西南財經大學
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS
1925 - 2021
经世济民 孜孜以求

机器学习基础

聚类方法

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 聚类的基本概念
- 层次聚类
- *k*均值聚类

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 聚类的基本概念
- 层次聚类
- *k*均值聚类

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

聚类方法

- 聚类是依据样本的相似度，将其归并到若干个类
  - 相似的在同一个类；不相似的在不同类
- 层次聚类
  - 聚合：自下而上
  - 分裂：自上而下
- k均值聚类

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

相似度或距离

- 假设样本的特征向量为 $x_i = (x_{1i}, x_{2i}, \cdots, x_{mi}),\ i = 1, \cdots, n$
- 闵可夫斯基距离 $d_{ij} = \left(\sum_{k=1}^{m} |x_{ki} - x_{kj}|^p\right)^{\frac{1}{p}},\ p \geq 1$
  - $p = 2$ 时，欧氏距离 $d_{ij} = \left(\sum_{k=1}^{m} |x_{ki} - x_{kj}|^2\right)^{\frac{1}{2}}$
  - $p = 1$ 时，曼哈顿距离 $d_{ij} = \sum_{k=1}^{m} |x_{ki} - x_{kj}|$
  - $p = +\infty$ 时，切比雪夫距离 $d_{ij} = \max_k |x_{ki} - x_{kj}|$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

马哈拉诺比斯距离

- 假设 $S=[Cov(X_i,X_j)]_{m\times m}$，其中 $X_i=(x_{i1},x_{i2},\cdots,x_{in})$，$i=1,\cdots,m$
- 马哈拉诺比斯距离 $d_{ij}=[(x_i-x_j)^{\top}S^{-1}(x_i-x_j)]^{\frac{1}{2}}$
  - 马氏距离越大，相似度越小
  - 当协方差矩阵 S 为单位矩阵时，马氏距离就是欧式距离
  - 马氏距离不受量纲的影响，还可以排除变量之间的相关性的干扰

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

相关系数

- 相关系数 $r_{ij}=\dfrac{Cov(X_i,X_j)}{\sqrt{Var(X_i)Var(X_j)}}$

$$=\frac{\sum_{k=1}^{m}\left(x_{ki}-\bar{x}_i\right)\left(x_{kj}-\bar{x}_j\right)}{\left[\sum_{k=1}^{m}\left(x_{ki}-\bar{x}_i\right)^2\sum_{k=1}^{m}\left(x_{kj}-\bar{x}_j\right)^2\right]^{\frac{1}{2}}}$$

- 相关系数的绝对值越接近 1 表示样本越相似
- 相关系数的绝对值越接近 0 表示样本越不相似

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

余弦夹角

- 余弦夹角 $s_{ij} = \frac{X_i \cdot X_j}{|X_i|\,|X_j|}$

$$= \frac{\sum_{k=1}^{m} x_{ki} x_{kj}}{\left[\sum_{k=1}^{m} x_{ki}^2 \sum_{k=1}^{m} x_{kj}^2\right]^{\frac{1}{2}}}$$

- 余弦夹角越接近 1 表示样本越相似
- 余弦夹角越接近 0 表示样本越不相似

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

类的定义

- 设 T 为给定正数，若集合 G 中任意两个样本 x_i, x_j 有 $d_{ij} \leq T$
- 集合中任意样本 x_i 一定存在 G 中另一个样本 x_j 有 $d_{ij} \leq T$
- 集合中任意样本 x_i 有 $\frac{1}{n_G - 1} \sum_{x_j \in G} d_{ij} \leq T$
- 设 T, V 为给定正数，集合中任意两个样本 x_i, x_j 有

$$\frac{1}{n_G(n_G - 1)} \sum_{x_i \in G} \sum_{x_j \in G} d_{ij} \leq T$$

$$d_{ij} \leq V$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

类的特征

- 类的均值，有称为类的中心 $\bar{x}_G = \frac{1}{n_G}\sum_{i=1}^{n_G} x_i$
- 类的直径，即类中任意两个样本之间的最大距离 $D_G = \max_{x_i, x_j \in G} d_{ij}$
- 类的样本散布矩阵 $A_G = \sum_{i=1}^{n_G}(x_i - \bar{x}_G)(x_i - \bar{x}_G)^\top$
- 类的样本协方差矩阵 $S_G = \frac{1}{m-1} A_G$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

类与类之间的距离

- 两类中样本的最短距离 $D_{pq}=\min\{d_{ij}|x_i\in G_p,\ x_j\in G_q\}$
- 两类中样本的最长距离 $D_{pq}=\max\{d_{ij}|x_i\in G_p,\ x_j\in G_q\}$
- 两类中心之间的距离 $D_{pq}=d_{\bar{x}_p\bar{x}_q}$
- 两类中任意两点之间距离的平均值

$$D_{pq}=\frac{1}{n_p n_q}\sum_{x_i\in G_p}\sum_{x_i\in G_q}d_{ij}$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 聚类的基本概念
- 层次聚类
- *k*均值聚类

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

层次聚类

- 聚合聚类
  - 开始时每个样本各自分为一类；之后将距离最近的两类合并，建立一个新类；重复此操作直到满足停止条件
- 分裂聚类
  - 开始时所有样本分到一类；之后将已有类中相距最远的样本分到两个新的类，重复此操作直到满足停止条件

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

层次聚类

- 聚合聚类和分裂聚类会得到同样的结果，但分裂聚类的计算量要远大于聚合聚类
- 优点：距离或相似度容易定义，限制少，不需要预先制定簇的个数，可以发现簇的层次关系。
- 缺点：计算复杂度太高，奇异值也能产生很大影响，算法很可能聚类成链状。

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

聚合聚类算法

- 计算 n 个样本两两之间的距离 d_{ij}
- 构造 n 个类，每个类只包含一个样本
- 合并类间距离最小的两个类，其中最短距离为类间距离，构建一个新类
- 计算新类与当前各类的距离，直到所有样本合并为一类
- 计算复杂度是 $O(n^3)$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 聚类的基本概念
- 层次聚类
- *k*均值聚类

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

模型

- 将 n 个样本划分为两两不相交的 k 类 $G_1, G_2, \cdots, G_k$
- 每一种划分对应函数 C 表示第 i 个样本所属的类为 $C(i)$
- 以欧式距离为距离度量，则每一种划分对应的损失函数为

$$W(C) = \sum_{l=1}^{k} \sum_{C(i)=l} \|x_i - \bar{x}_l\|^2$$

  - 其中 $\bar{x}_l$ 是第 l 的类中心

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

策略

- k 均值聚类就是找出最佳划分

$$C^* = \arg\min_C W(C) = \arg\min_C \sum_{l=1}^{k} \sum_{C(i)=l} \| x_i - \bar{x}_l \|^2$$

- 将 n 个样本分到k类，所有可能分法的数目是

$$S(n,k) = S(n-1,k-1) + k \times S(n-1,k)$$

$$= \frac{1}{k!} \sum_{l=0}^{k} (-1)^l C_k^l (k-l)^n$$

  - 要在指数级的候选函数中找到最优是NP难问题

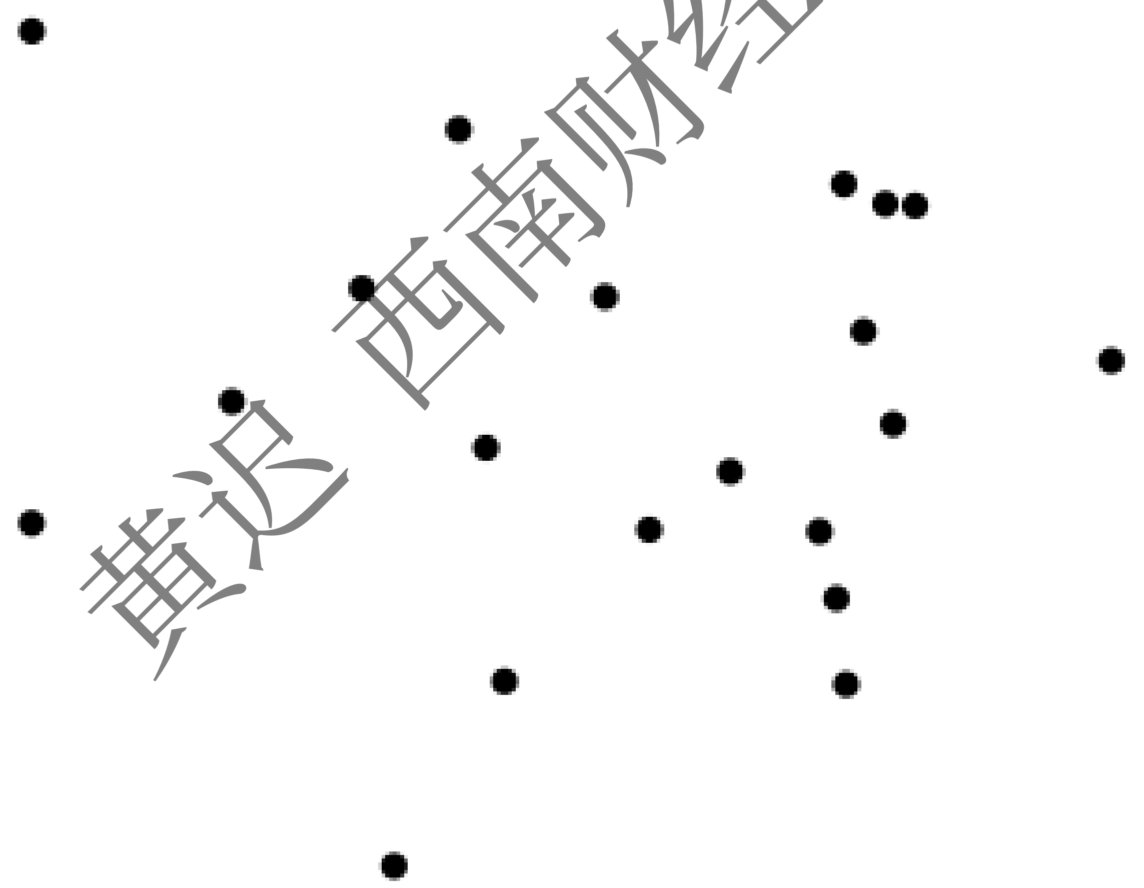
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

