

# K-Means算法

聚类算法

Supervised  $\Rightarrow (x, y)$   
特征  $\downarrow$  label

# 聚类分析

$(x, *)$   $x \rightarrow y$

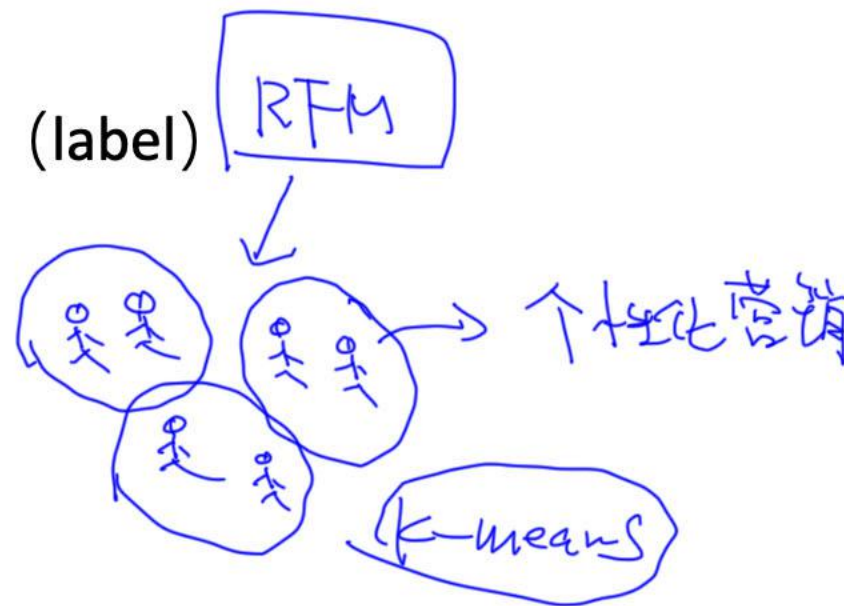
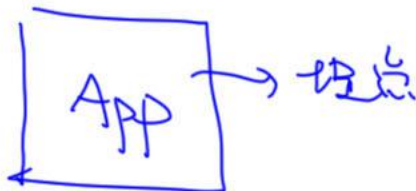
**无监督学习**：需要数据，但没有数据标签 (label)

发现数据中的规律 (模式)

- 用户分群
- 行为聚类
- ....

CRM  $\rightarrow$

ID	city	salary	...



Euclidean Distance

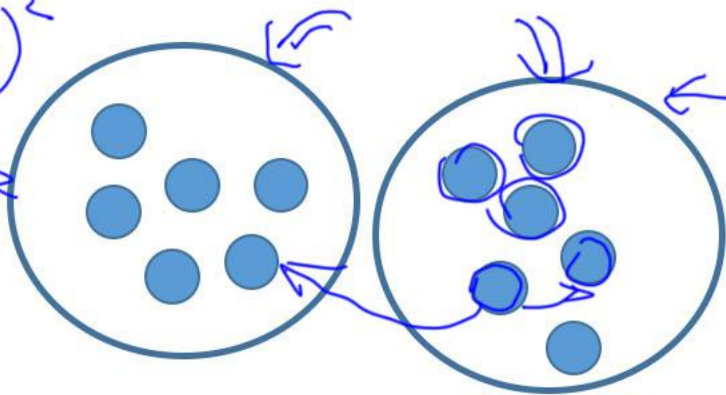
$$x = (x_1, x_2, \dots, x_d)$$

$y = (y_1, y_2, \dots, y_d)$  简单来讲, 把相似的物体聚在一起

# 聚类分析 $\Rightarrow$

$$d_E(x, y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_d - y_d)^2}$$

$$d_E^2(x, y) = (x_1 - y_1)^2 + \dots + (x_d - y_d)^2$$



K-means

Data

如何评估相似性?

