1. **Bayes Classifier**

目录

[1. Bayes Rule 3](#_Toc60073945)

[2. Outline 4](#_Toc60073946)

[3. Bayesian Decision Based on Minimum Error Rate 4](#_Toc60073947)

[3.1 What is it? 4](#_Toc60073948)

[3.2 Example 5](#_Toc60073949)

[3.3 Theoretical Model 6](#_Toc60073950)

[4. Maximum Likelihood Estimation 7](#_Toc60073951)

[5. Naïve Bayes Classifier 9](#_Toc60073952)

[5.1 Assumption 9](#_Toc60073953)

[5.2 Expression 10](#_Toc60073954)

[5.3 Naïve Bayes Classifier 10](#_Toc60073955)

[5.4 Example 10](#_Toc60073956)

[5.5 Laplacian Correction 12](#_Toc60073957)

[5.5.1 Example 13](#_Toc60073958)

[6. Bayes Classifier Extension 13](#_Toc60073959)

[6.1 Bayesian Decision Based on Minimal Risk 13](#_Toc60073960)

[6.1.1 Example 14](#_Toc60073961)

[6.2 Parameter Estimation 16](#_Toc60073962)

[6.2.1 Supervised Parameter Estimation 16](#_Toc60073963)

[6.2.2 Unsupervised Parameter Estimation 17](#_Toc60073964)

[6.2.3 Non-parametric Estimation: 17](#_Toc60073965)

[7. Semi-naïve Bayes Classifier 20](#_Toc60073966)

[7.1 ODE 20](#_Toc60073967)

[7.2 SPODE 20](#_Toc60073968)

[7.3 AODE 20](#_Toc60073969)

[7.3.1 Laplacian Correction 21](#_Toc60073970)

[7.3.2 Example 21](#_Toc60073971)

[8. Conclusions 22](#_Toc60073972)

[9. How to get the P(c|x)? 22](#_Toc60073973)

[10. Extension 23](#_Toc60073974)

[10.1 Bayesian network 23](#_Toc60073975)

[10.1.1 Definition 23](#_Toc60073976)

[10.1.2 Example 23](#_Toc60073977)

[10.1.3 Independence 23](#_Toc60073978)

[10.1.4 Structure 24](#_Toc60073979)

[10.2 Expectation-Maximization 25](#_Toc60073980)

[10.2.1 What is EM? 25](#_Toc60073981)

[10.2.2 Basic Idea 25](#_Toc60073982)

[11. Homework 25](#_Toc60073983)

[11.1 Question 26](#_Toc60073984)

[11.1.1 Naïve Bayesian Advantages and Disadvantages? 26](#_Toc60073985)

[11.1.2 Three Conditions of Naïve Bayesian? 28](#_Toc60073986)

[11.1.3 What is MLE? 149-150 28](#_Toc60073987)

[11.1.4 What is Naïve Bayes? 150-151 28](#_Toc60073988)

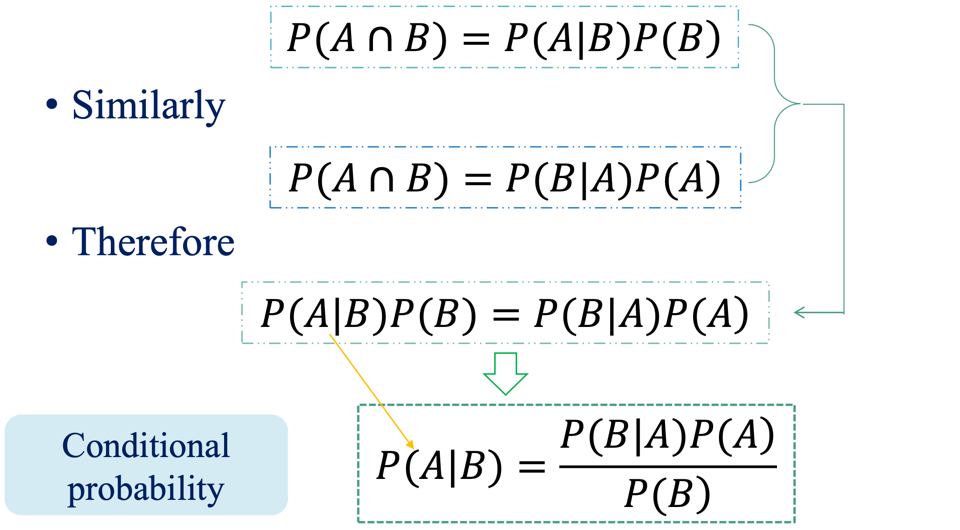
[11.1.5 What is EM? 163 29](#_Toc60073989)

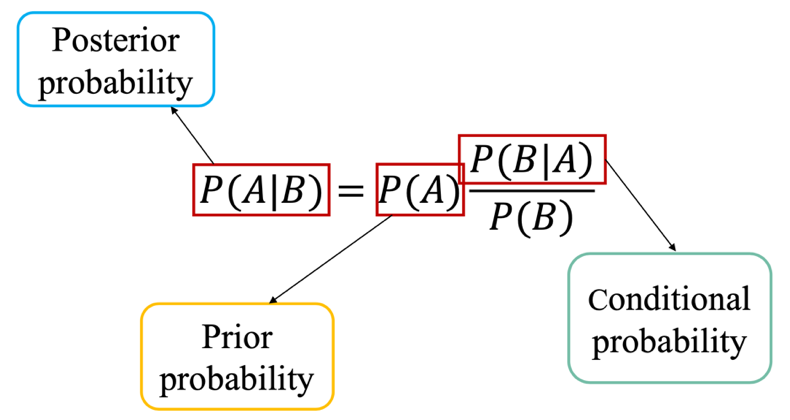
[11.2 水果糖问题 29](#_Toc60073990)

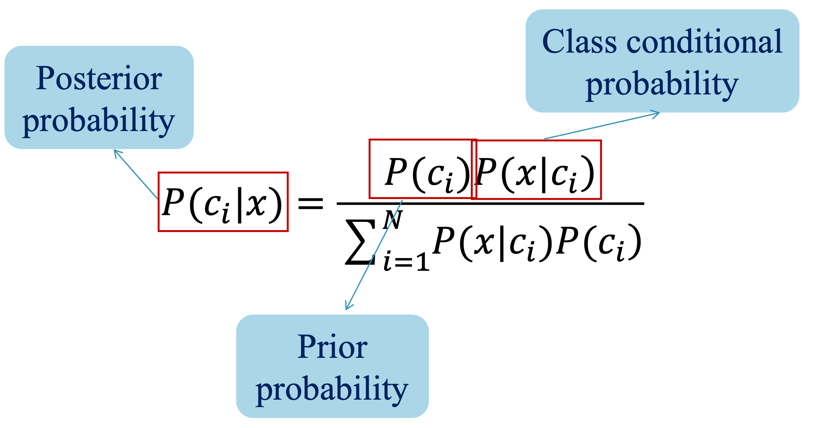
[11.3 假阳性问题 30](#_Toc60073991)

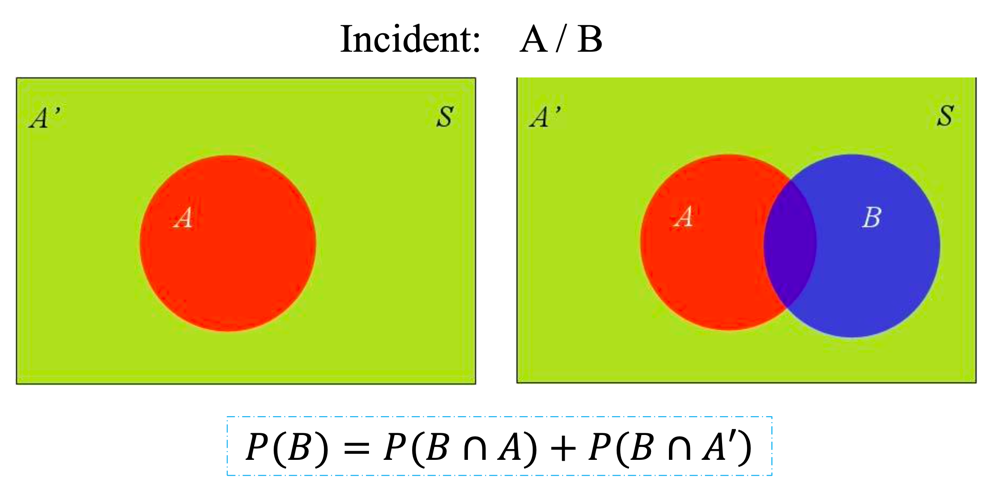
[11.4 射击问题 30](#_Toc60073992)

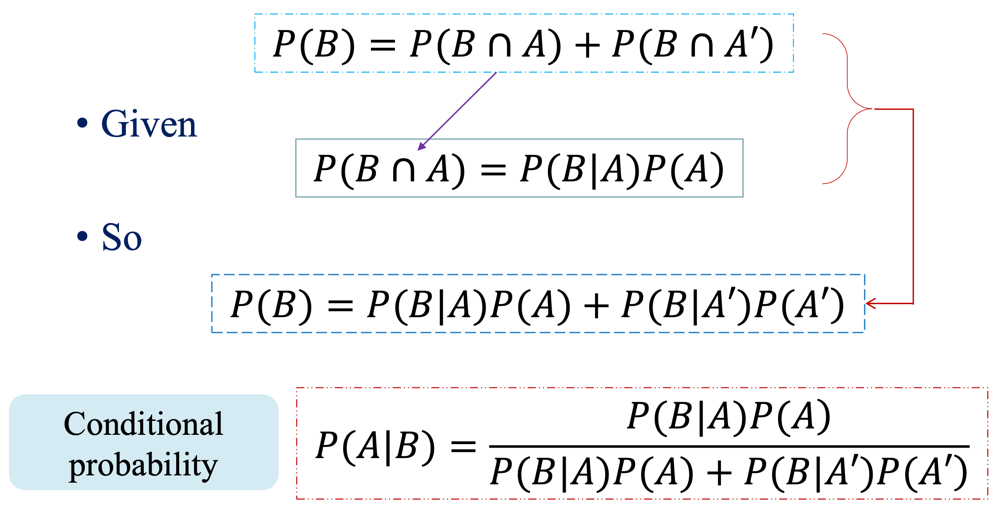
# Bayes Rule

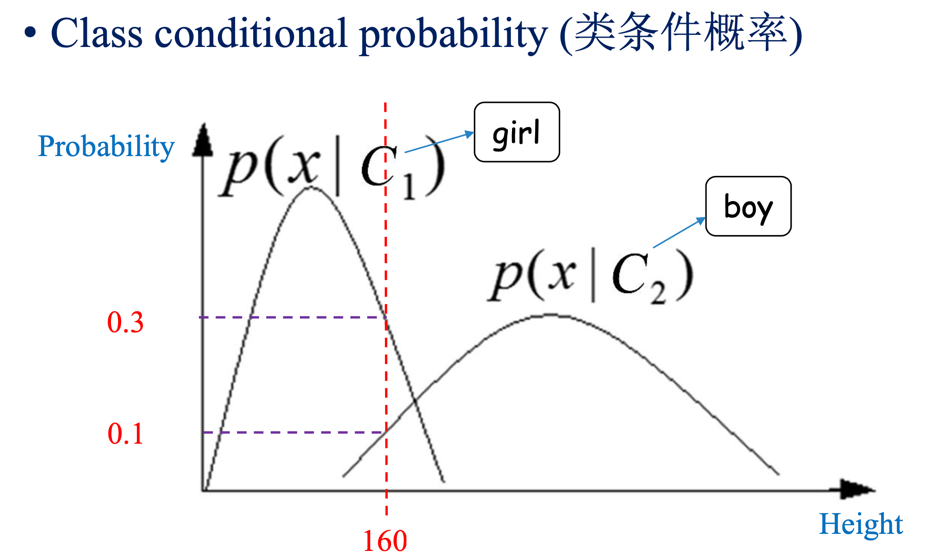












# Outline

* Bayesian Decision Based on Minimum Error Rate
* Maximum Likelihood Estimation
* Naïve Bayes Classifier
* Bayes Classifier Extension
* Semi-naïve Bayes Classifier
* Bayesian Application Examples