算法

- 初始化：随机选取 k 个样本为类中心 $m^{(0)}=(m_1^{(0)},m_2^{(0)},\cdots,m_k^{(0)})$
- 第 $t+1$ 轮聚类：对当前类中心 $m^{(t)}=(m_1^{(t)},m_2^{(t)},\cdots,m_k^{(t)})$，计算样本到类中心的距离，将样本指派到与其最近的类，得到样本划分 $C^{(t+1)}$
- 计算新的类中心：对新的划分 $C^{(t+1)}$ 重新计算各个类的中心，作为新的类中心 $m^{(t+1)}=(m_1^{(t+1)},m_2^{(t+1)},\cdots,m_k^{(t+1)})$
- 重复以上过程直到算法收敛或满足停止条件（类中心改变小）
- 算法的计算复杂度为 $O(mnk)$

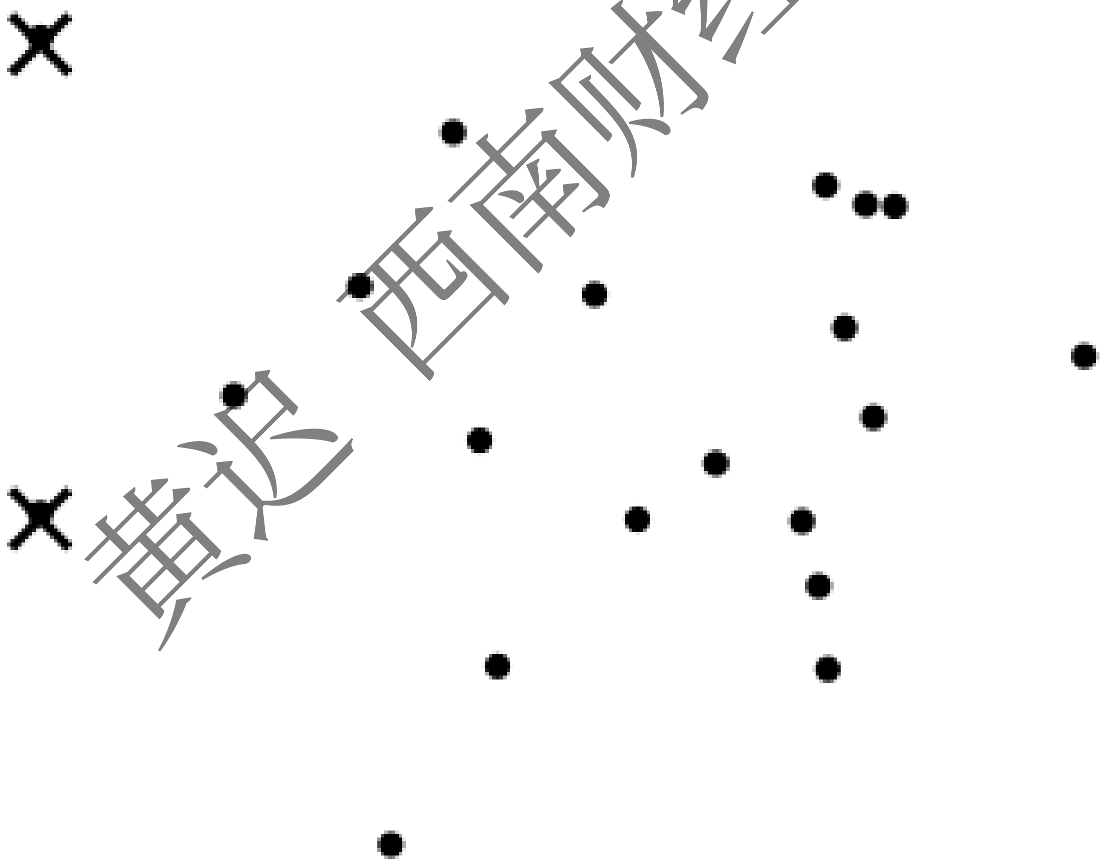
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子



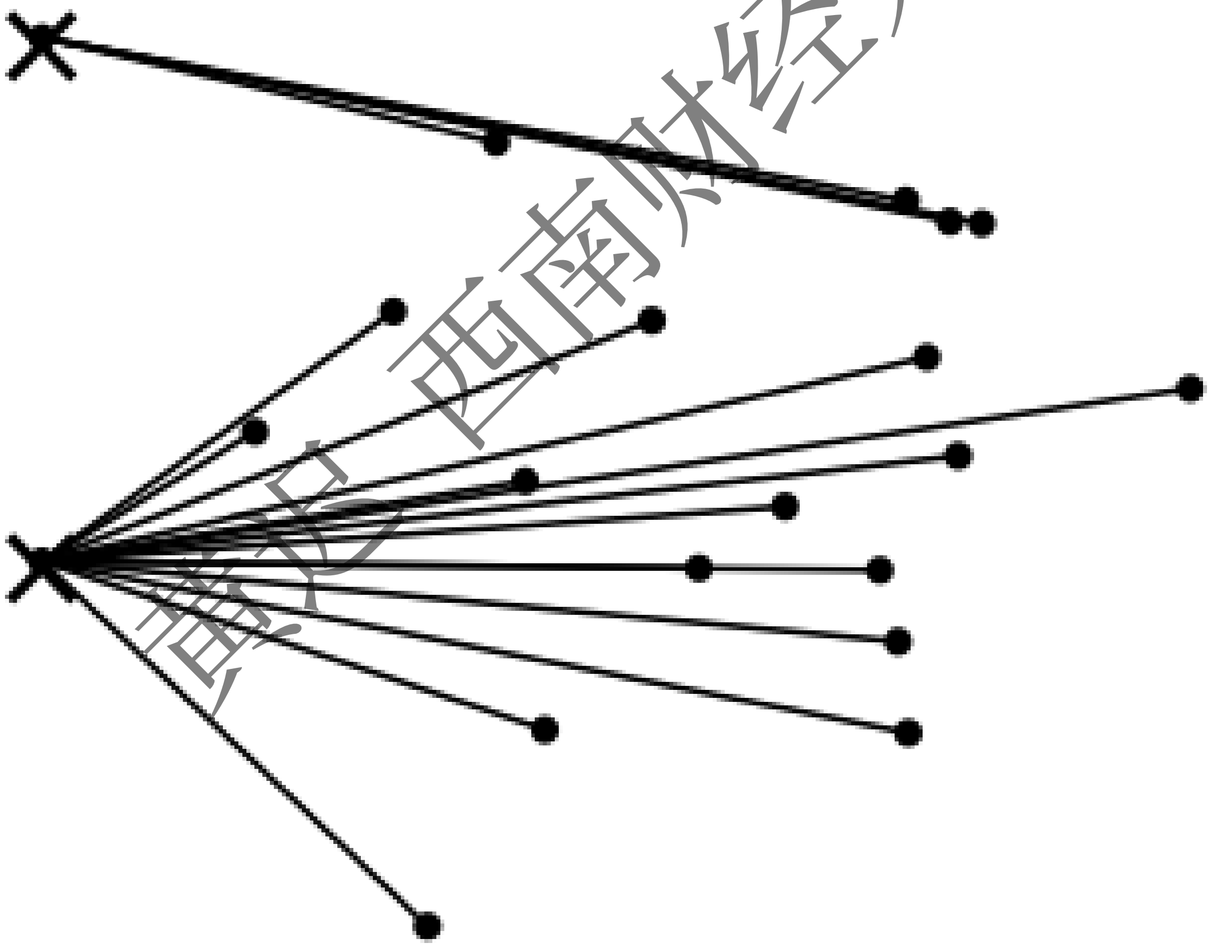
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：随机选择两个种子