KNN

距离:

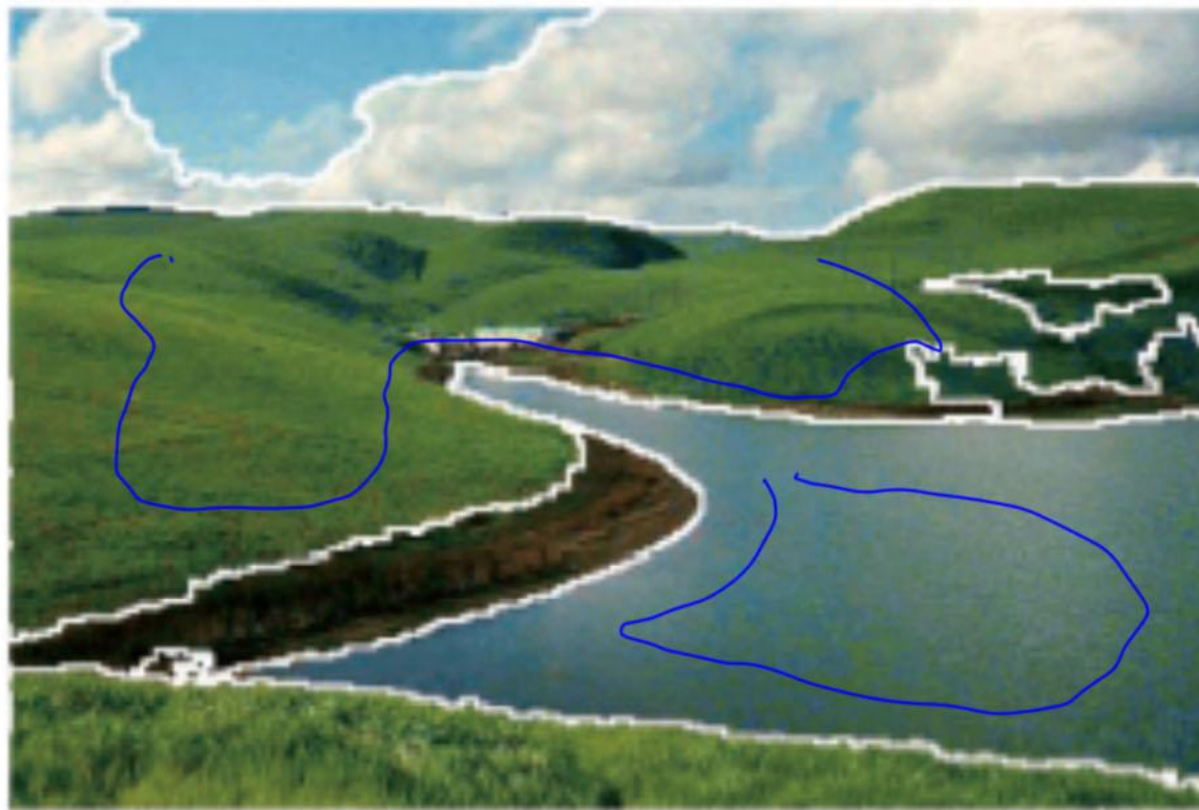
$\hat{x}$

$y$

$$d(x, y) = \|x - y\|_2^2 \downarrow$$

# 应用场景

图像分割：把图像分成相似的区域 *Image Segmentation*



*Vector Quantization*

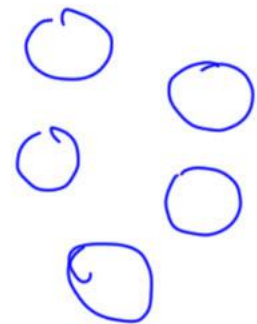


# 应用场景 (非常经典)

用户分层：把拥有类似兴趣的用户聚在一起



# K-means算法 $\Rightarrow$ Data $\Rightarrow$



循环迭代式的算法

