# Probability Theory in brief

## Fundamental Rules

Given two events and , we define the probability of as follows:

Especially, if and are mutually exclusive, then

We define the **probability of the joint event** as follows:  
ALSO CALLED **product rule**

Also, given a joint distribution on the two events as , we define the **marginal distribution** as follows:  
ALSO CALLED **sum rule**

For the probability of the joint event of more than 2 events , we yield the **chain rule** of probability:

注：为了避免故弄玄虚的符号定义，这里借用了matlab的表示方法。表示序列。

We define “the **conditional probability** of event , given that event is true” as follows:

## Bayes rule

Combining the definition of conditional probability with the product and sum rules yields **Bayes rule**, also called Bayes Theorem:

**EXAMPLE**

Suppose you decide to have a medical test for a breast cancer called a mammogram. If the test is positive, what is the probability you have cancer? That obviously depends on how reliable the test is. Suppose you are told the test has a **sensitivity** of 80%, which means “if you do have cancer, the test will report positive with probability 0.8”. In other words,

where is the event “the mammogram is positive”, and is the event “you have breast cancer”.

Many people conclude they are therefore 80% likely to have cancer. But this is wrong! It ignores the prior probability of having breast cancer, which fortunately is quite low:

Ignoring the prior(前提) is called the base rate fallacy. We also need to take into account the fact that the test may be a **false positive**, a.k.a. **false alarm**. In this example it has a rate of

Combining the three terms above using Bayes rule, we yield,

**EXAMPLE : Generative Classifiers**

We can generalize the previous example to classify **feature vector**s of arbitrary type as follows,

Where is the vector of parameters. 如果觉得看着太烦，不妨把公式里的全删掉，不影响整体理解。

This is called a generative classifier, since it specifies how to generate the data using the **class-conditional density** and the **class prior** .

## Independence

We say and are **unconditionally independent** or **marginally independent**, denoted , **iff** we can represent the joint as the product of two marginals, i.e.

In reality, unconditional independence is rare, because things influent things. Usually, this influence is **mediated** via other variables rather than being direct. We therefore say and are **conditionally independent (CI)** given **iff** the conditional joint can be written as the product of conditional marginals:

There is an important characteristic of CI, indicating iff , then can be decomposed as a product of function and , where is **only variate with** and is **only variate with** , i.e.

**THEOREM: Conditional Independent Joint p.f is decomposible**

iff there exist function and such that

for all such that .

For proof of this theorem, see (Murphy. *Machine Learning – A Probabilistic Perspective*: Exercise 2.8)

This theorem allows us to build large probabilistic models from small pieces.