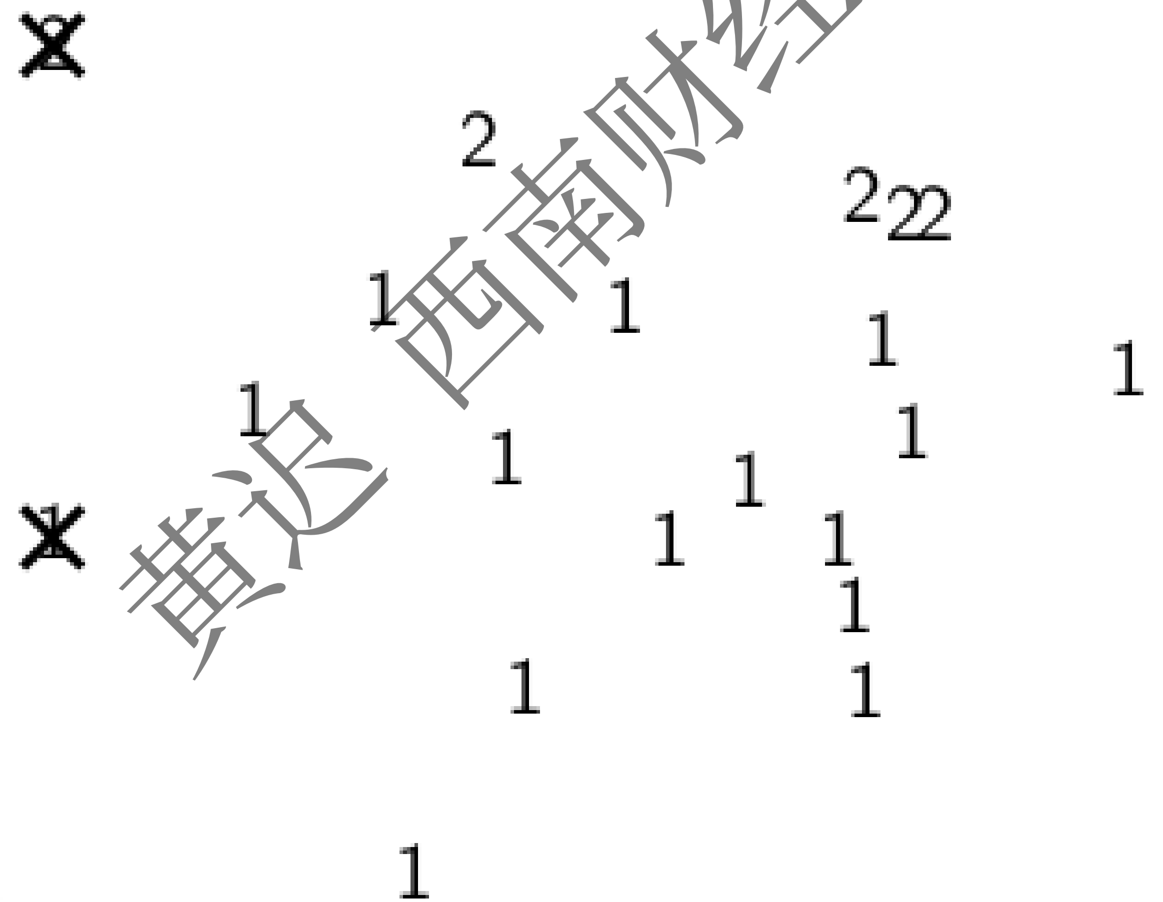
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：将样本分配给离它最近的中心(第一次)



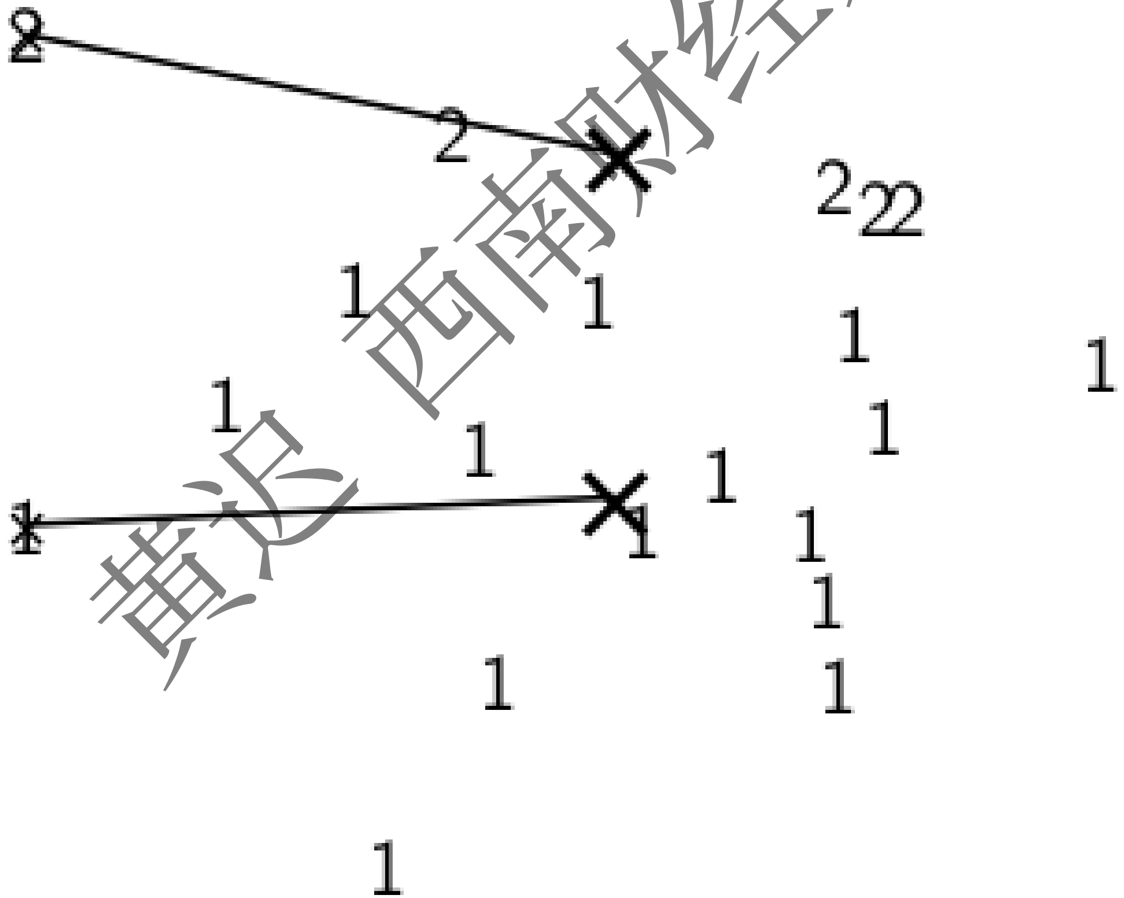
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：分配后的类(第一轮)



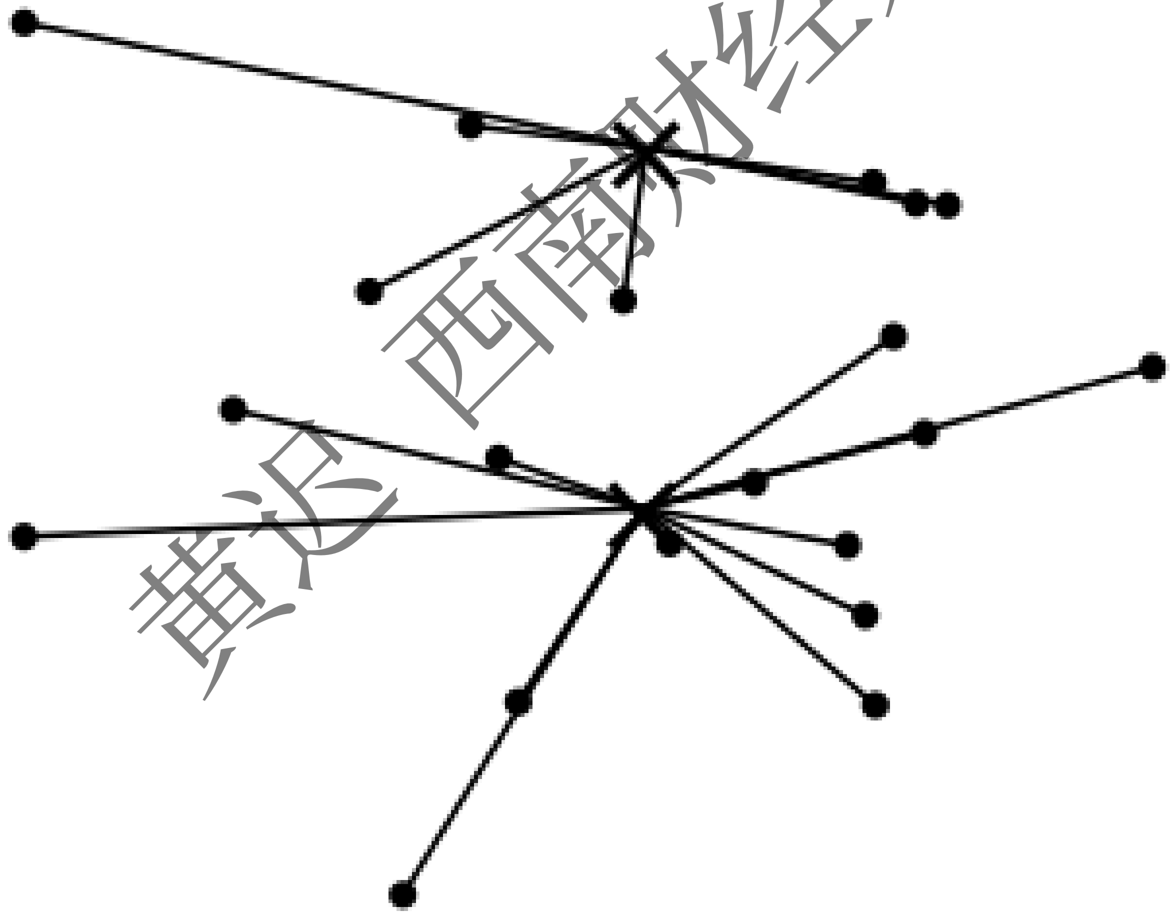
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：重新计算类中心



