30	M
Yes	No	Yes	Yes	35	M
No	No	No	No	42	M
Yes	No	No	No	43	M
Yes	No	No	No	45	M

V_1= Yes, V_2 = Yes,
V_3= No, V_4= No,

Age= 45

Gender = ?

Example 5

Age	Gender
38	F
40	F
41	F
55	F
27	M
30	M
35	M
42	M
43	M
45	M

For P(Age | Gender=F);
$\mu=43.5$ and $\sigma=7.77$
For P(Age | Gender=M);
$\mu=37$ and $\sigma=7.46$

According to Normal distribution;

P(Age=45 | Gender=F) = 0.0504 ~ 5/100
P(Age=45 | Gender=M) = 0.0301 ~ 3/100

Example 5

v_1	v_2	v_3	v_4	Age	Gender
Yes	No	Yes	No	38	F
Yes	Yes	Yes	No	40	F
Yes	Yes	Yes	No	41	F
No	No	No	No	55	F
No	Yes	No	No	27	M
Yes	Yes	Yes	Yes	30	M
Yes	No	Yes	Yes	35	M
No	No	No	No	42	M
Yes	No	No	No	43	M
Yes	No	No	No	45	M

$P(C \mid \text{Yes,Yes,No,No,45}) = ?$

$$P(C = M \mid v_1, v_2, v_3, v_4, Yaş) = \frac{3}{5}\left(\frac{2}{3}\frac{1}{3}\frac{2}{3}\frac{2}{3}\frac{3}{100}\right)$$
$$\cong 0.0018$$
$$P(C = F \mid v_1, v_2, v_3, v_4, Yaş) = \frac{2}{5}\left(\frac{3}{4}\frac{1}{2}\frac{1}{4}1\frac{5}{100}\right)$$
$$\cong 0.0019$$

So, the result must be **F.**

Zero Frequency Problem

If a given class and a feature value never occur together in the training data, then the frequency-based probability estimate will be zero. This is problematic because it will wipe out all information in the other probabilities when they are multiplied.

Zero Frequency Problem

The solution to the zero frequency problem would be to shift your zero frequency to slightly higher.

So, if you have counts of some occurrence in your data, use Laplacian smoothing.

MATLAB Application

```
>edit NBayes_ornek.m
```

Student should study with this sample by using given datasets.

Presentation Task

You can choose one of two algorithms listed below.

- Lojistik Regresyon
- EM (Expectation Maximization)