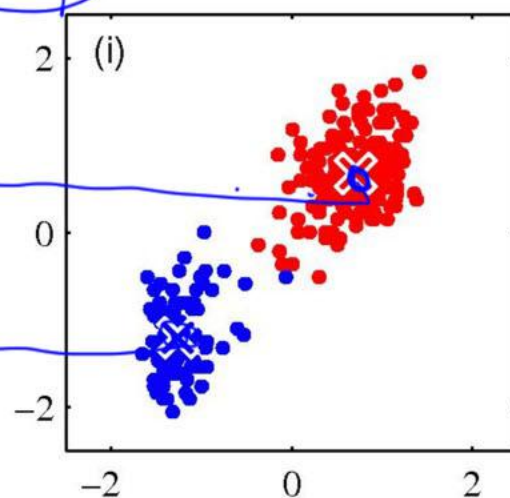
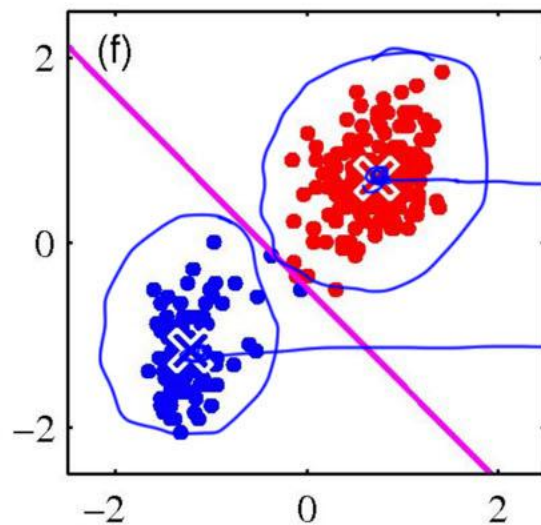
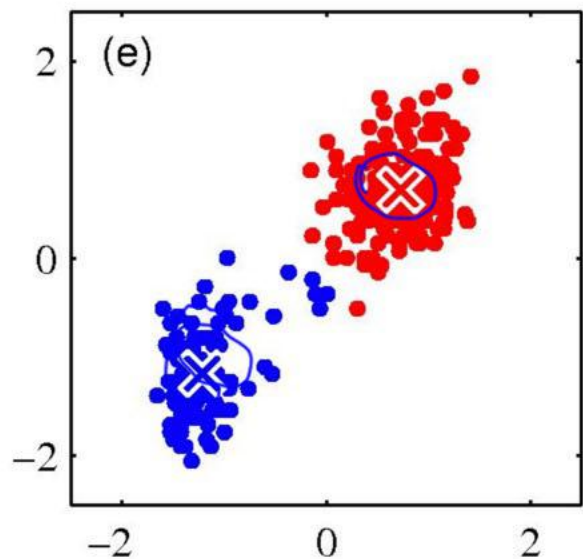
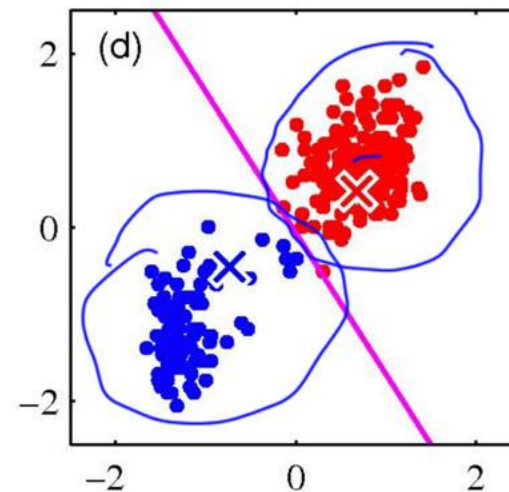
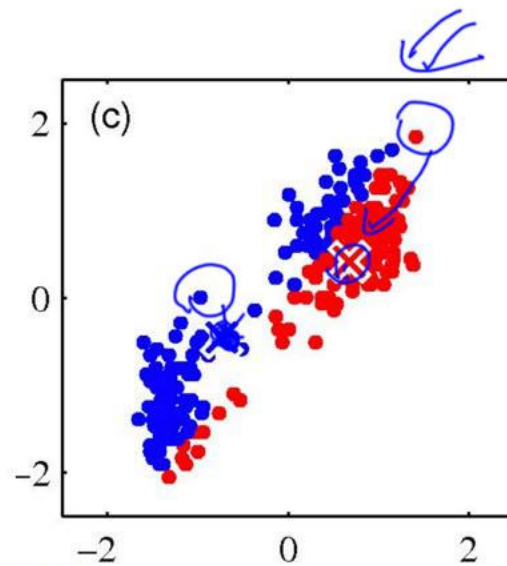
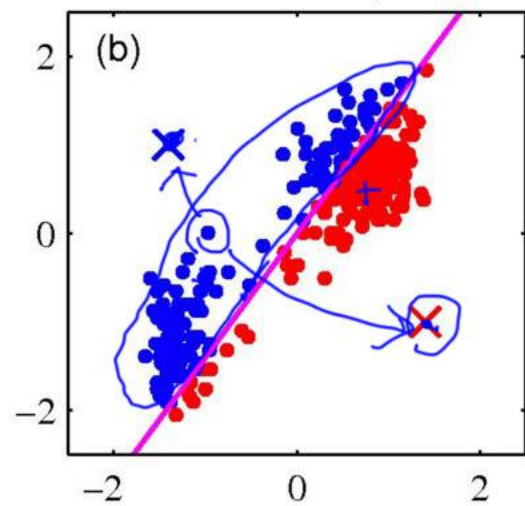
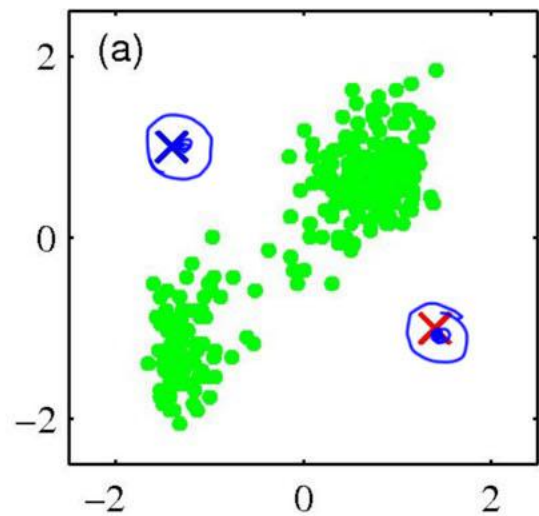
**初始化：**