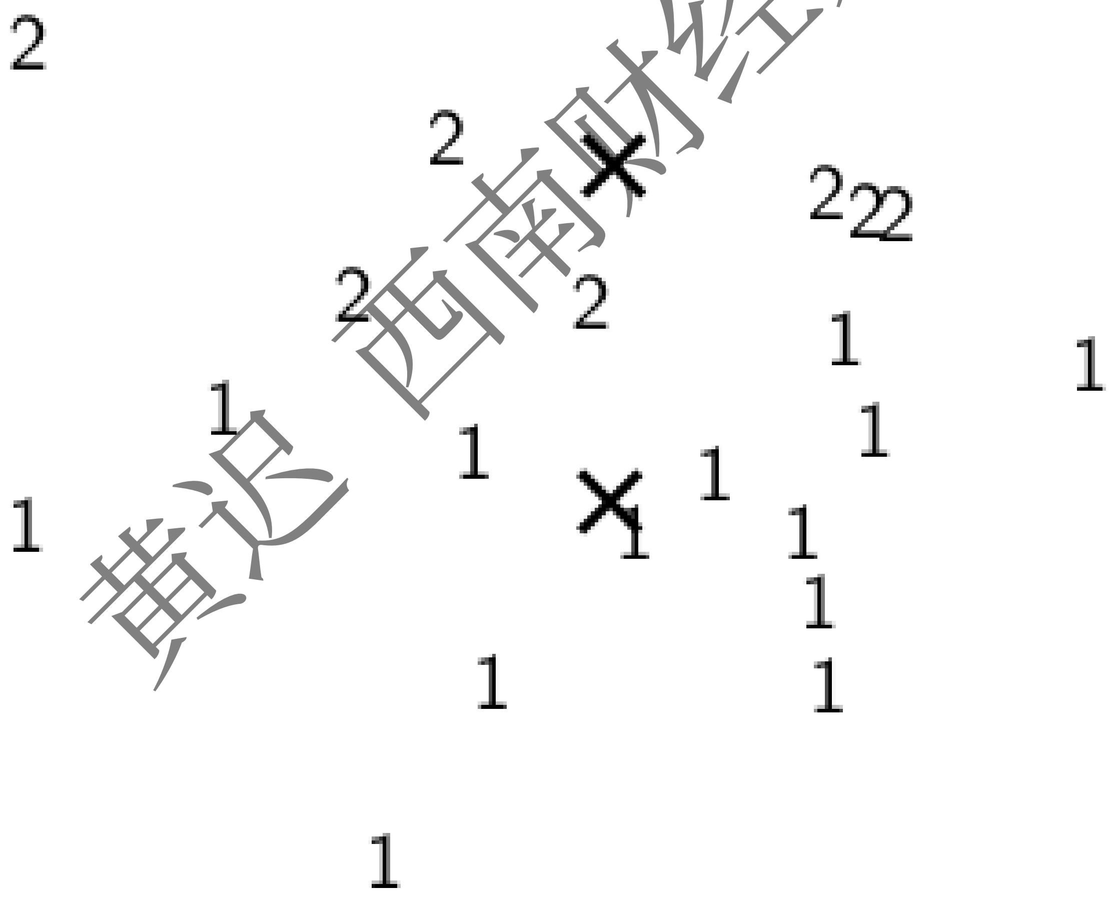
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：将样本分配给离它最近的中心(第二轮)