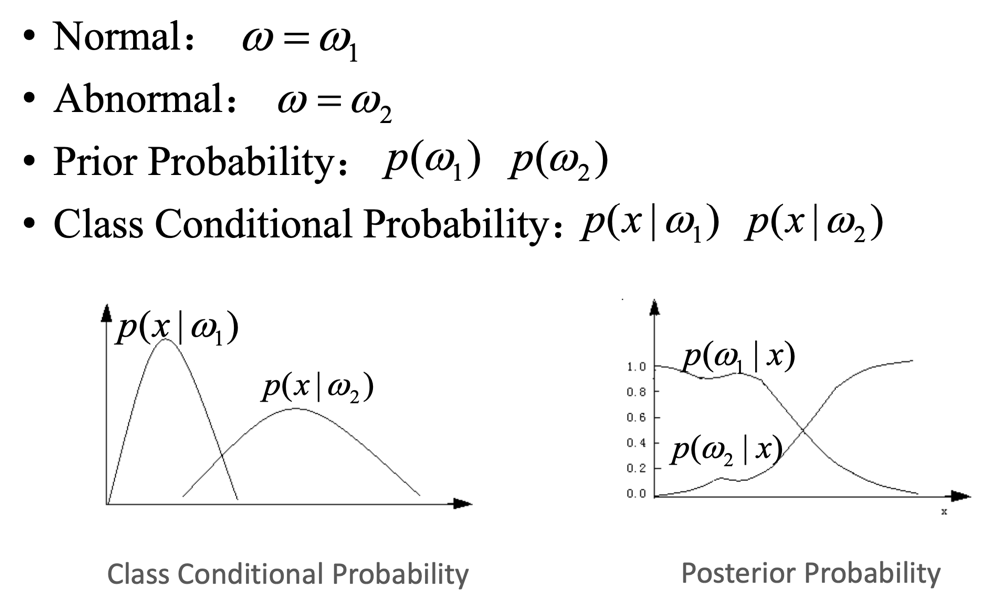
# Bayesian Decision Based on Minimum Error Rate

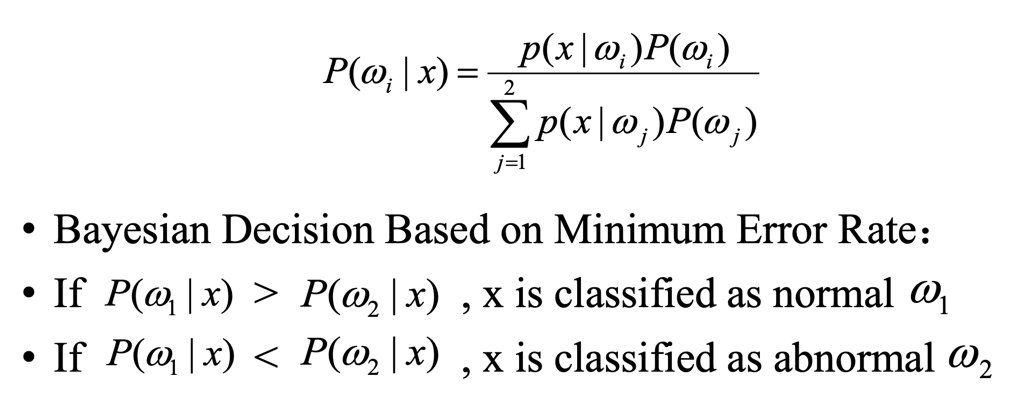
## What is it?

* In the problem of pattern classification, based on the Bayesian formula in probability theory to minimize the error of classification, the classification rule that minimizes the error rate can be obtained, which is called **Bayesian decision based on minimum error rate**.

## Example

* An *example* of cancer cell recognition illustrates the problem-solving process. Assuming that each cell to be identified has been preprocessed, d features representing the basic characteristics of the cell are extracted and become a vector x of the d-dimensional space. The purpose of the identification is to classify x as normal or abnormal cells.





* Assuming that in a local area, the prior probabilities of normal and abnormal in cell recognition are:

Normal: 𝑃(𝑐1) = 0.9

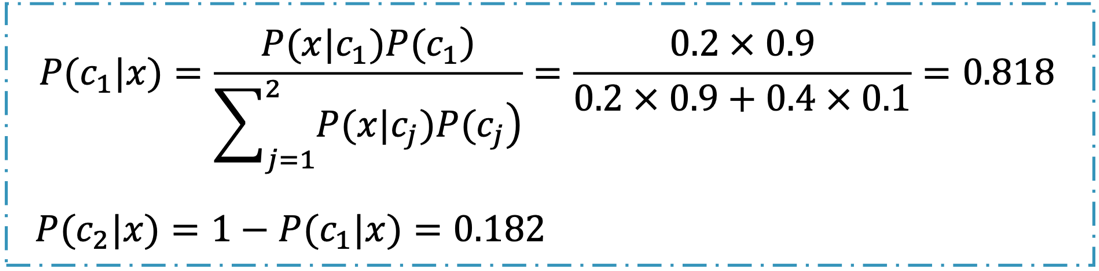
Abnormal: 𝑃(𝑐2) = 0.1

* + There is a cell to be identified, the observed value is x, from the class condition probability density distribution curve

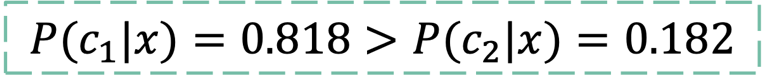


* + Try to judge whether the cell is normal or abnormal?
  + The posterior probability of 𝑐1 and 𝑐2 is calculated by

Bayesian formula :



* According to Bayesian decision rules



* Decision rules: Normal
* From this example, it can be seen that the decision outcome depends on both the **observed conditional probability density** **and the** **prior probability**. In this example, because the prior probability of state 1 is several times greater than the prior probability of state 2, **the prior probability** plays a dominant role in making decisions.

## Theoretical Model

* Prior probability:

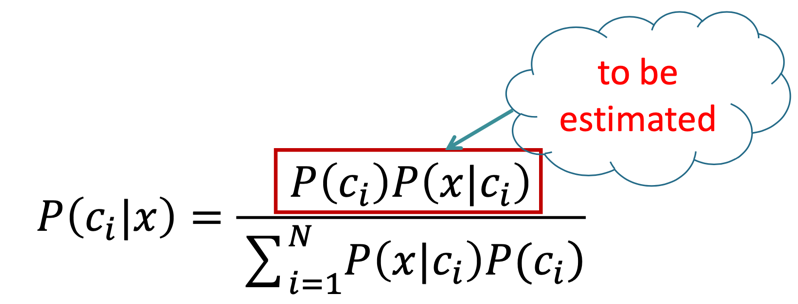
page18image53792320page18image53795584

* Class conditional probability:

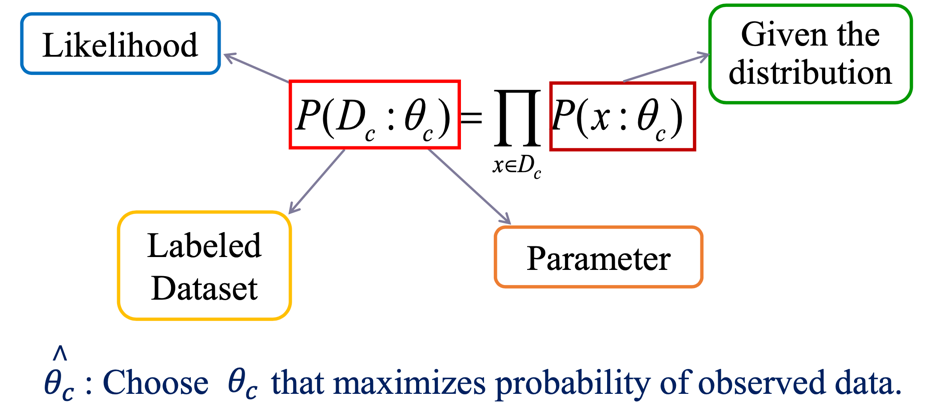
 --In fact, they are unknown.

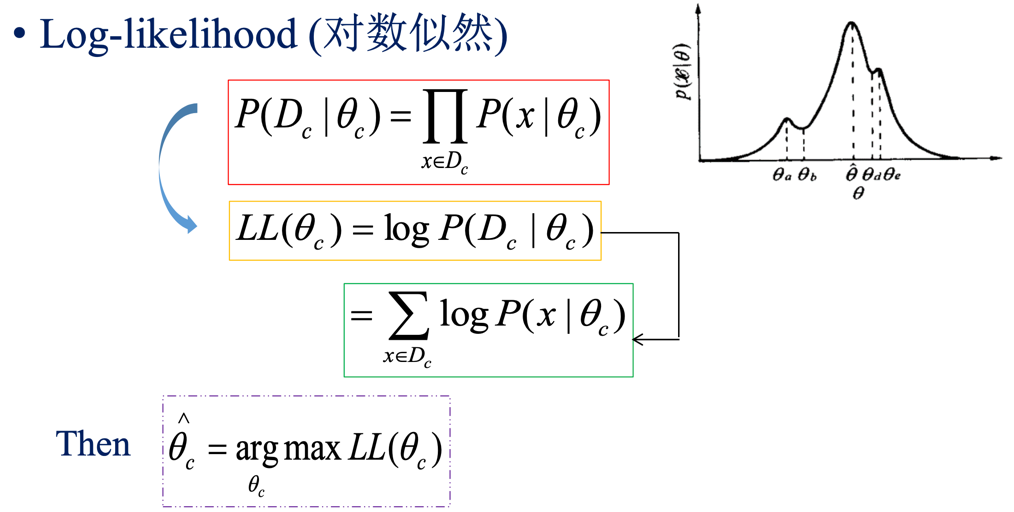
* Bayesian model is a **theoretical model**. It is difficult to obtain prior probability and class conditional probability realistically, so their values need to be estimated.

# Maximum Likelihood Estimation

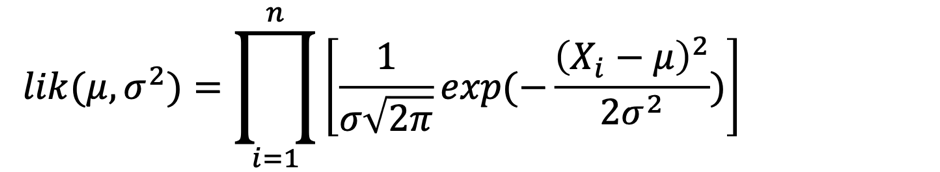


* MLE (Maximum Likelihood Estimation): A general method for estimating parameters in a model.