随机选择K个点，作为初始中心点，每个点代表一个group.

**交替更新：**

- A. 计算每个点到所有中心点的距离，把最近的距离记录下来并赋把group赋给当前的点
- B. 针对于每一个group里的点，计算其平均并作为这个group的新的中心点。

$k=2$   $\leftarrow \hat{\mu}$



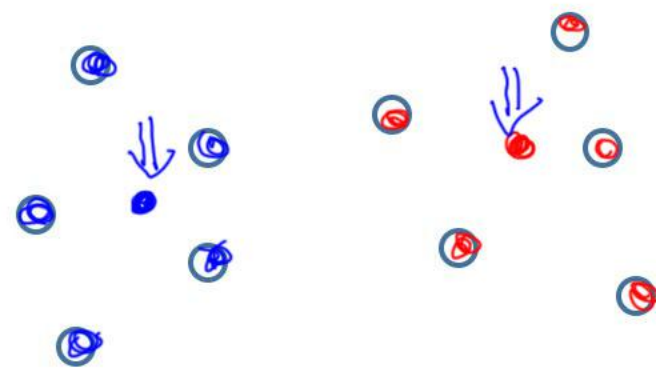
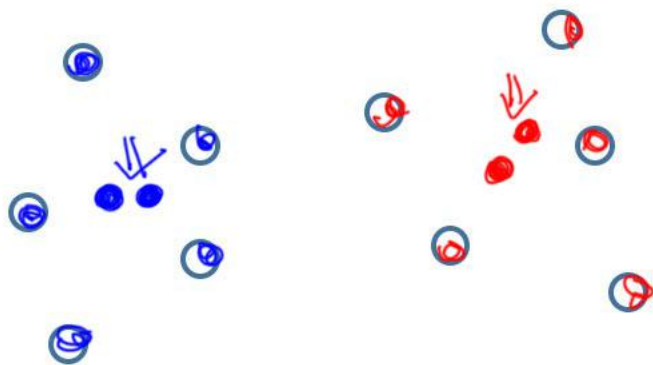
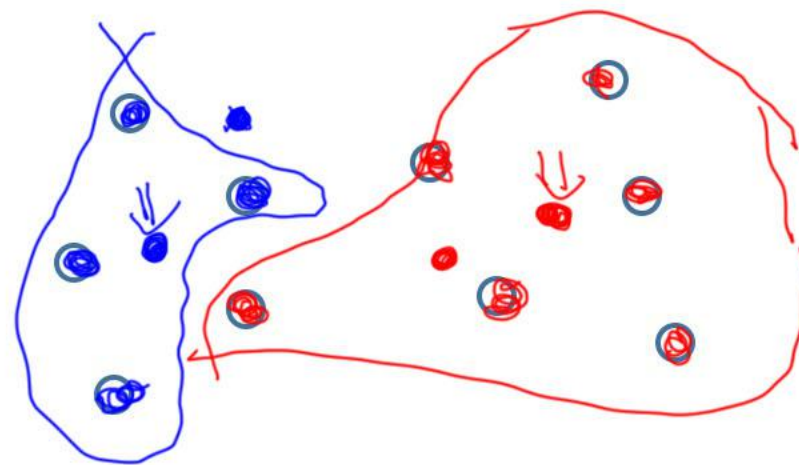
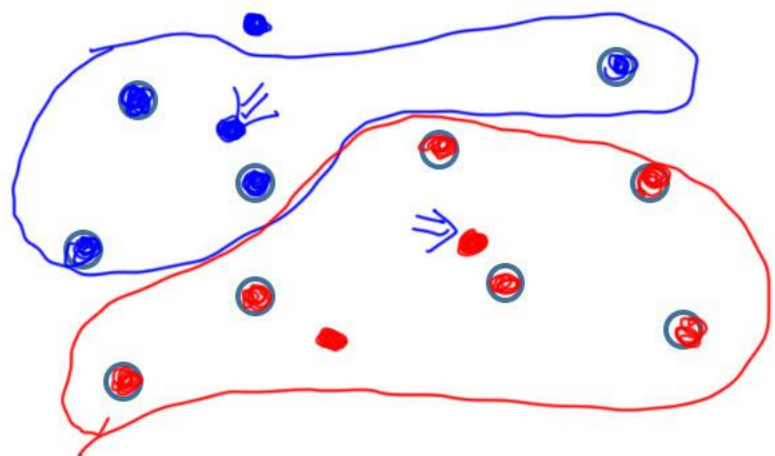
Stop