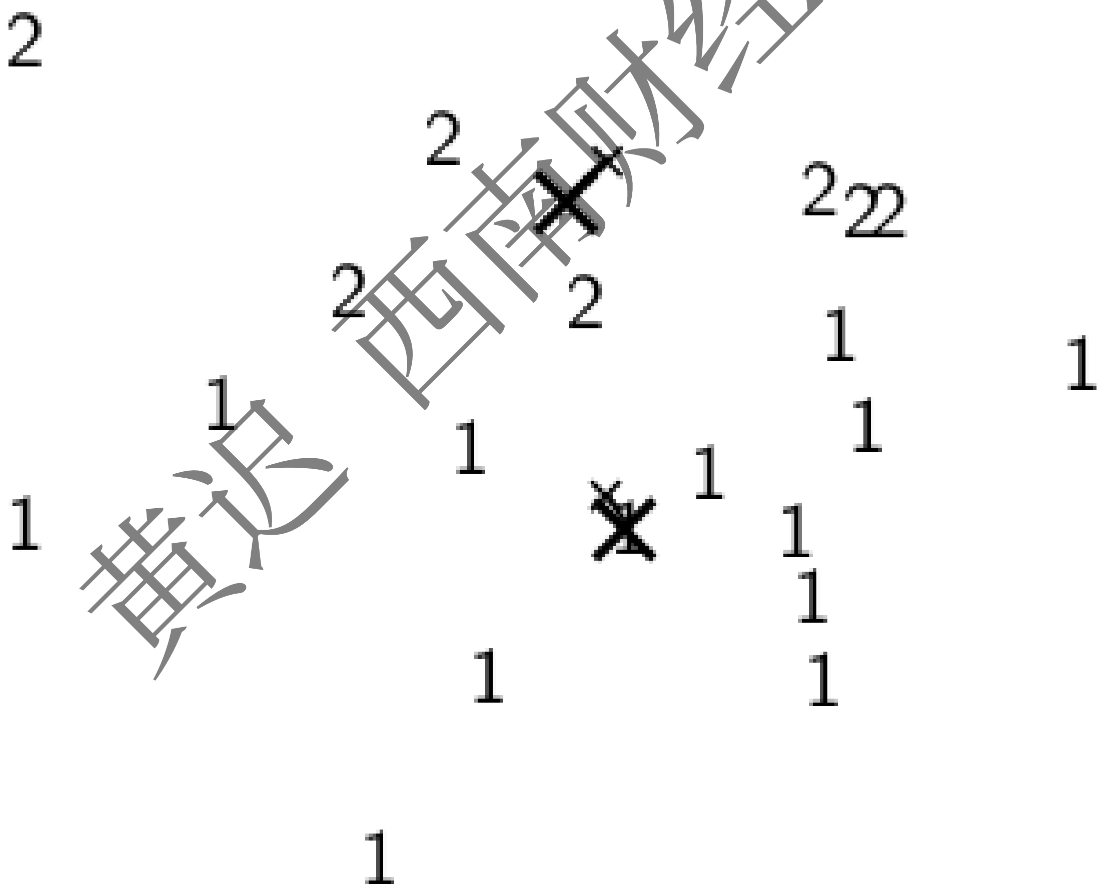


黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：重新分配的结果

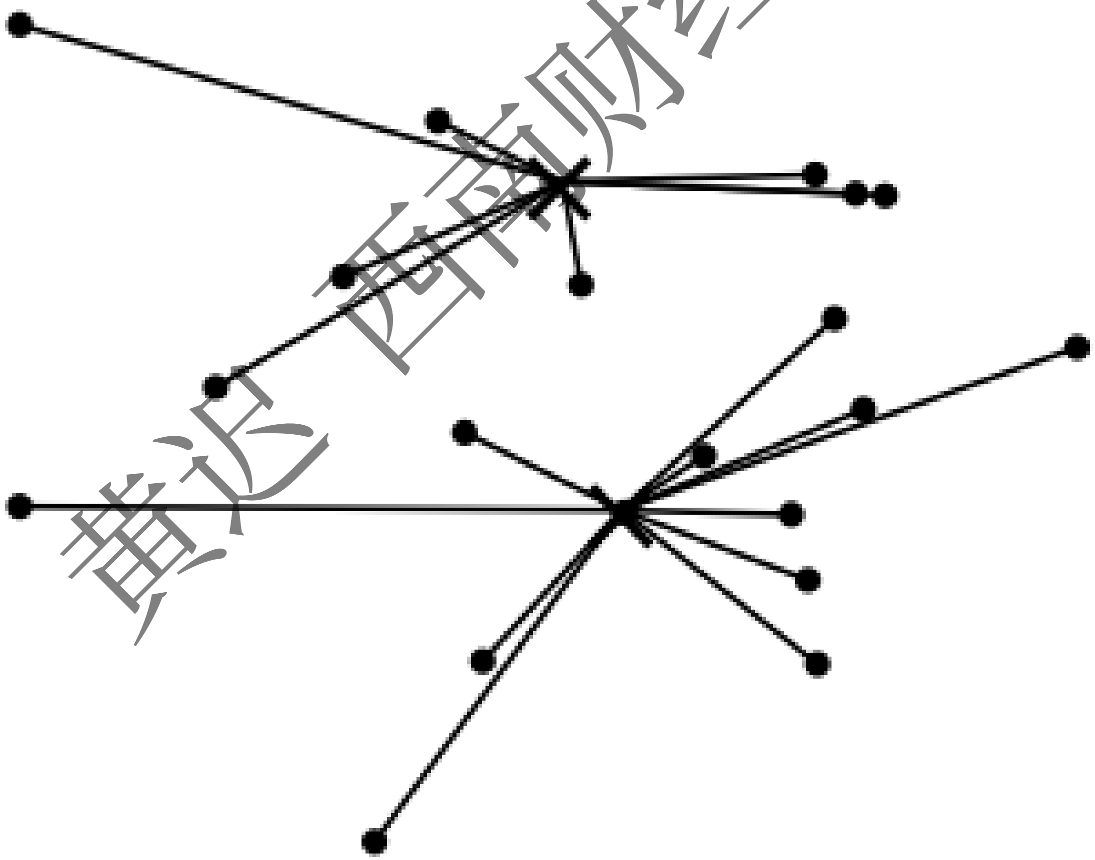


例子：重新计算类中心



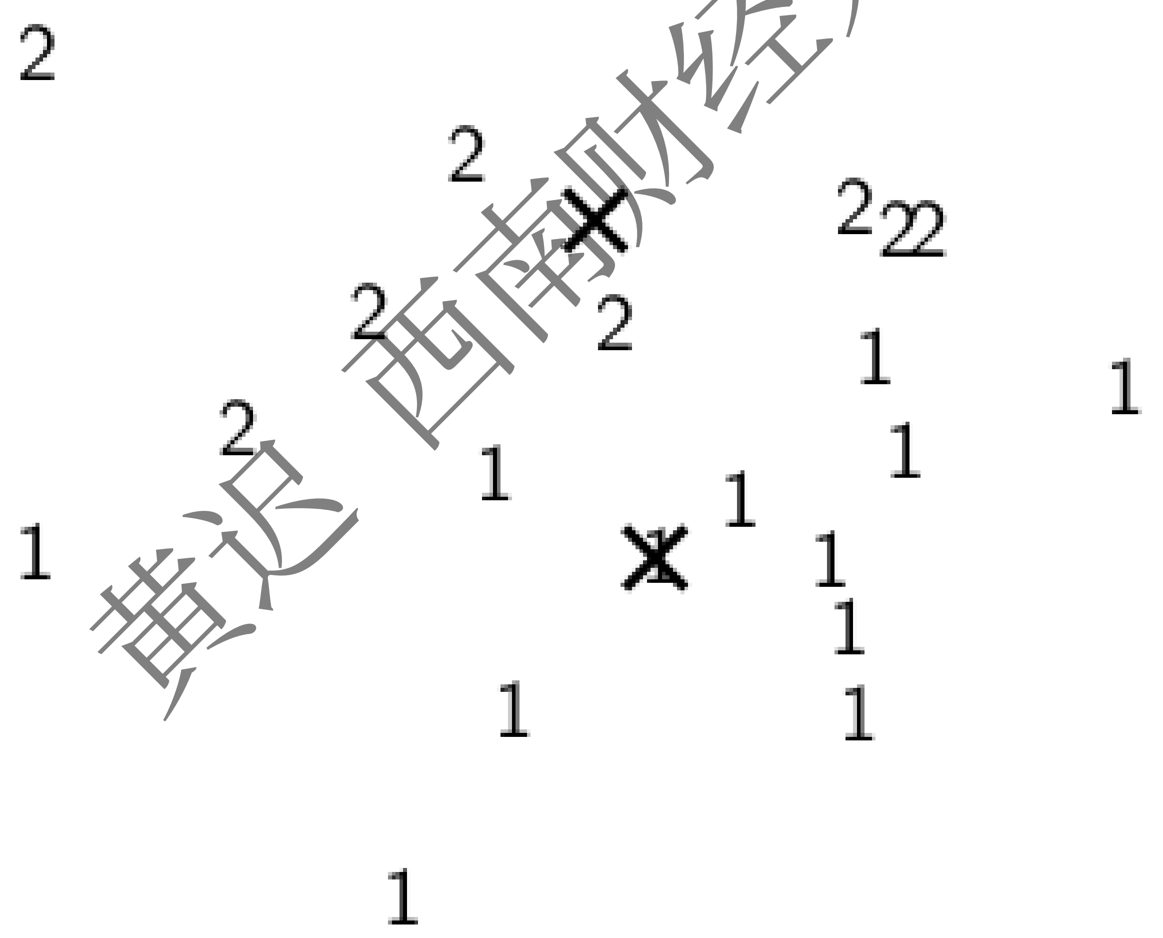
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：再重新分配(第三轮)



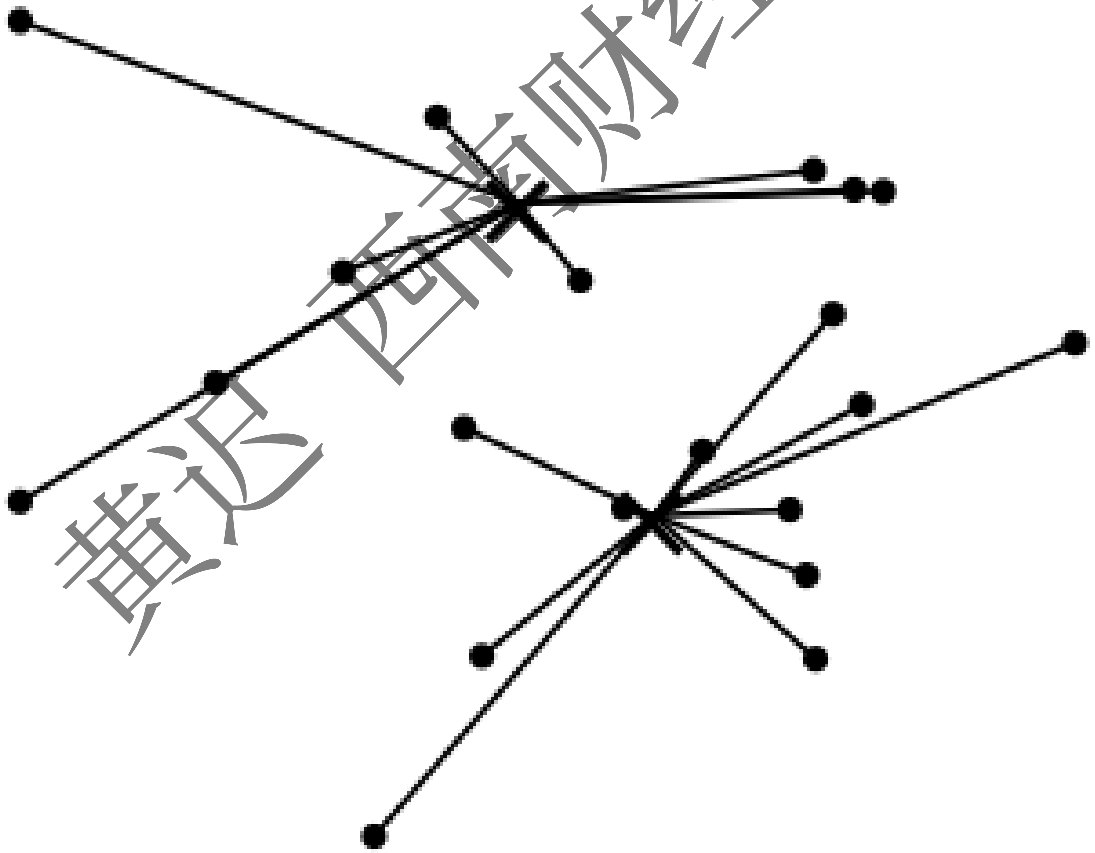
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：分配结果



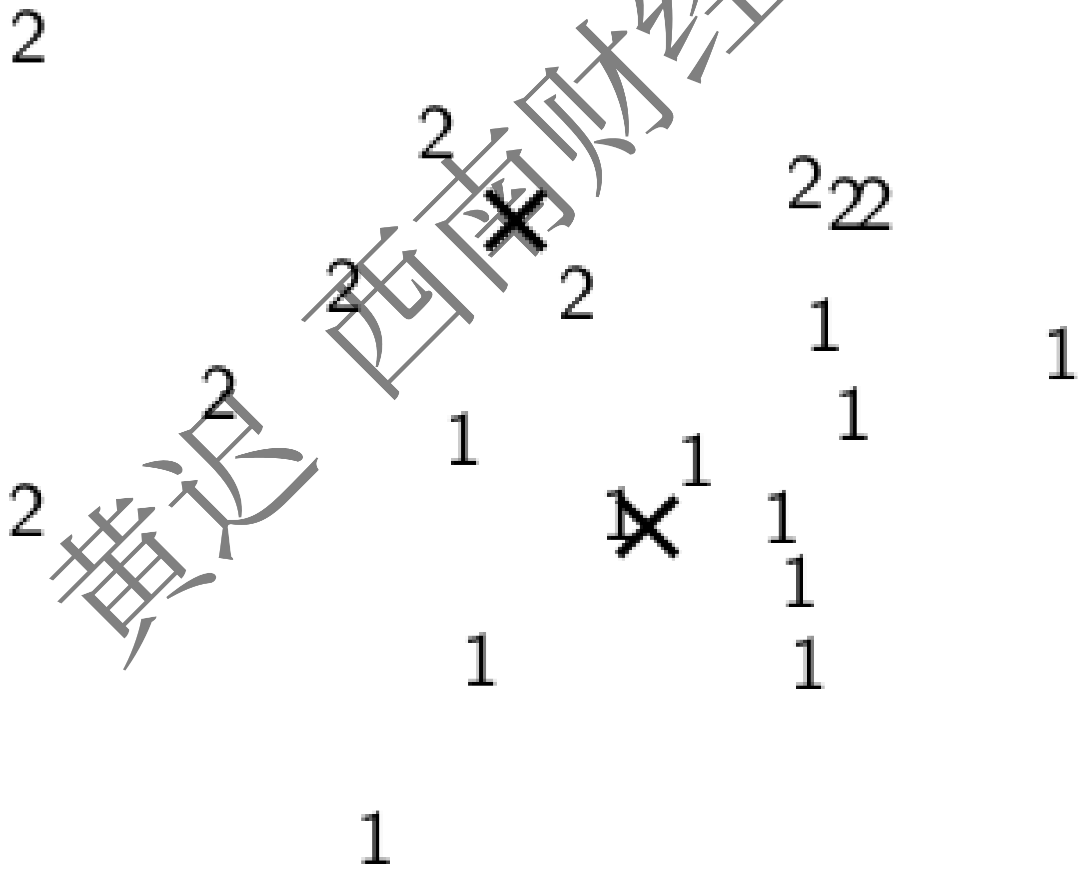
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：再重新分配(第四轮)