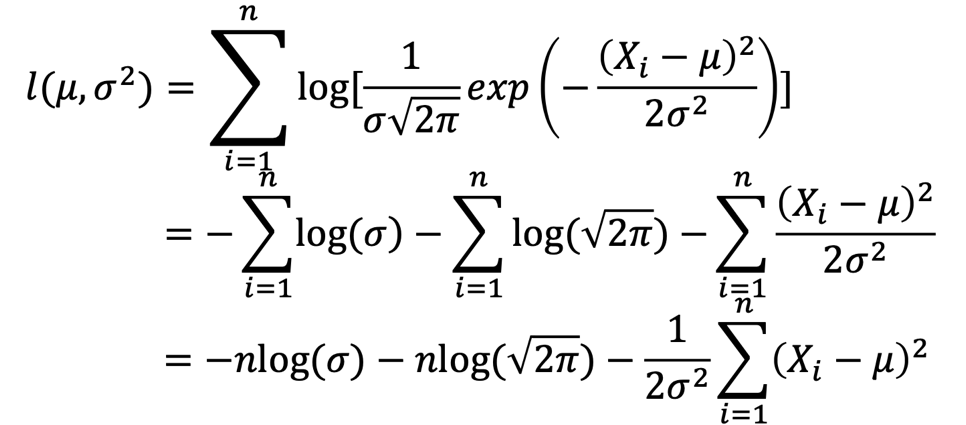




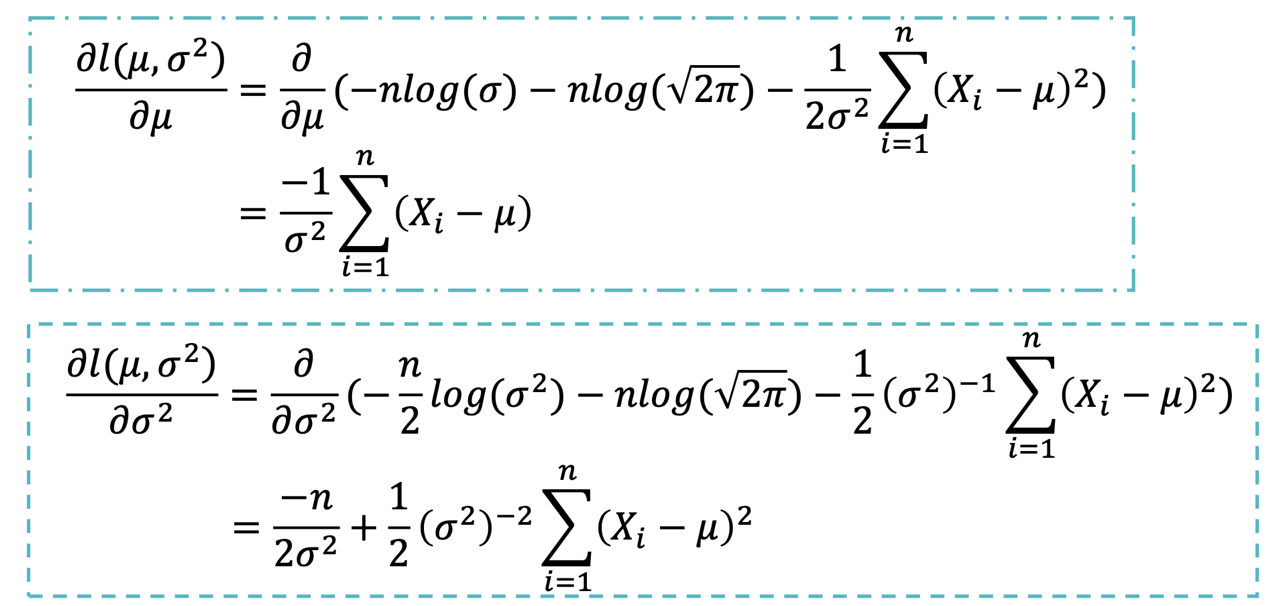
* Suppose 𝑋i ~ 𝑁(𝜇, 𝜎2) and i.i.d.
* What is the likelihood function?



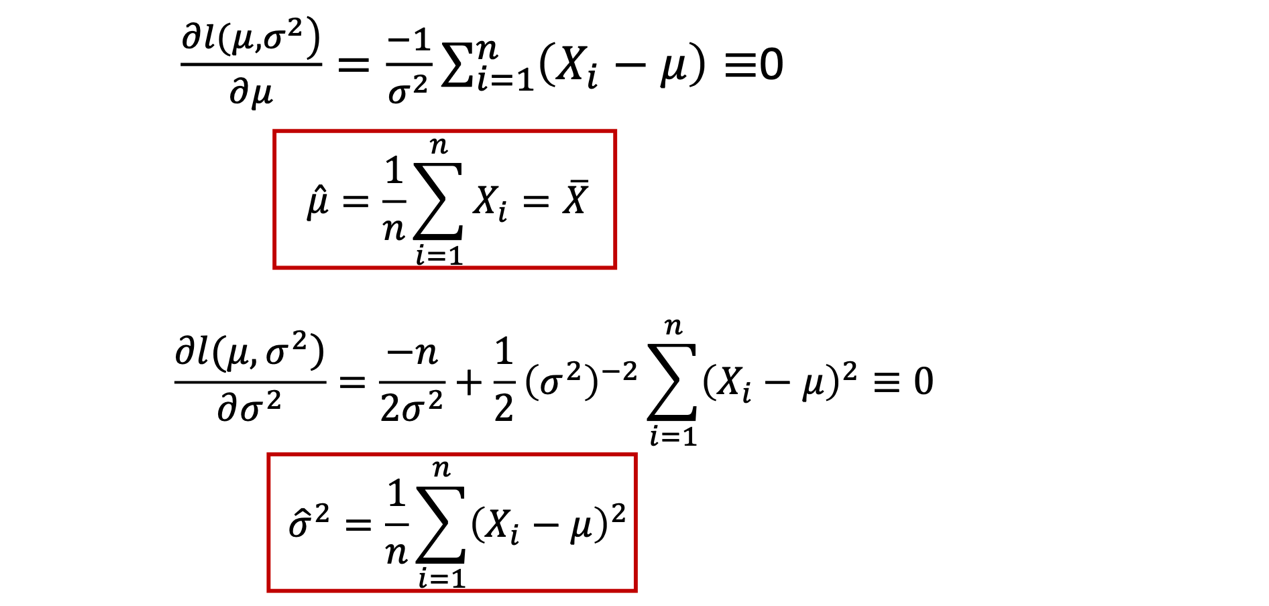
* What is the log-likelihood function?



* Now there are two unknown parameters so we will need to find the separate partial derivatives:



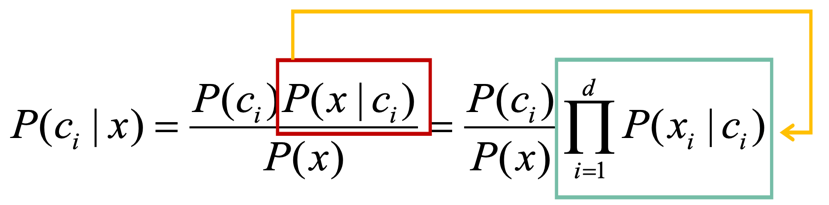
* Set the separate partial derivatives to zero and solve for the specific parameter:



# Naïve Bayes Classifier

## Assumption

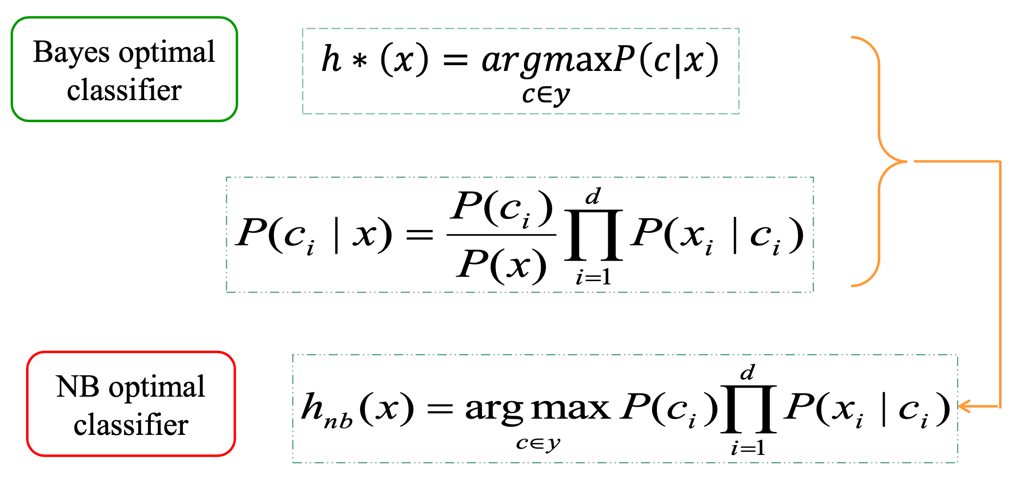
* A simplified assumption: **attributes are conditionally independent**:



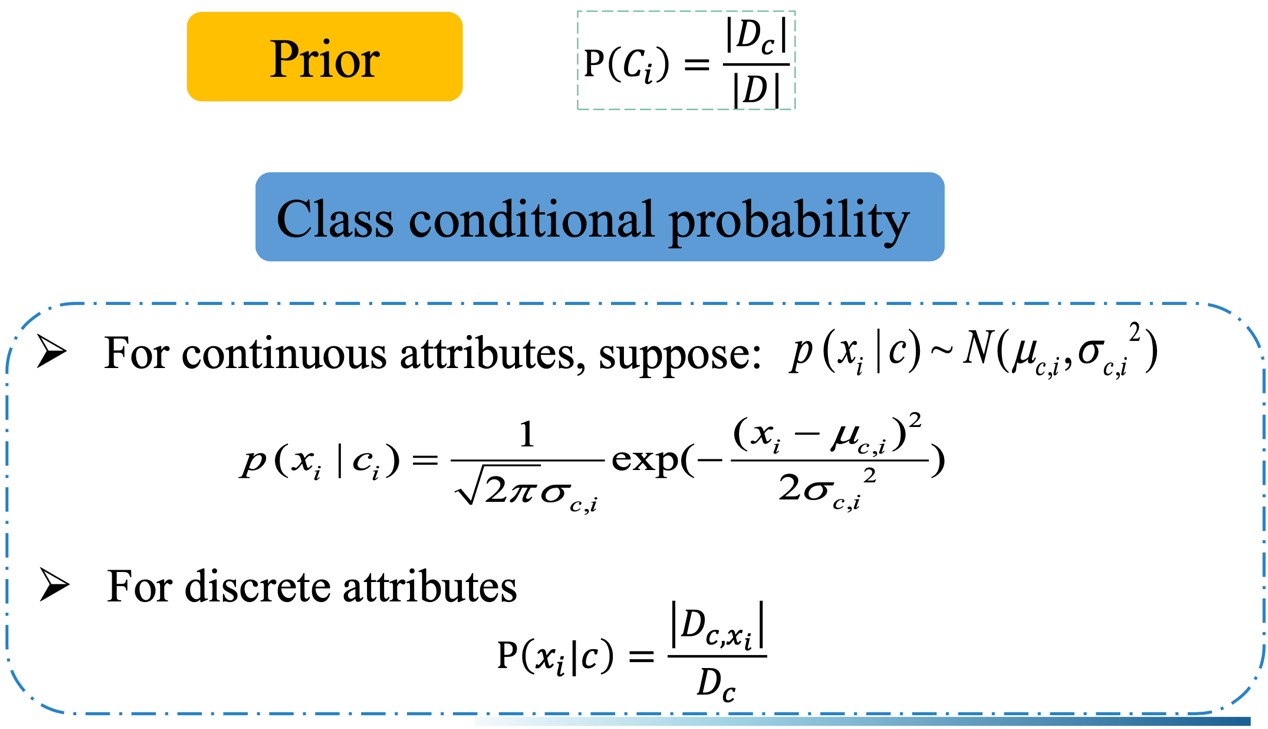
𝑑 The dimension of attributes

𝑥𝑖 The value of 𝑥 on the attribute of 𝑖

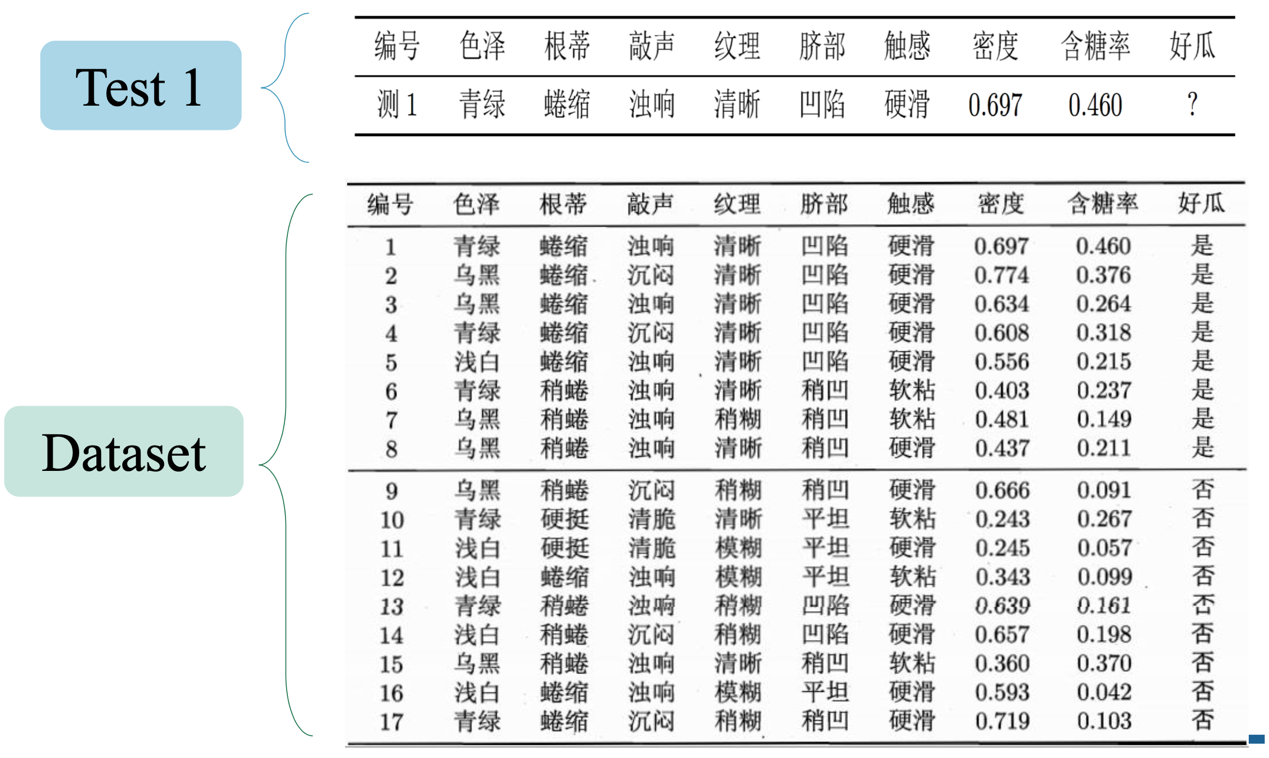
## Expression

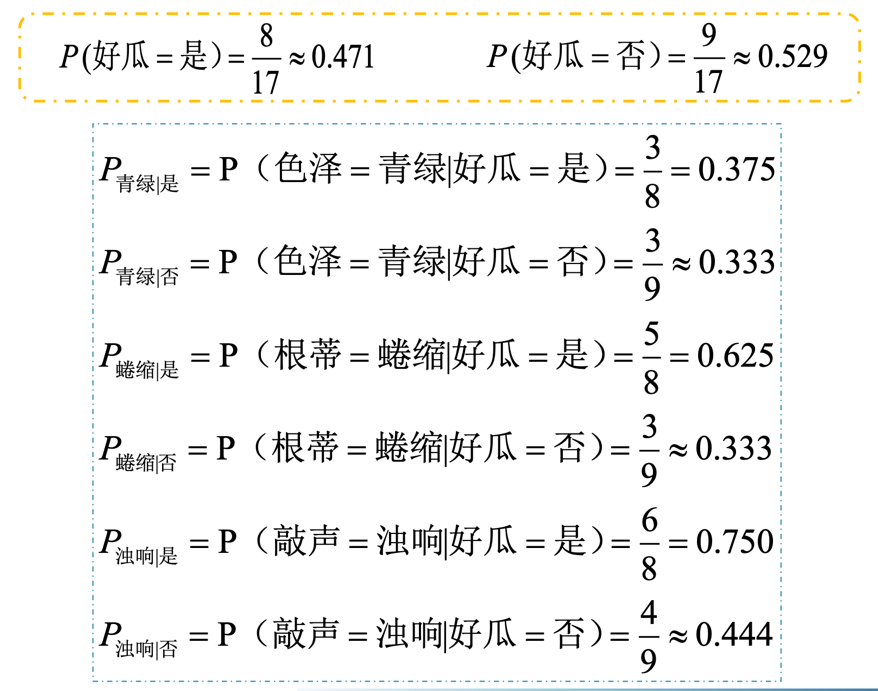


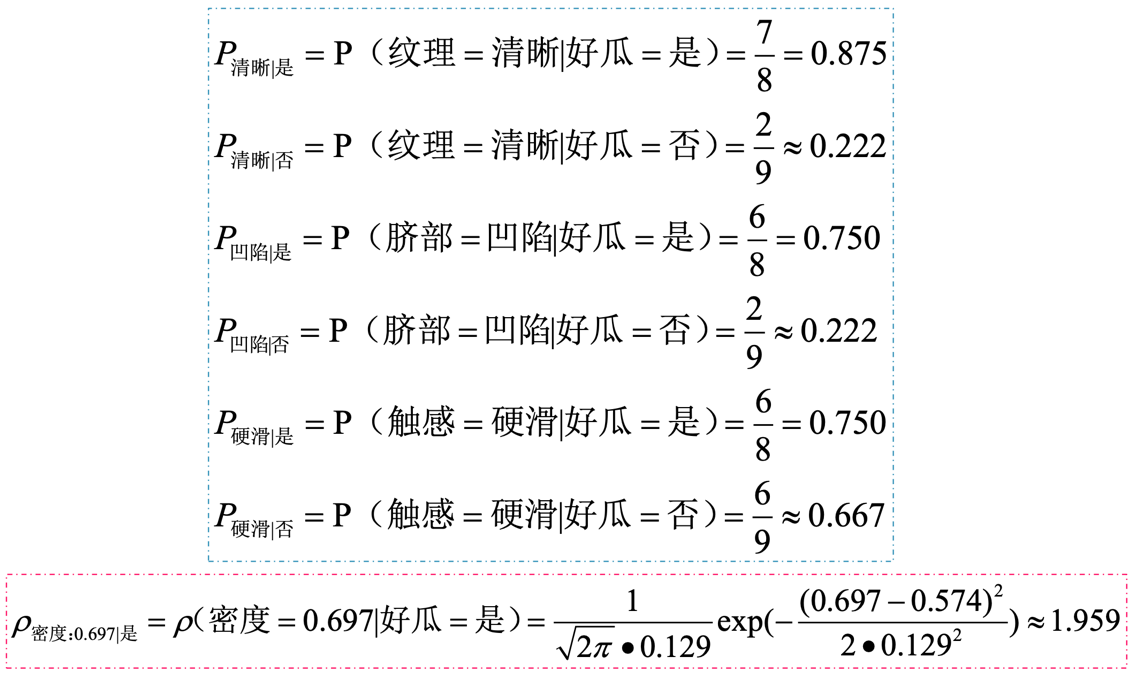
## Naïve Bayes Classifier

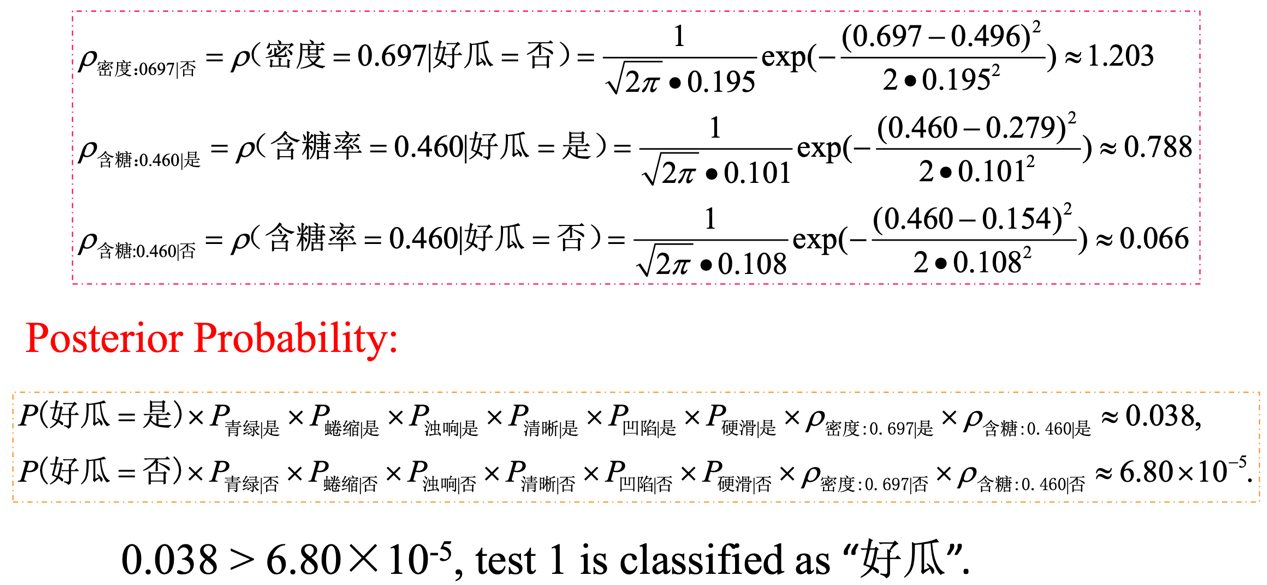


## Example





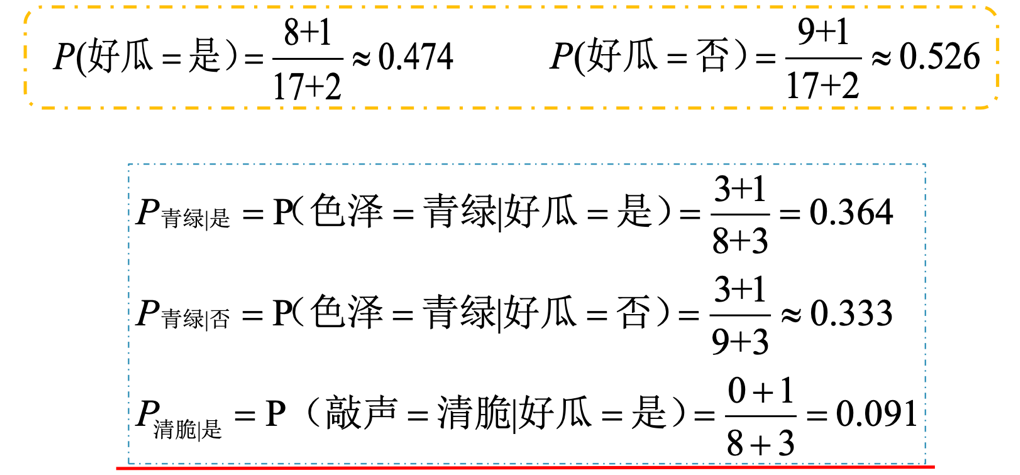




## Laplacian Correction

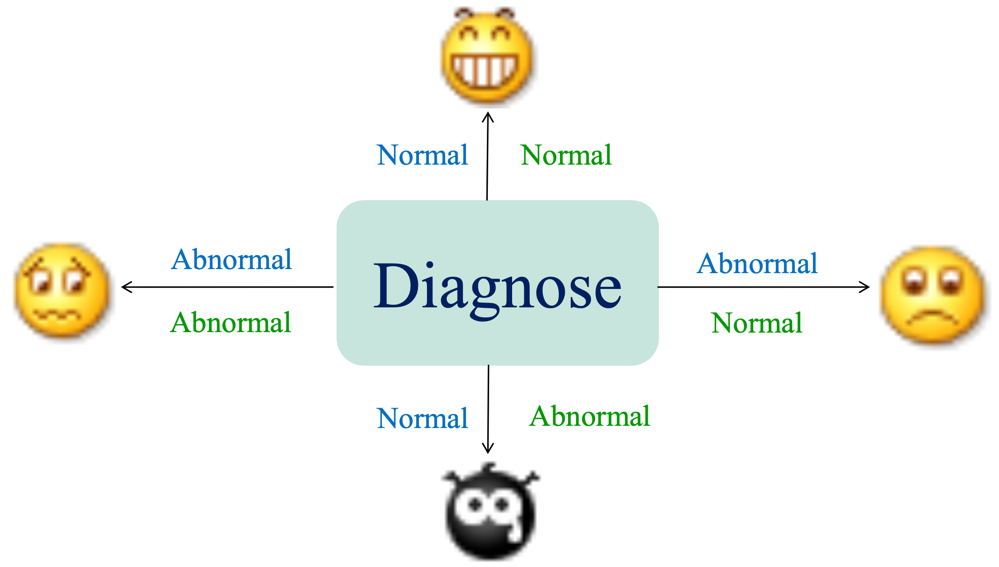


### Example

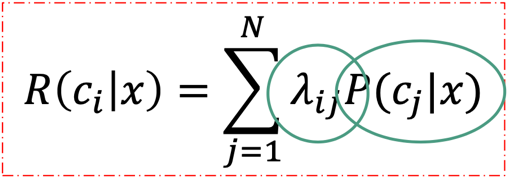


# Bayes Classifier Extension

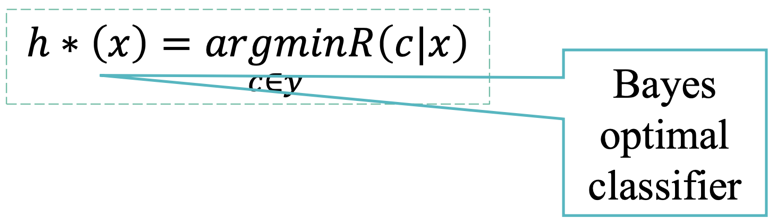
## Bayesian Decision Based on Minimal Risk