k-means

中心点:  $(x_1, y_1), (x_2, y_2) \Rightarrow \frac{x_1+x_2, y_1+y_2}{2}$

# K-means 算法过程

$k=2, k=3, k=20 \dots$



Done



# K-means算法的一些特性

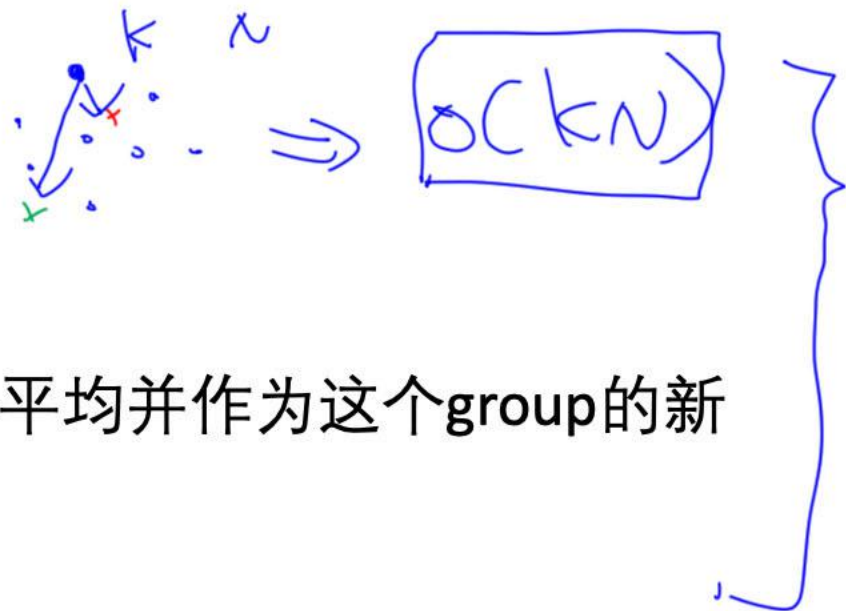
Big-O notation

## 每一次迭代的复杂度

$O(m \cdot n)$  { for  $i$  in  $\text{rag}(n)$ :  
for  $j$  in  $\text{rag}(m)$ :  
 $a = a + i \cdot j$

- 
- ① - 计算每个点到所有中心点的距离，把最近的距离记录下来并赋把group赋给当前的点

$O()$  ?

  $O(KN)$

- ② - 针对每一个group里的点，计算其平均并作为这个group的新的中心点

$O()$  ?

$O(N)$

# 关于K-means几个问题

- ✓ 1. 一定会收敛吗? ← 收敛 EM (结果一样)
- ✓ 2. 不同的初始化结果, 会不会带来不一样的结果? • 变化
- ✓ 3. K-Means的目标函数是什么? LR, SVM, 等等 Objective
- 4. K如何选择? 进阶  
↓  
超参

问题1：是否一定会收敛？

问题2：不同的初始化是否带来不一样的结果？

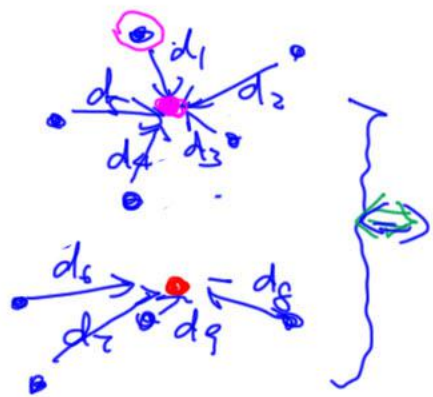


算法:

### 问题3: K-Means的目标函数

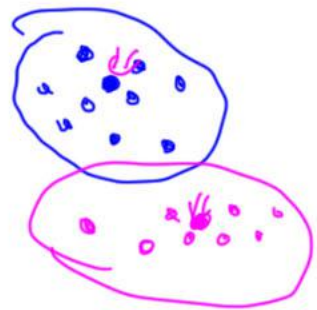
$k=3$

- ? [ ① 中心点,  
② 每一个点到底属于哪个类别



Objective:

$$\text{minimize } d_1 + d_2 + \dots + d_9$$



假设  $D = \{x_1, x_2, \dots, x_n\}$   $x_i \in \mathbb{R}^d$

参数1  $\Rightarrow$  中心点  $\mu_1, \mu_2, \dots, \mu_k$

参数2  $\Rightarrow$  属于哪个中心点,  $r_{ik} = \begin{cases} 1, & \text{if } x_i \text{ 属于第 } k \text{ 个类} \\ 0 & \text{otherwise} \end{cases}$   
Indicator Function

$$r_i = (0, 1, 0)$$

$$\text{minimize } \sum_{i=1}^n \sum_{k=1}^k r_{ik} (x_i - \mu_k)^2$$

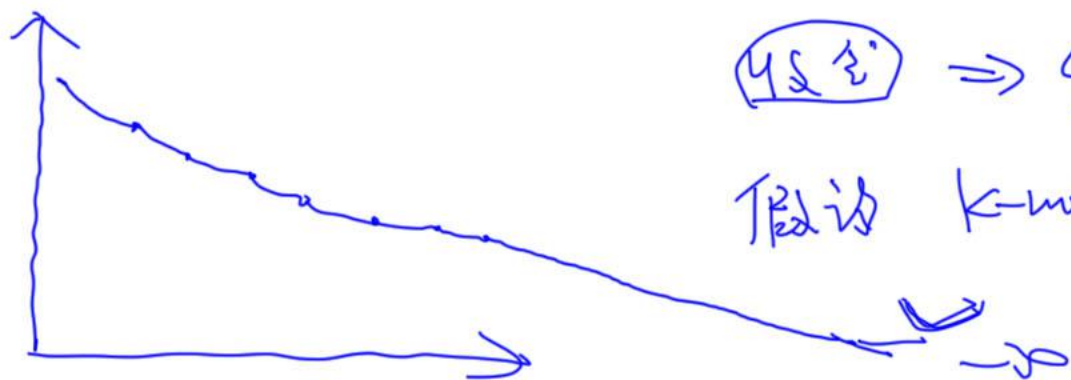
算法  $\rightarrow$  求解  $\{\mu, r\}$

EM  $\Rightarrow$  ① 假定  $\mu$  是已知的  $\Rightarrow r$  的最优  $\rightarrow$  所有的点  $\rightarrow$  分类  
 $\Rightarrow$  ② 假定  $r$  是已知的  $\Rightarrow \mu$  的最优  $\rightarrow$  平均步

每一次的Iteration 一定是变好

# 问题4: K值如何选择?

EM  $\Rightarrow$



(4, 5)  $\Rightarrow$  Converge

假设 K-means objective 不是



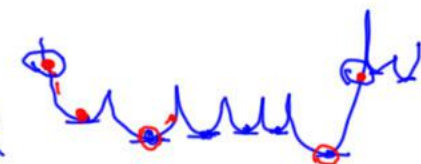
Q: 不同的初始值  $\Rightarrow$  不同的结果?

minimize

$$\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K r_{ik} (x_i - \mu_k)^2$$