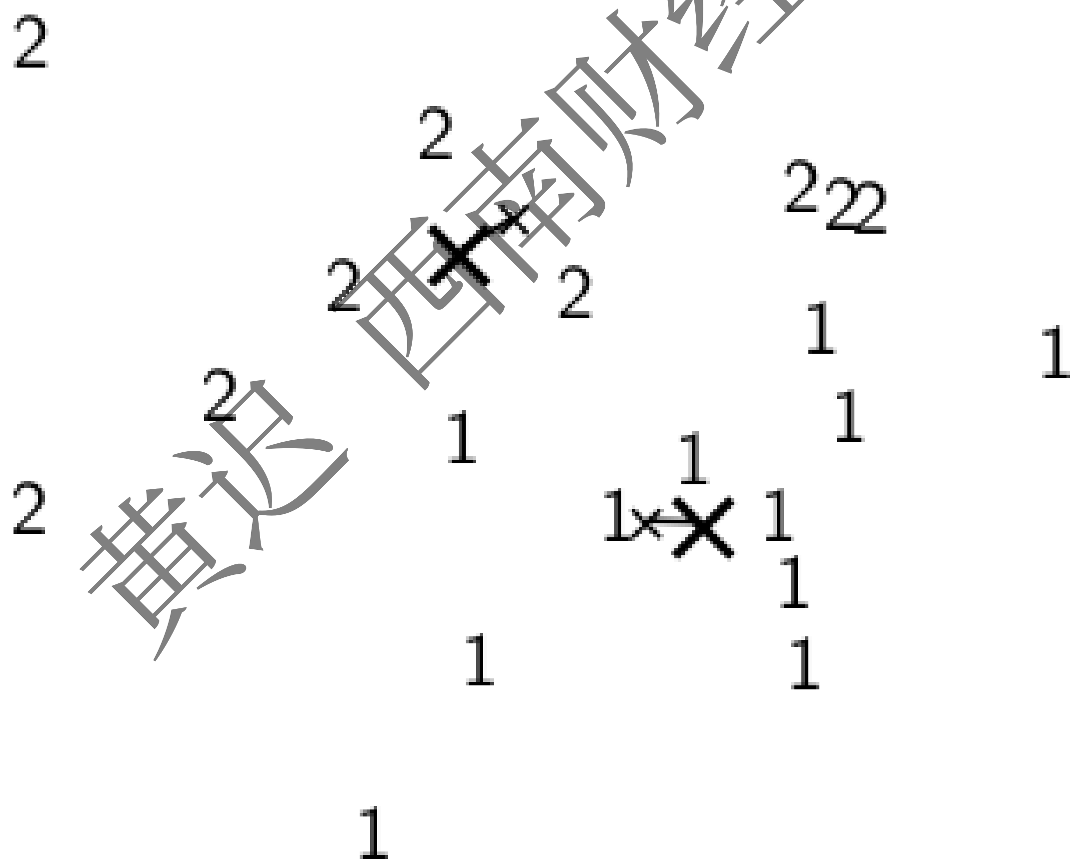
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：分配结果



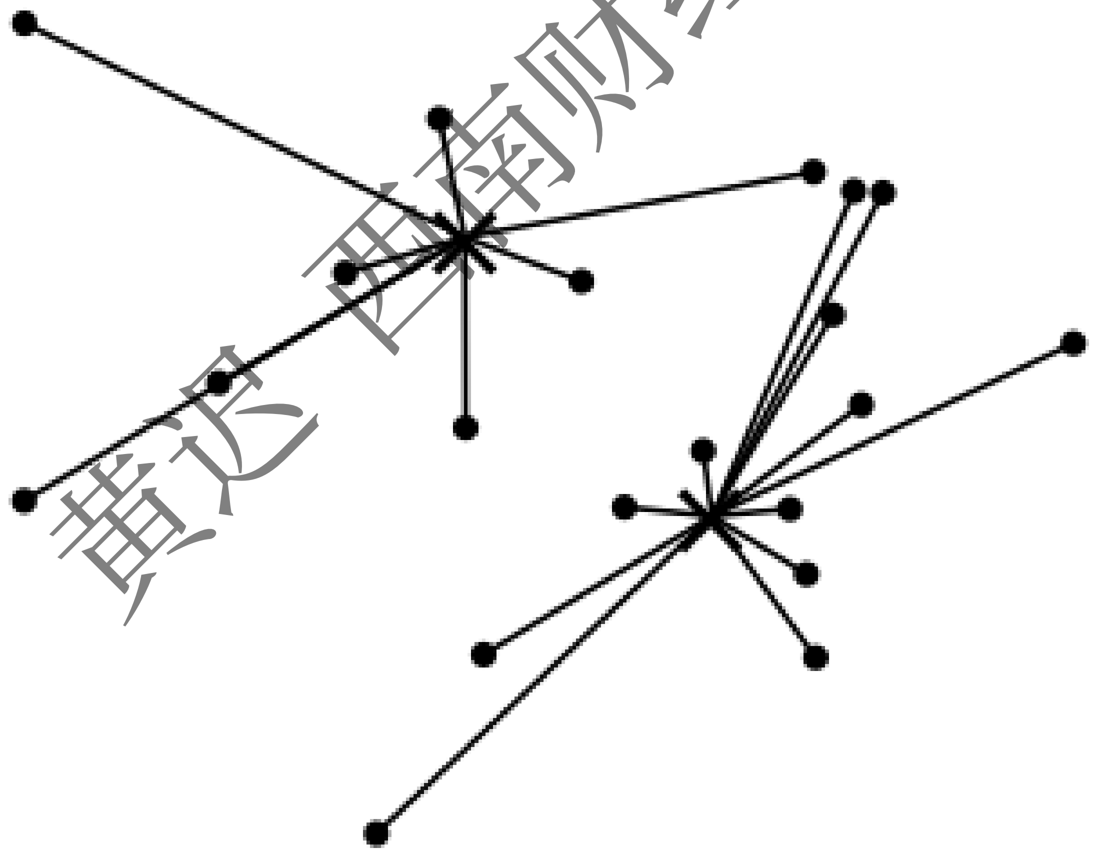
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：重新计算类中心



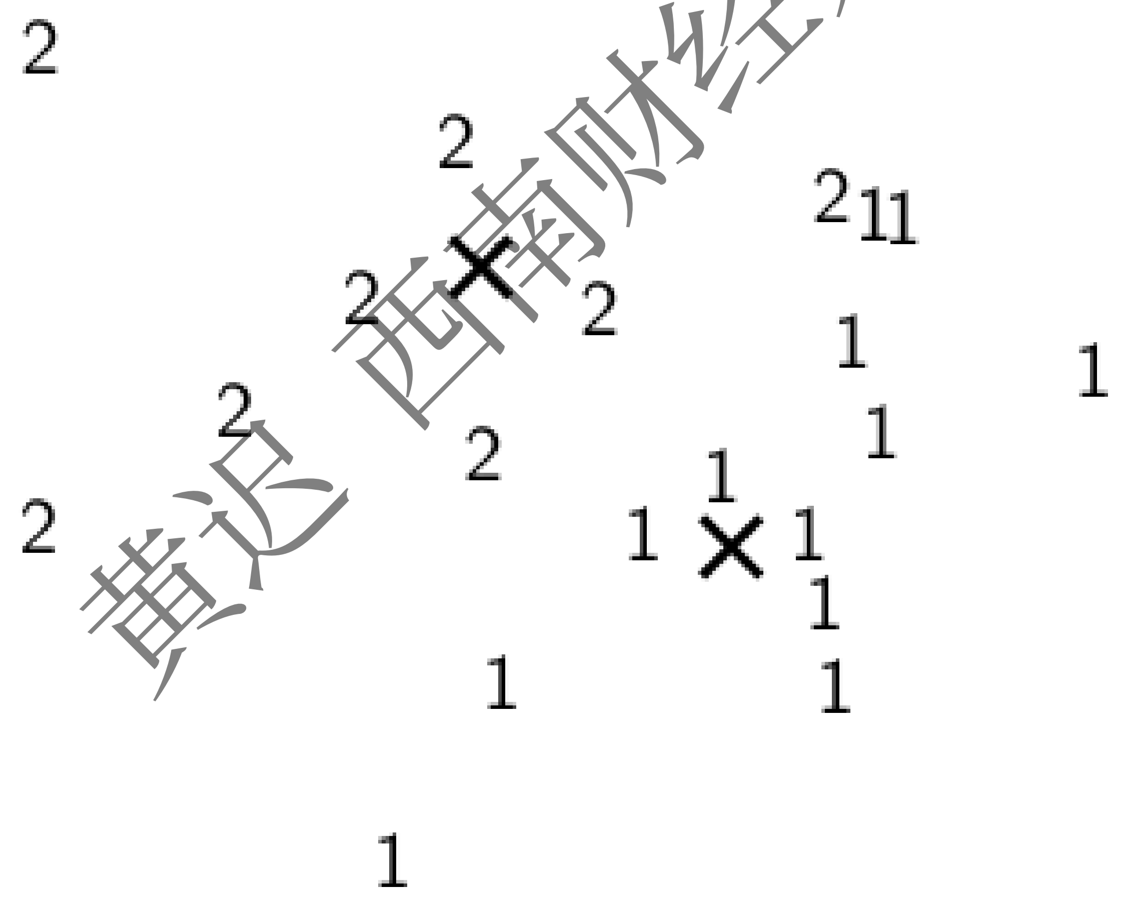
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：重新分配(第五轮)



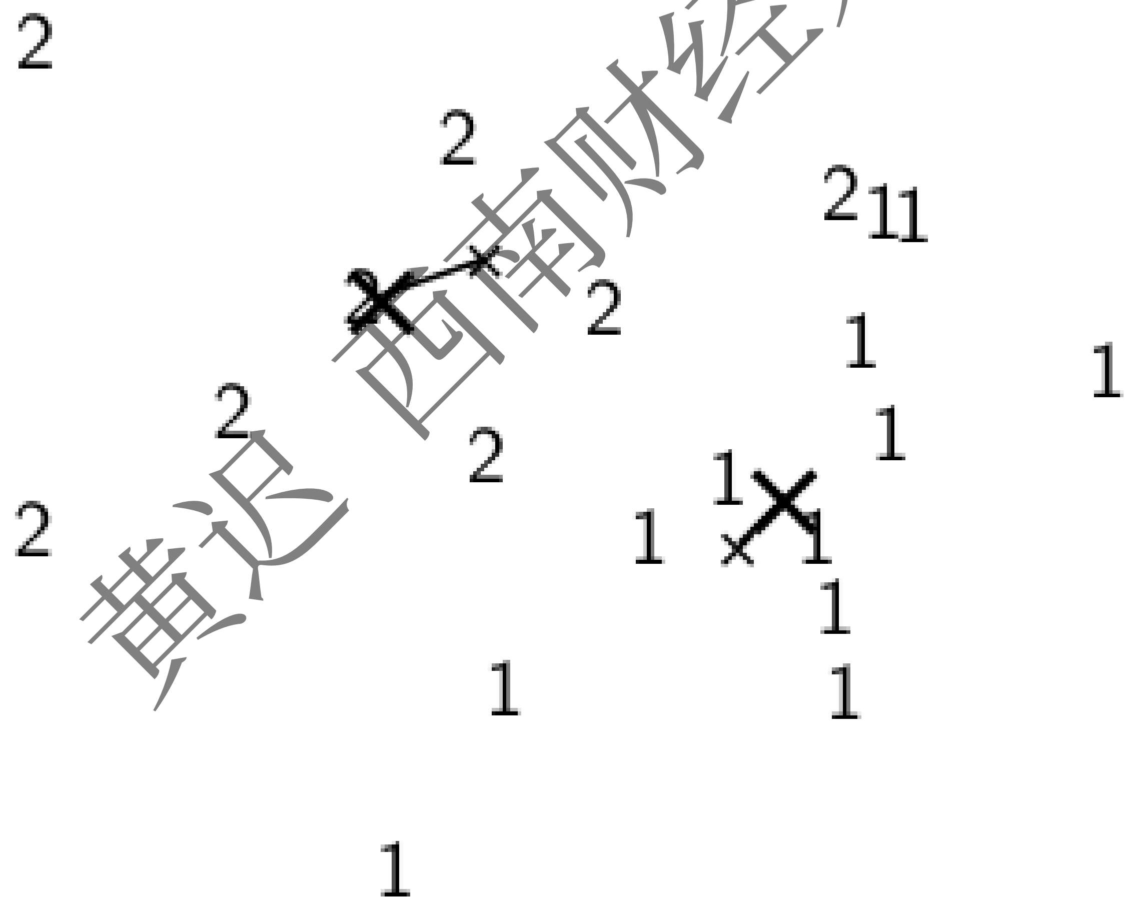
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：分配结果