* 𝑦 = {𝑐1,𝑐2,...,𝑐N}: the finite set of N states of labels
* 𝜆ij: the loss incurred by mistaking 𝑐i for 𝑐j
* Given 𝑥 = [𝑥1,𝑥2,⋯𝑥𝑑]𝑇
* Conditional risk:



* Find a **decision rule** h: 𝑋 ↦ 𝑌 to minimize the overall risk.
* Choose the label which can minimize the conditional risk for each



### Example

#### Repeat

* Assuming that in a local area, the prior probabilities of normal and abnormal in cell recognition are:

Normal: 𝑃(𝑐1) = 0.9

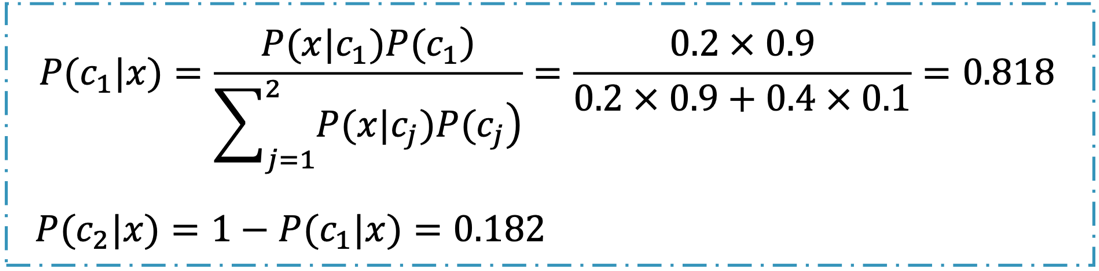
Abnormal: 𝑃(𝑐2) = 0.1

* There is a cell to be identified, the observed value is x, from the class condition probability density distribution curve

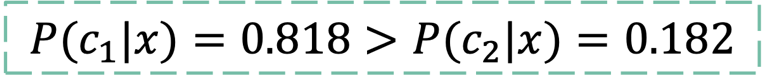


* Try to judge whether the cell is normal or abnormal?
  + The posterior probability of 𝑐1 and 𝑐2 is calculated by

Bayesian formula:

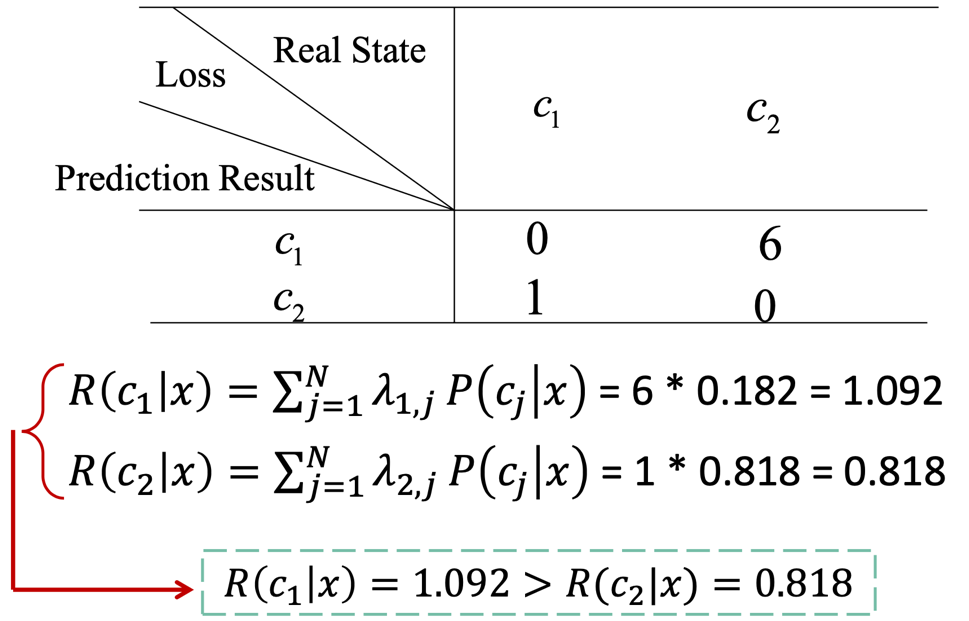


* According to **Bayesian decision rules**



* Decision rules: Normal

#### Based on Minimal Risk

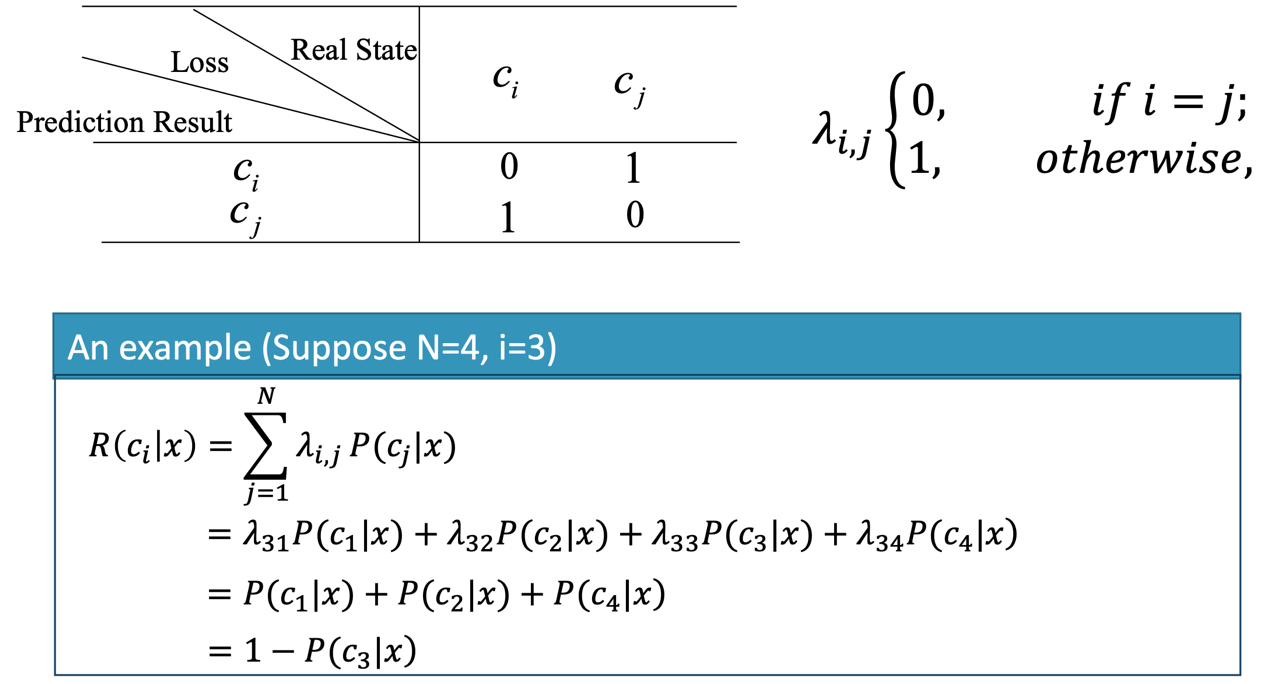


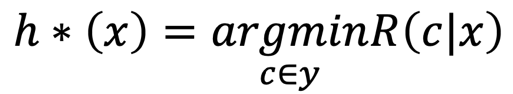
* Decision rules: Abnormal

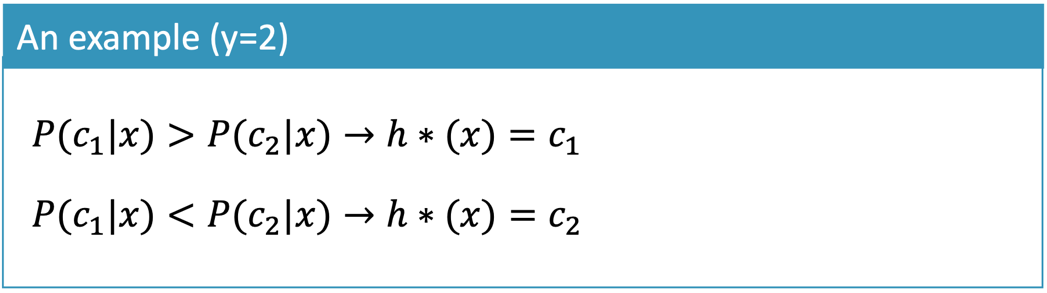
#### Comparison

* The result of the classification is just the opposite. This is because there is one more factor affecting the decision-making result, namely "loss". And the losses caused by the two types of wrong decisions are very different, so "loss" has played a leading role.

#### Minimal Classification Error Rate

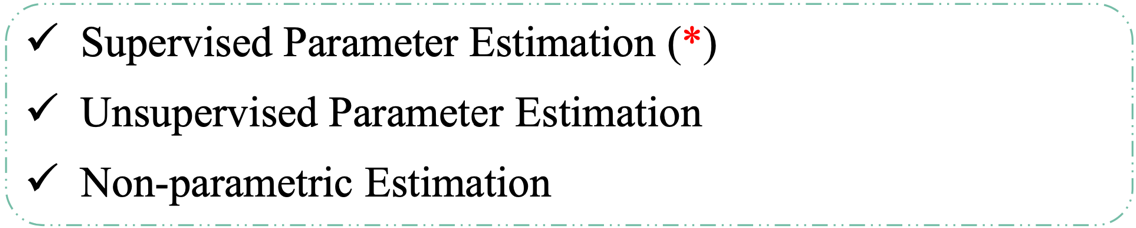


* Conditional risk: 𝑅(𝑐|𝑥) = 1 − 𝑃(𝑐|𝑥)
* Recall: 
* Bayes optimal classifier : 



## Parameter Estimation

The method of estimating the overall probability distribution from the sample set can be summarized as follows:



### Supervised Parameter Estimation

The **categories and conditions** to which the sample belongs **are known** in the form of the overall probability density function, and some of the **parameters** that characterize the **probability density function are unknown**.

**For example**, only the overall **distribution** of the sample is **known**, and the **parameters** of the normal distribution are **unknown**. Our goal is to statistically judge some of the population distribution from a set of samples of a known category. The estimate in this case is called the parameter estimation under supervision.

### Unsupervised Parameter Estimation

The overall **probability density function is known**, but the **category** to which the sample belongs is **unknown** and it is required to **determine some parameters** of the probability density function.

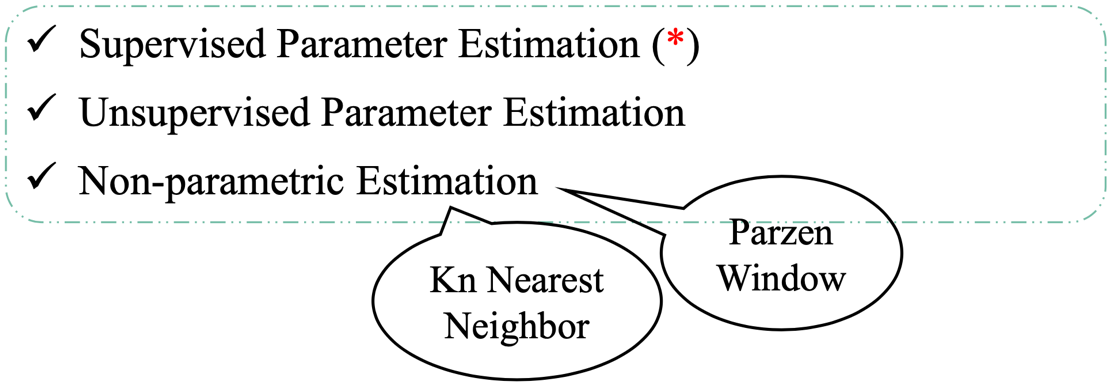
Supervised and unsupervised means whether the category to which the sample belongs is known or unknown. There are two commonly used methods, one is **the maximum likelihood estimation**, and the other is **Bayesian estimation**.