凸函数

非凸函数



如何判断 f 是非凸  
 $\downarrow$  凸性  
convex optimization

非凸函数

EM

局部  
 $\uparrow$

初始化

HMM, CRF, LDA,  
Graphical Model  
GMM,

# 其他聚类算法

Soft clustering

hard clustering

1. GMM

← k-means

2. 层次聚类

3. Spectral Clustering

4. DBSCAN

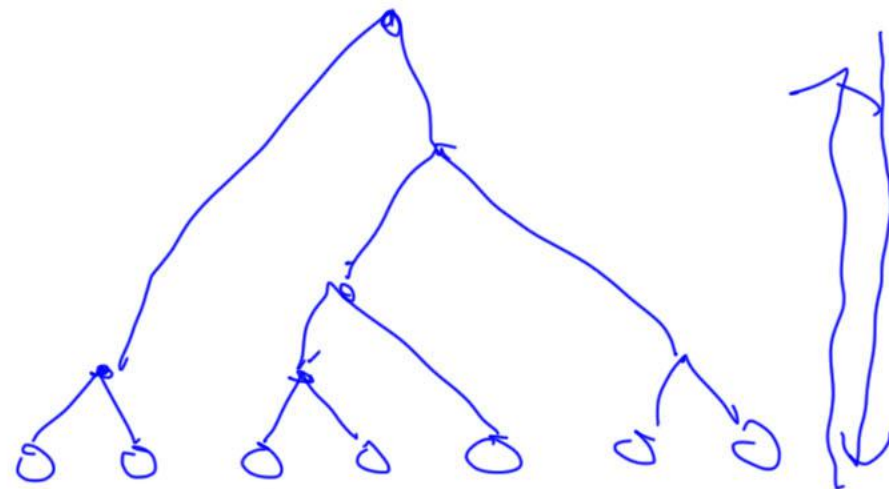
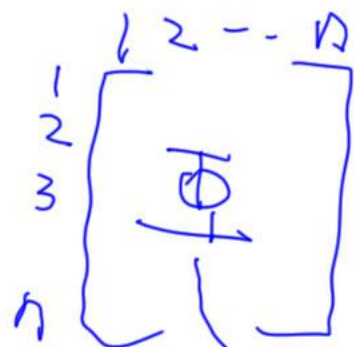
→ neighbor

5. Kernel K-Means

⇒ k-Means + kernel  
(SVM)

6. .... ✓

LDA (主题)





如何选择  $k$ ?

$k=2,$

$$\frac{\sum_{i=1}^n \sum_{k=1}^K y_{ik} (X_i - \mu_k)^2}{M}$$

$$= 27$$

$k=3,$

$$= \dots$$

$$= 20$$

$k=4,$

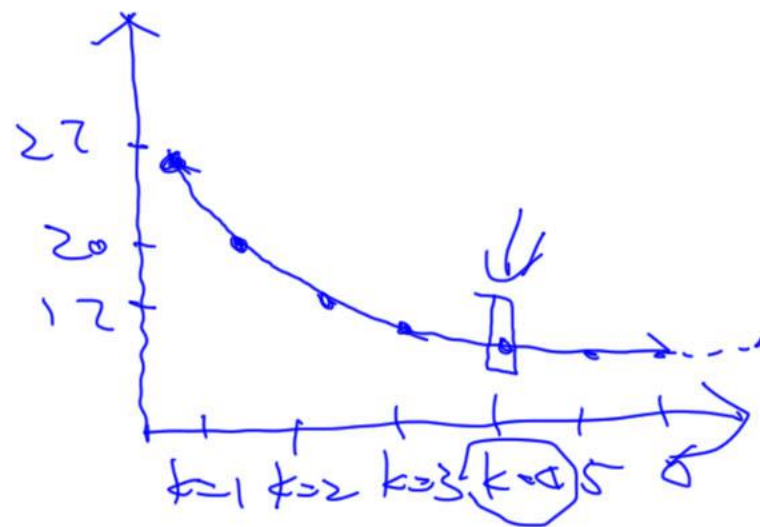
$$= 17$$

$k=5$

$$= 16$$

$k=6$

$$= 15.5$$



$k=4$



## Data Types:

① real-valued data: 175cm, 60kg, 36.5°

归一化  $\rightarrow N(0,1)$   
 $\rightarrow [0,1]$

离散化 | 0, 0, 1, 1, 2, 3, 3.5, 4 |  
 $\downarrow$   
非线性  $\downarrow$  A  $\downarrow$  B  $\downarrow$  C  
本科 > 硕士 1 2 3

② categorical data (类别型): 男/女 [本科/硕士/博士]

$\downarrow$  独热编码

本科  $\Rightarrow (1, 0, 0)$ , 硕士  $\Rightarrow (0, 1, 0)$  博士  $\Rightarrow (0, 0, 1)$

③ Ordinal data: 绩点  $\Rightarrow$  ☆☆☆☆ 3, 2, 5,  $\Leftarrow$

1, 2, 3, ...