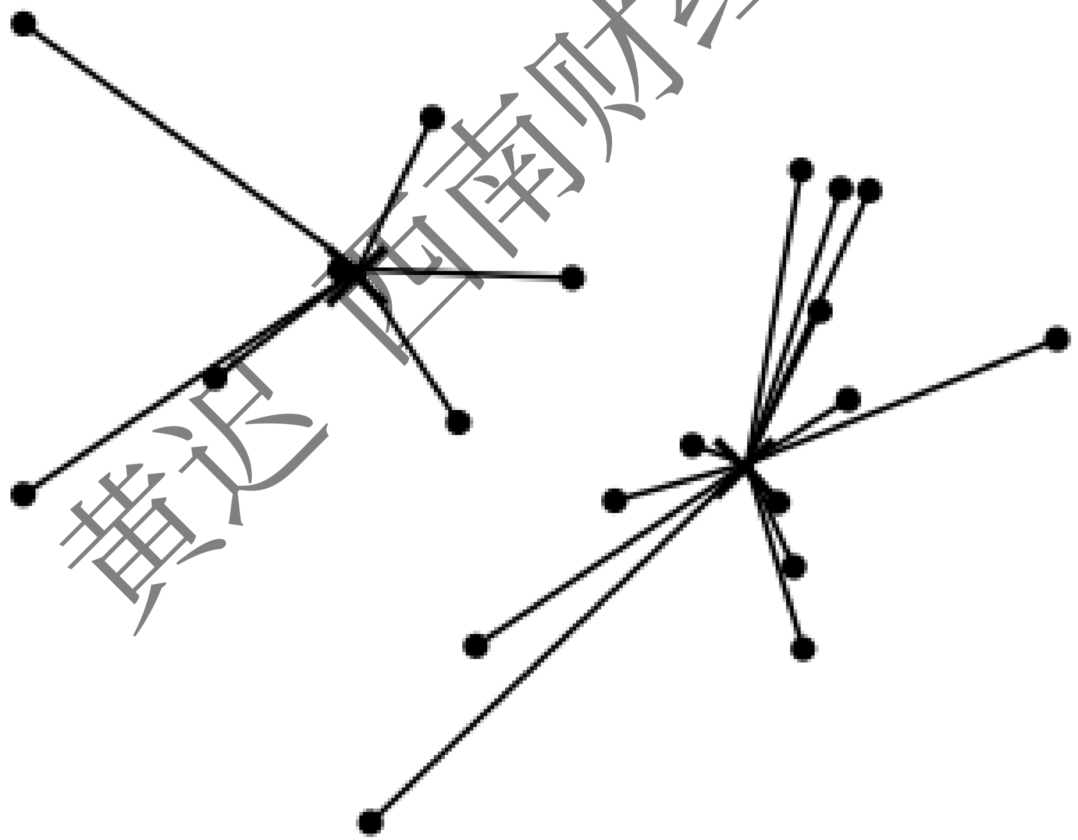
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：重新计算类中心



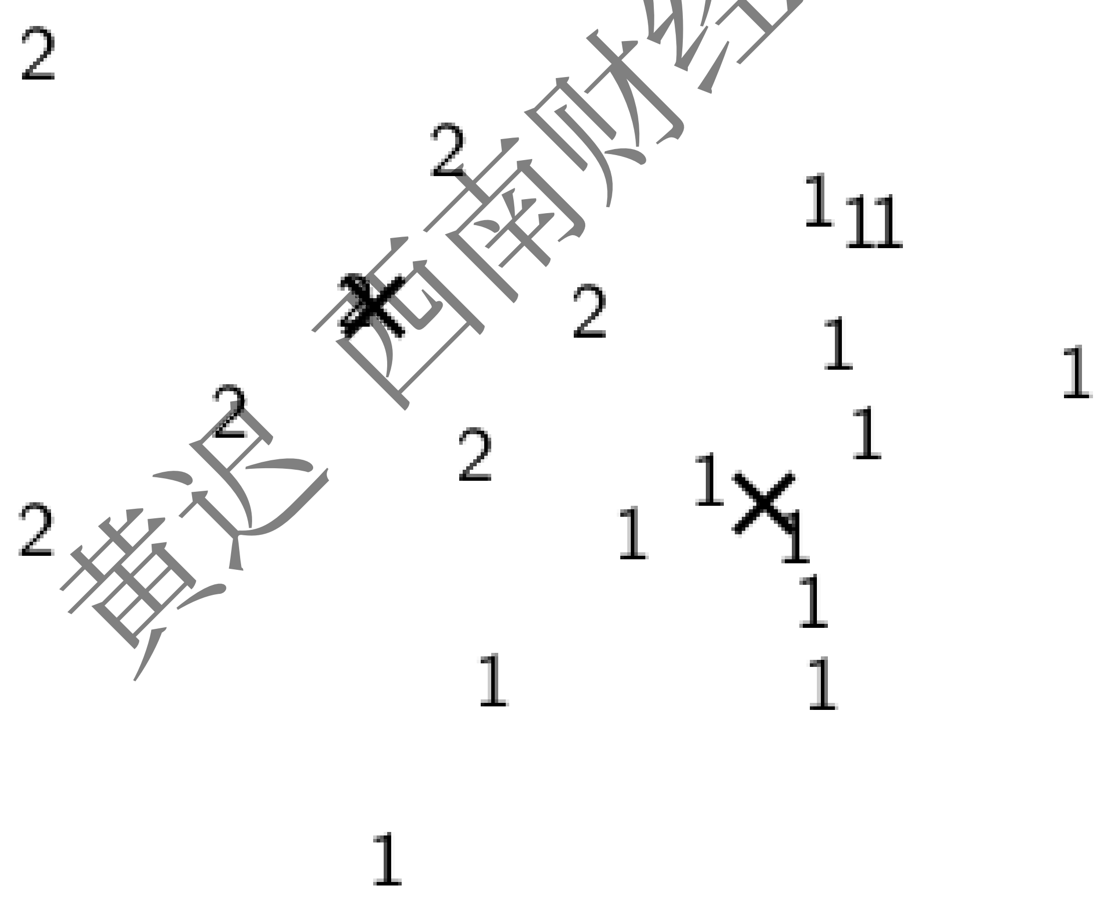
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：重新分配(第六轮)



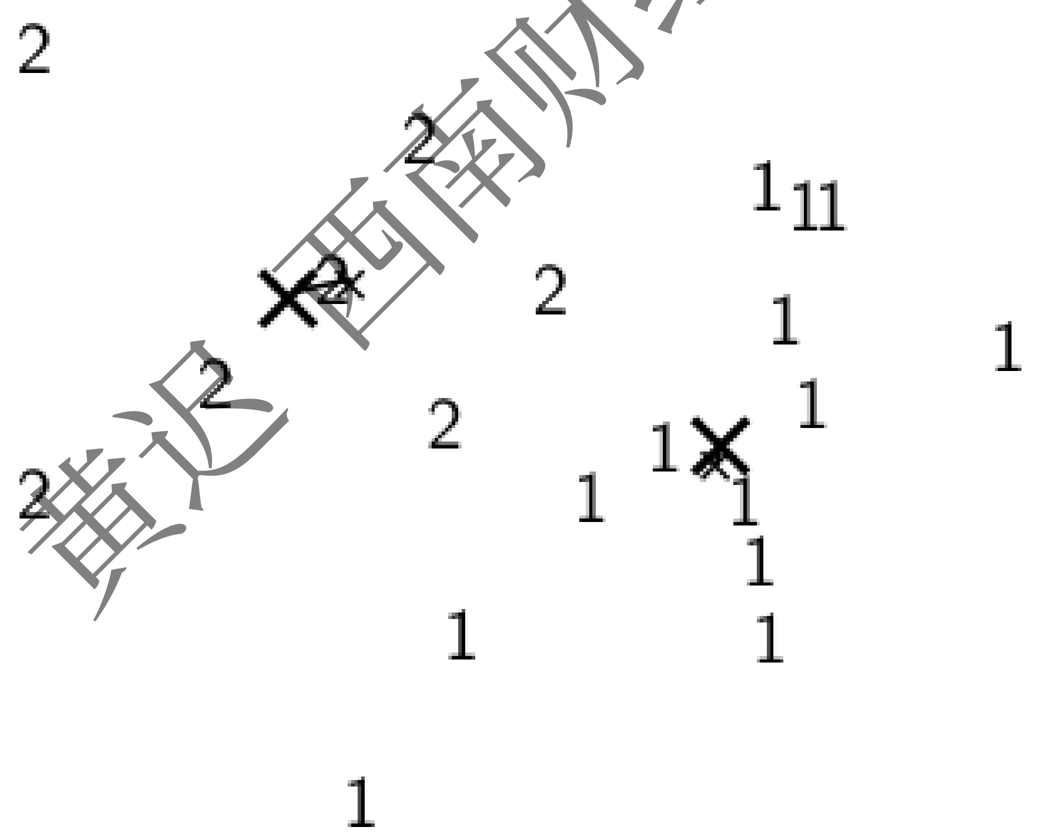
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：分配结果