* **The maximum likelihood estimation** is that the **parameters** are regarded as definite and **unknown**, and the best **estimate** is obtained under the condition that the **probability** of obtaining the actual observed sample is **maximum**.
* **The Bayesian estimation** treats the unknown parameters as a random variable with a certain distribution. The observed result of the sample transforms the prior distribution into a posterior distribution, and then corrects the original estimate of the parameter based on the posterior distribution.

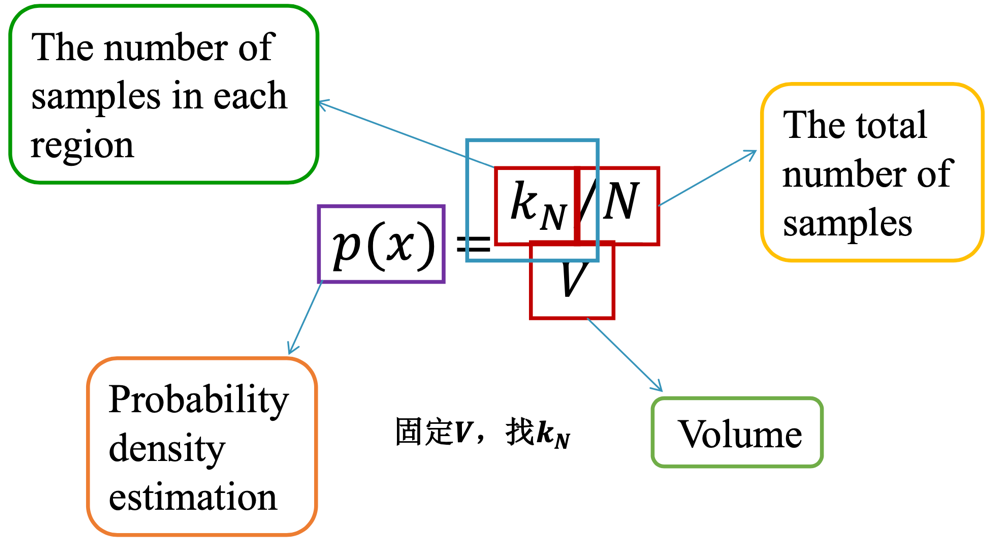
### Non-parametric Estimation:

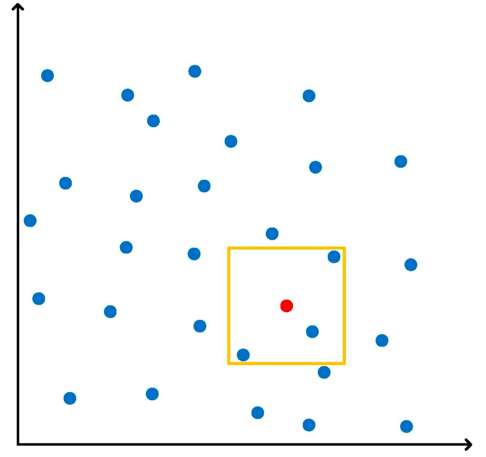
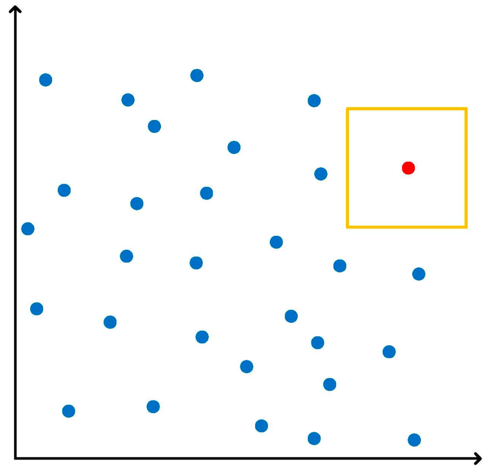
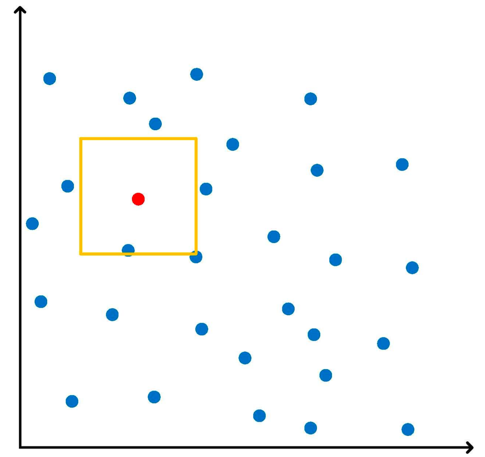
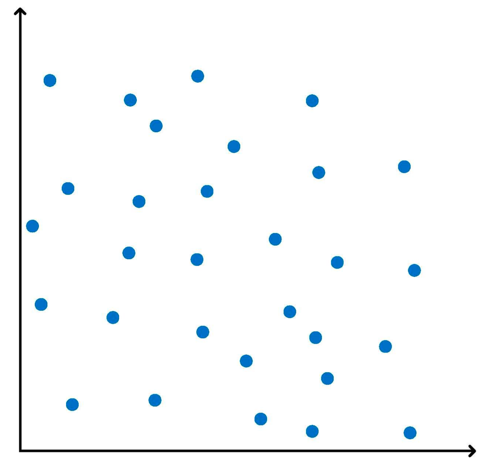
The **category** to which the sample belongs is **known**, but **the form of probability density function is unknown**, so it requires us to directly infer the probability density function itself.

The method of estimating the overall probability distribution from the sample set can be summarized as follows:



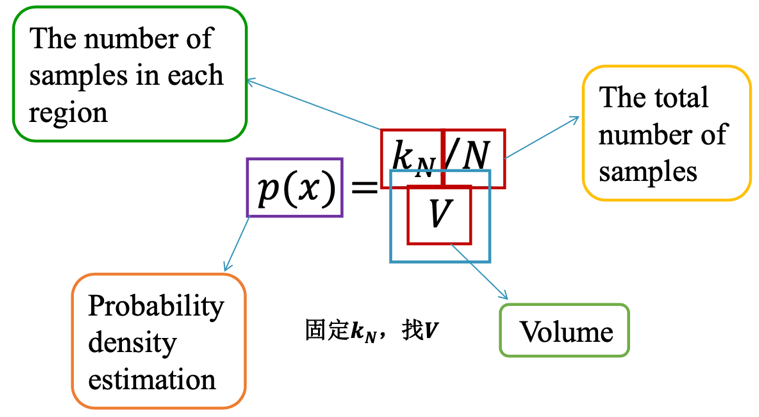
#### Parzen Window

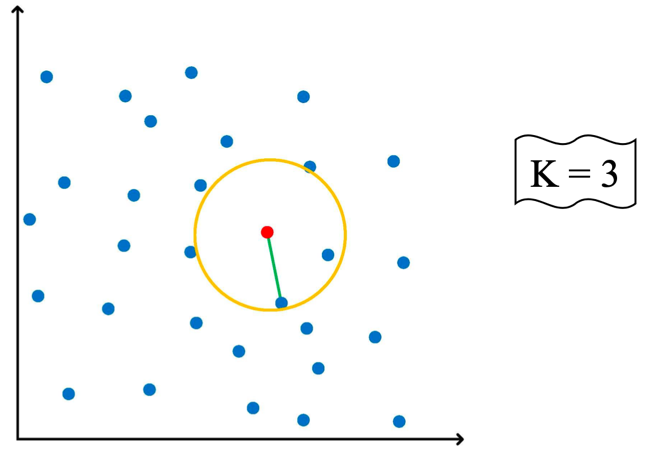


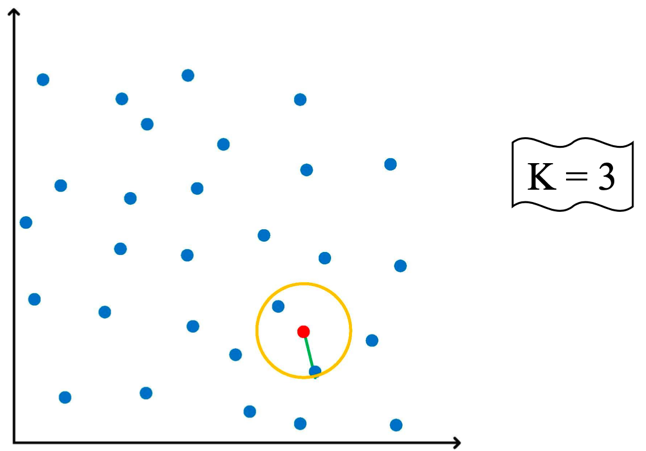
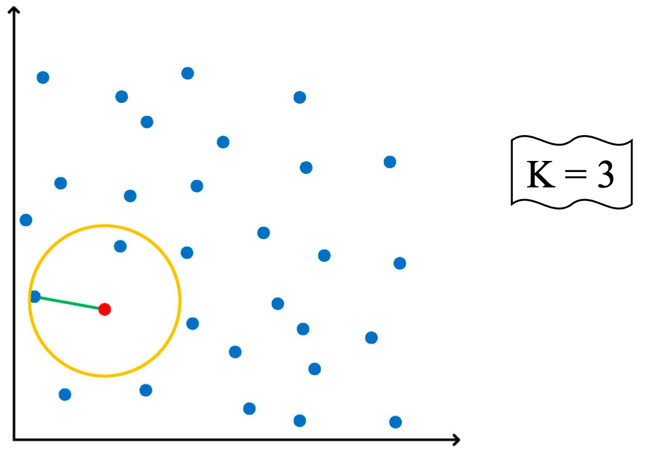


* Two-dimensional plane: Square
* Three-dimensional plane: Cube
* N-dimensional plane: Hypercube

#### Kn Nearest Neighbor



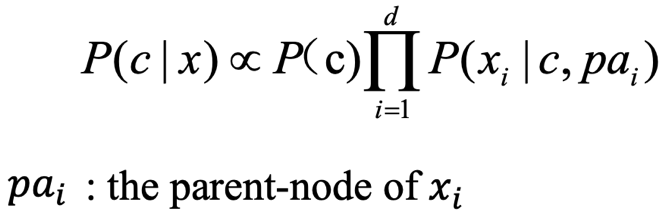




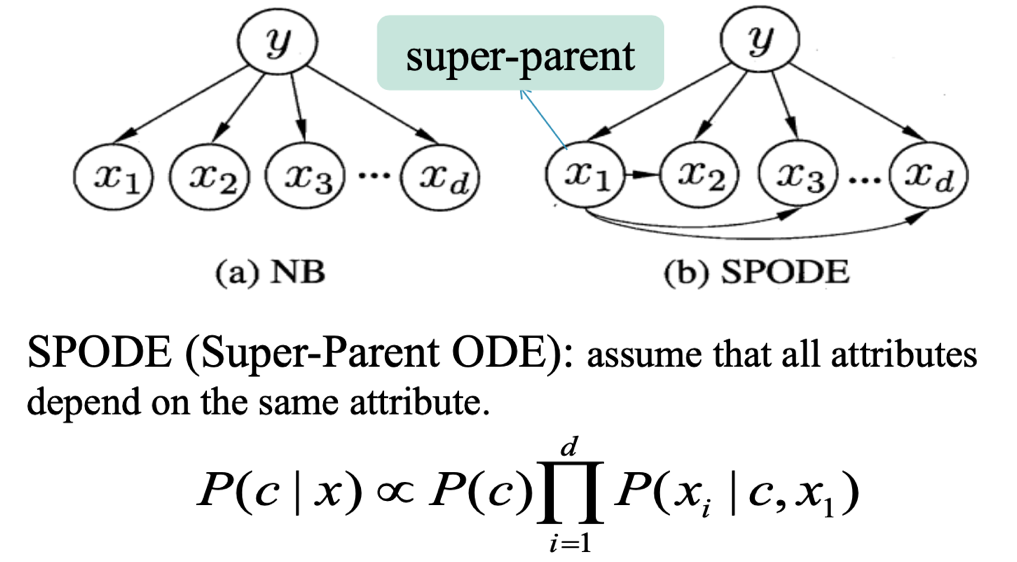
# Semi-naïve Bayes Classifier

## ODE

* In practical scenarios, the attribute independence assumption is often **violated**!
* ODE (**One-Dependent** Estimator): Suppose an attribute only relies on a most other properties.