One-hot Encoding

绩  $\Rightarrow$  非常差, 一般, 不好, 非常不好

60-75: D

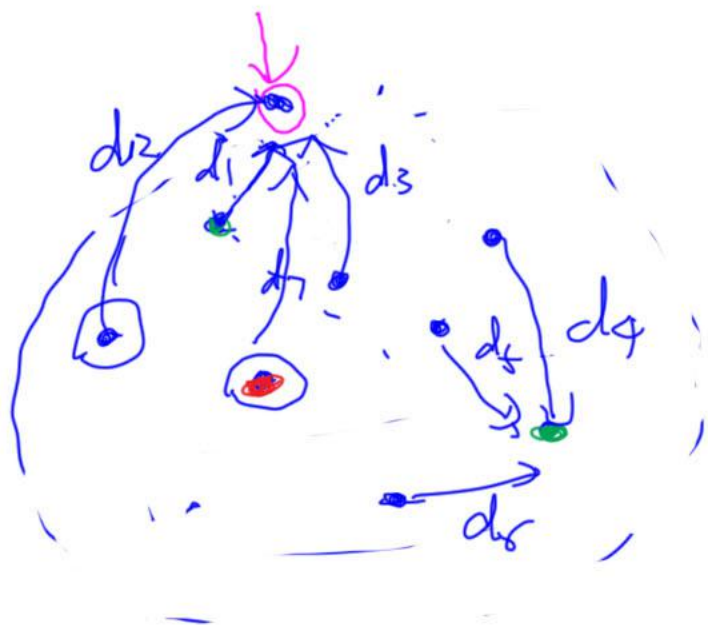
75-85: C

85-90: B

90-100: A

k-means +

$k=3$



$(d_1, d_2, d_3, \dots, d_n)$

↓ multinomial distribution

$(d_1, d_2, d_3, \dots, d_r)$

↓ sample

---

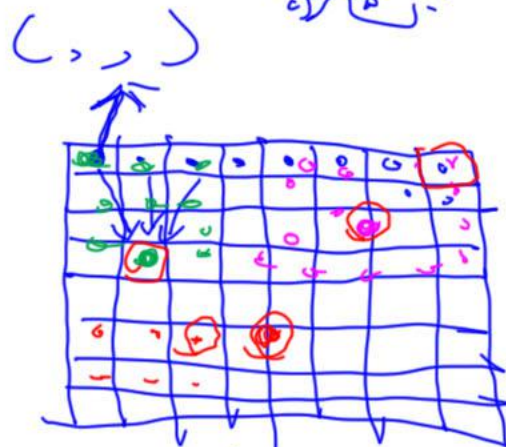
并行计算  $\Rightarrow$  Map-reduce

图片:



~~16x16~~  
8x8  
\* \*

imread  
⇒



⇓  
(8x8x3)  
↑  
(R, G, B)

黑白: 0/1  
彩色: RGB

(230, 56, 0)  
↑     ↑     ↑  
R     G     B

K=J

64

K=J

(3, 3)

64 ⇒ J

(64 / 5)

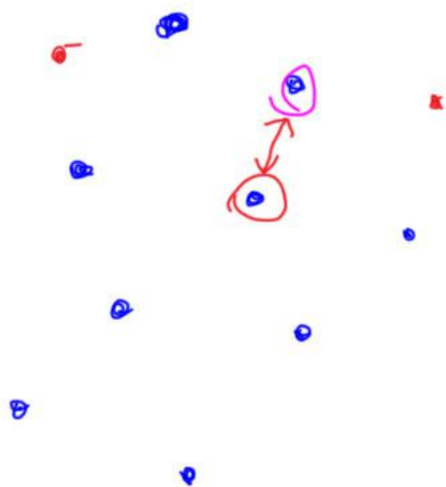
8x8x3

⇓

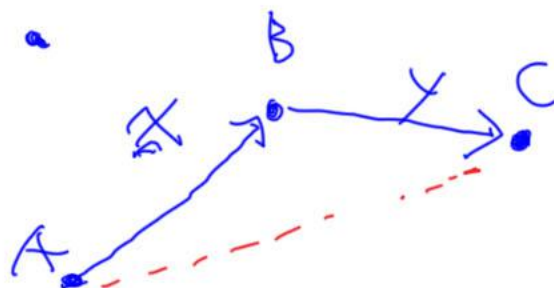
64x3

(8x8, 3)  
64, 3

k-means



k-means++



Ex: Distance  
is a metric

$$d(x, y) + d(y, z) \geq d(x, z)$$

$d(A, C) \leq x + y$