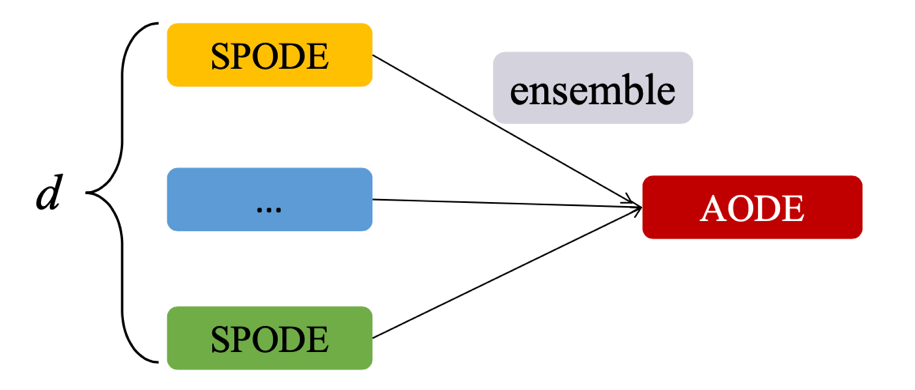


## SPODE



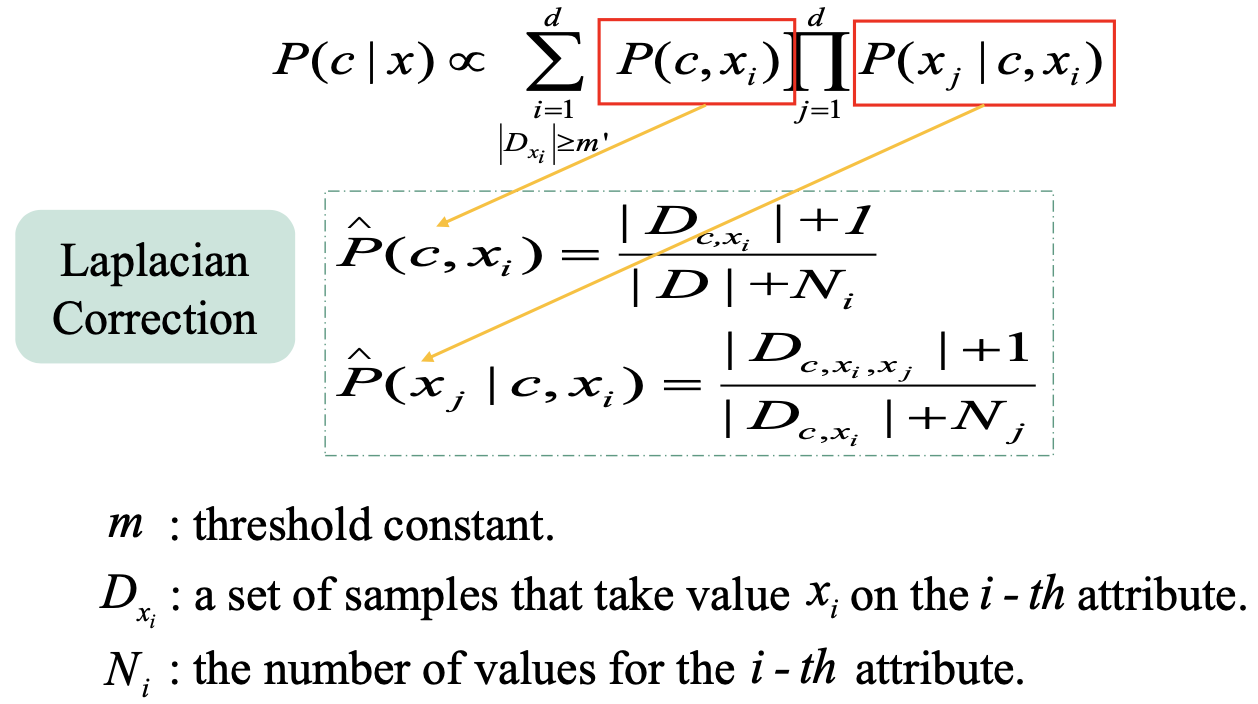
## AODE

* AODE: Averaged One-Dependent Estimator

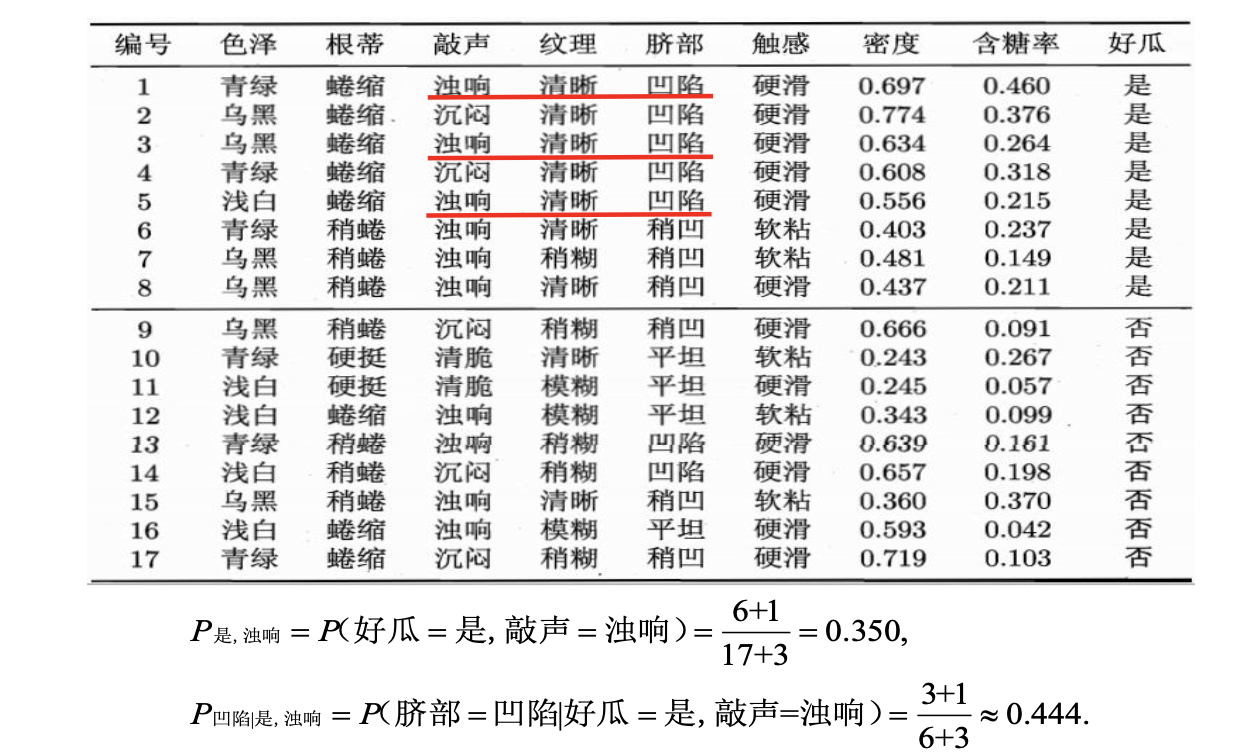


SPODE which has enough statistical data to support.

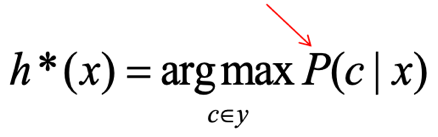
### Laplacian Correction



### Example



# Conclusions

* 
* MLE (parameter estimate)
* Naïve Bayes Classifier (attribute conditional independence assumption)
* Bayes Extension
* Semi-naïve Bayes Classifiers (ODE)

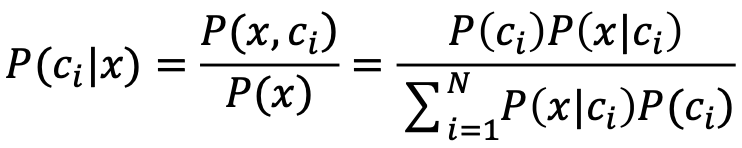
# How to get the P(c|x)?

* Discriminative models (判别式模型)

eg: Support Vector Machines, Decision Tree, BP Neural Network

* Generative models (生成式模型)

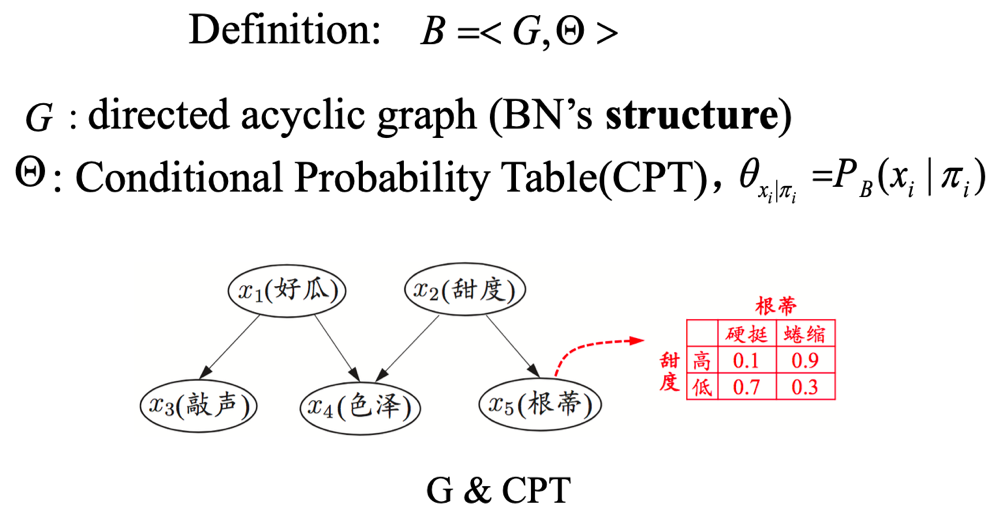
eg: Naïve Bayes, AODE, Restricted Boltzmann Machine



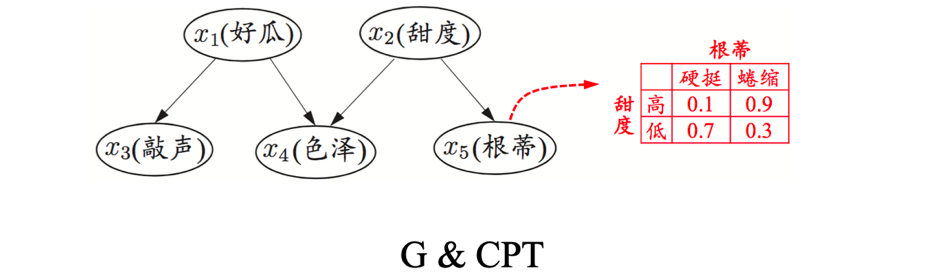
# Extension

## Bayesian network

### Definition



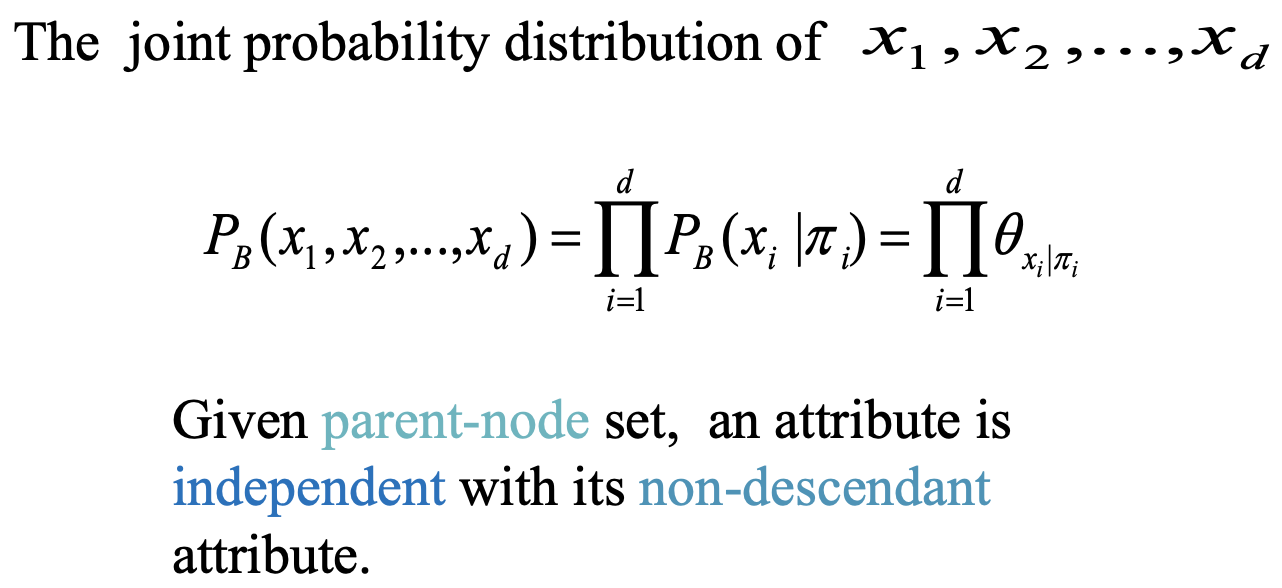
### Example



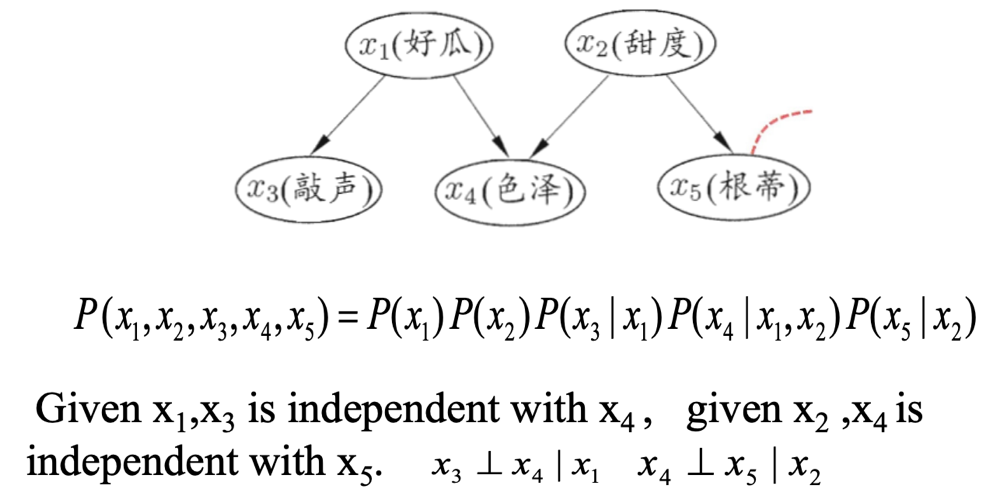
From G -> “色泽”直接依赖于“好瓜”和“甜度”

From CPT -> “根蒂”对“甜度”的量化依赖关系 P(根蒂=硬挺|甜度=高)=0.1

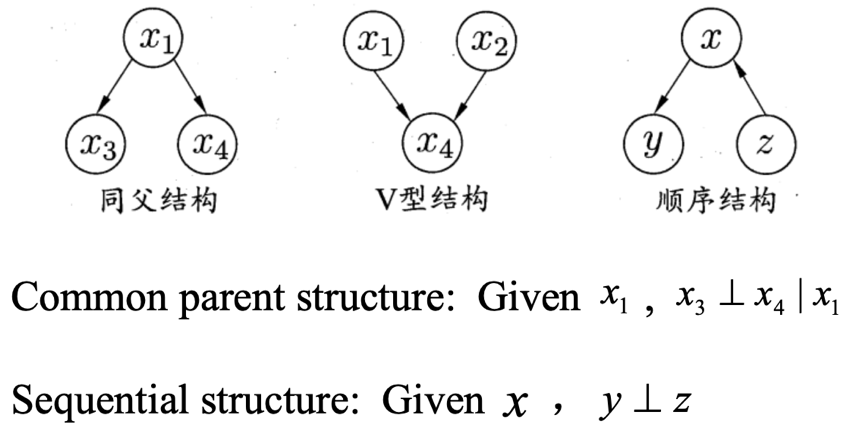
### Independence

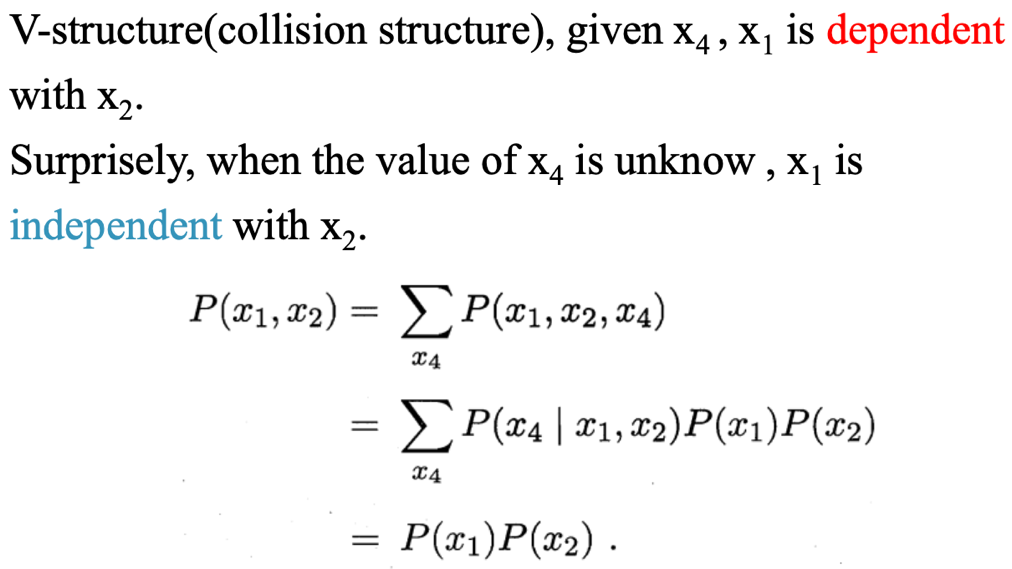


#### Example



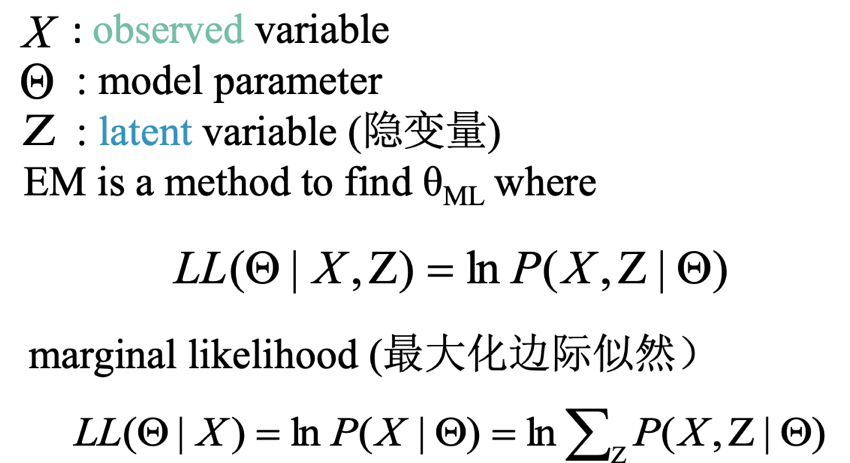
### Structure



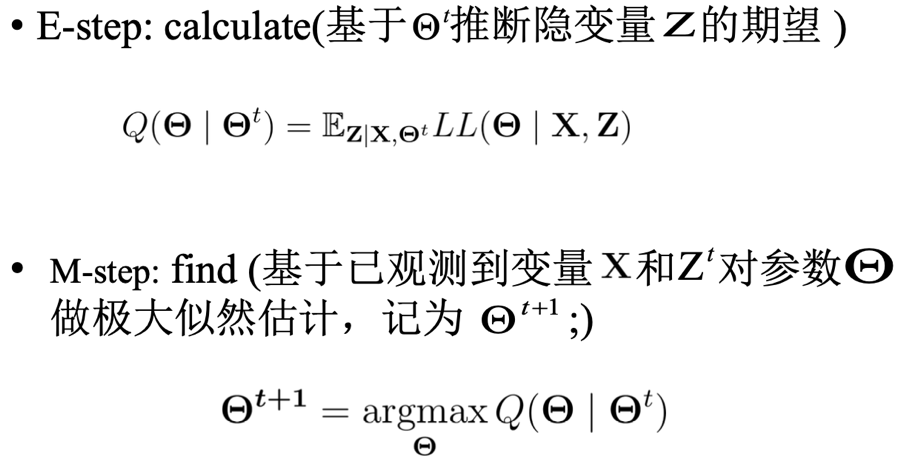


## Expectation-Maximization

### What is EM?



### Basic Idea



# Homework

## Question

### Naïve Bayesian Advantages and Disadvantages?

**Answer1:**

**Advantage:**

1. If its assumption of the independence of features holds true, it can perform better than other models.
2. Naive Bayes requires a small amount of training data to estimate the test data. So, the training period is less.
3. Easy to implement as only the probability is to be calculated.
4. It works well with high dimensions such as text classification.
5. Naive Bayes is better suited for categorical input variables than numerical variables.
6. Naive Bayes is suitable for solving multi-class prediction problems.

**Disadvantages:**

1. If the independent assumption does not hold then performance is very low.
2. ‘zero-frequency problem’. Smoothing turns out to be a over-head and a must to do step when probability of a feature turns out to be zero in a class.
3. Vanishing value is also a problem due to product of many small probability(eg. 0.05³).

### Three Conditions of Naïve Bayesian?

**属性条件独立性假设**：对已知类别，假设所有属性相互独立。

**？？？**

**？？？**

### What is MLE? 🍉149-150

🍉149-150:

极大似然估计是根据数采样来估计概率分布参数的经典方法。

…

(7.11)

### What is Naïve Bayes? 🍉150-151

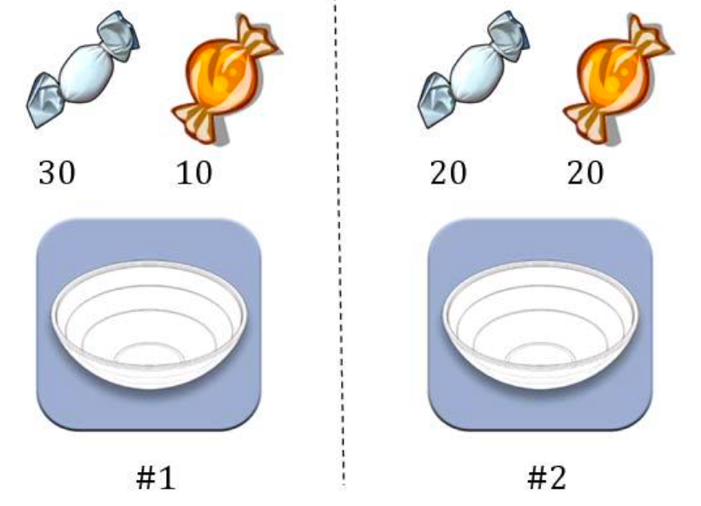
### What is EM? 🍉163

最大期望（EM）算法是在概率模型中寻找参数最大似然估计或者最大后验估计的算法，其中概率模型依赖于无法观测的隐变量。

🍉163：它是一种迭代式的方法，其基本思想是：…极大似然估计（M步）。

## 水果糖问题

两个一模一样的碗，一号碗有30颗水果糖和10颗巧克力糖，二号碗有水果糖和巧克力糖各20颗。现在随机选择一个碗，从中摸出一颗糖，发现是水果糖。请问这颗水果糖来自一号碗的概率有多大?



## 假阳性问题

已知某种疾病的发病率是0.001，即1000人中会有1个人得病。现有一种试剂可以检验患者是否得病，它的准确率是0.99，即在患者确实得病的情况下，它有99%的可能呈现阳性。它的误报率是5%，即在患者没有得病的情况下，它有5%的可能呈现阳性。现有一个病人的检验结果为阳性，请问他确实得病的可能性有多大?

## 射击问题

8支步枪中有5支已校准过，3支未校准。一名射手用校准过的枪射击，中靶概率为0.8；用未校准的抢射击，中靶概率为0.3；现从8支抢中随机取一支射击，结果中靶。求该枪是已校准过的概率。