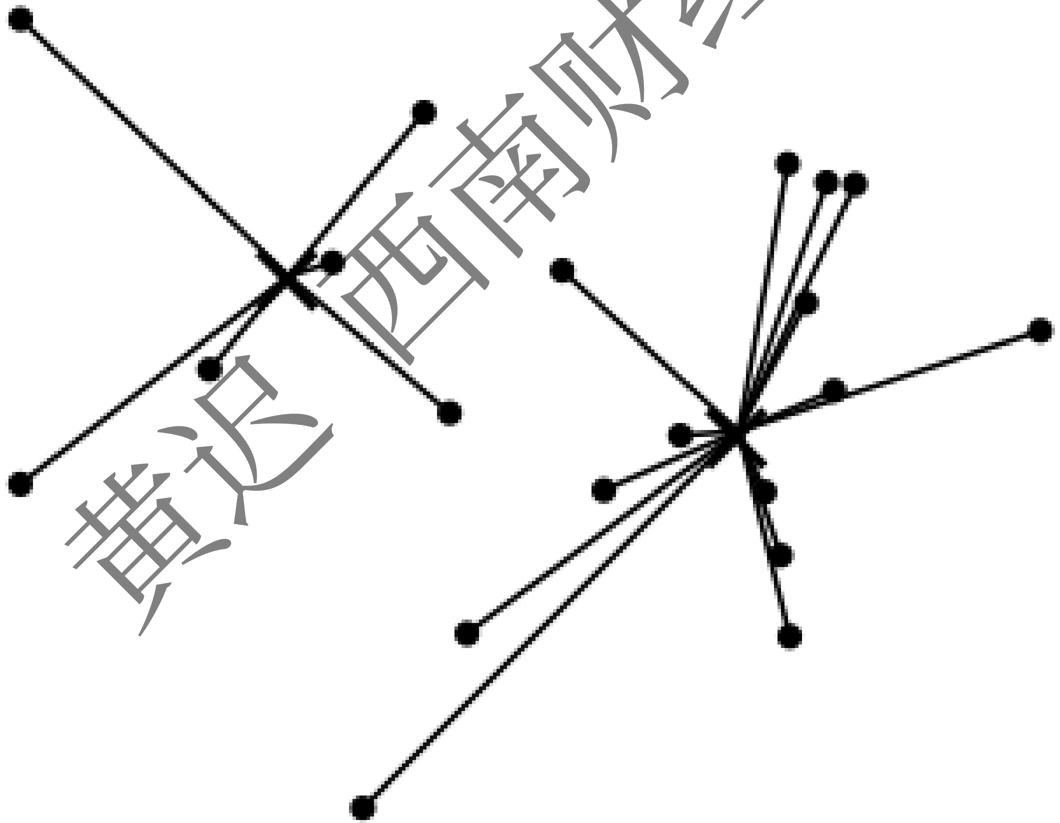
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：重新计算类中心



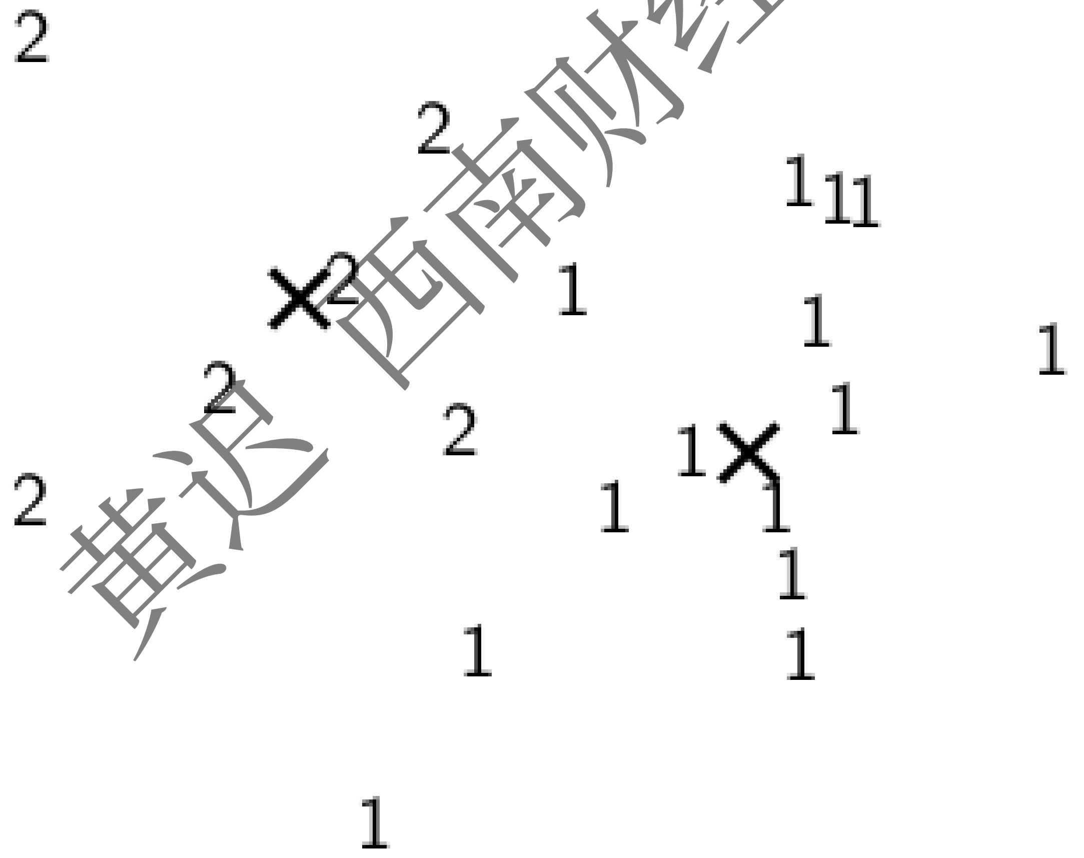
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：重新分配(第七轮)

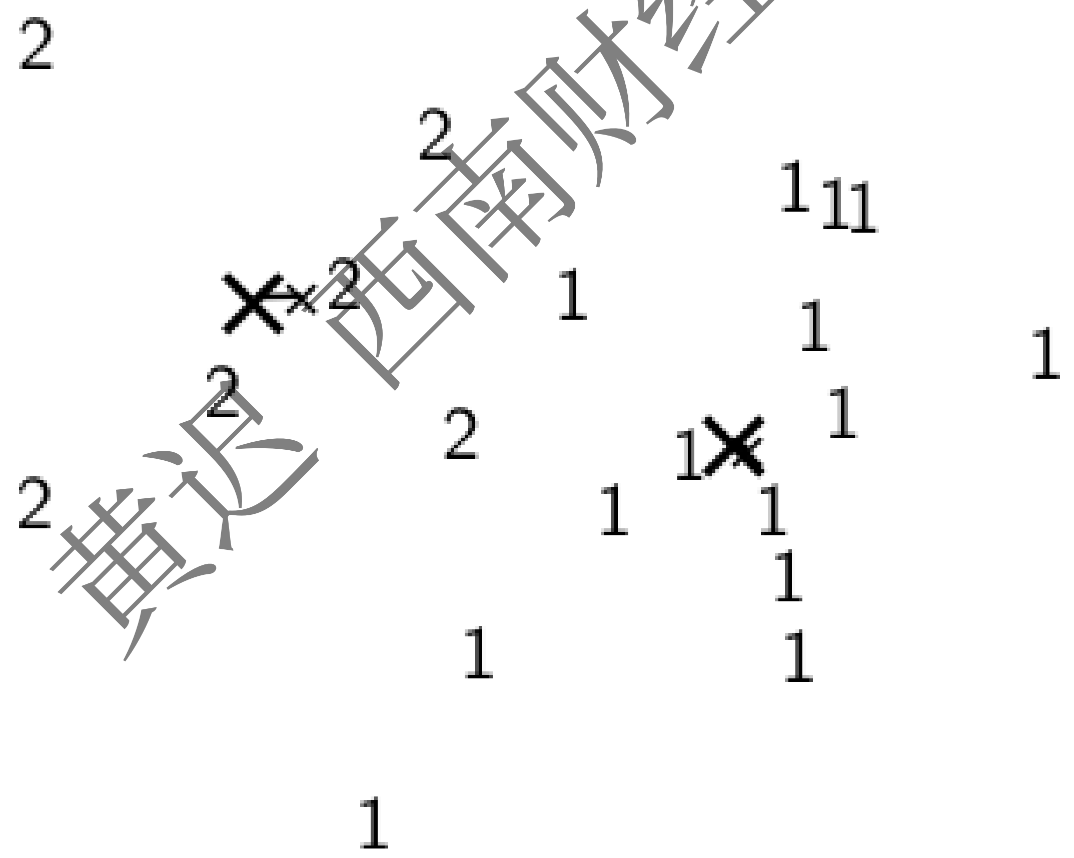


黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：分配结果

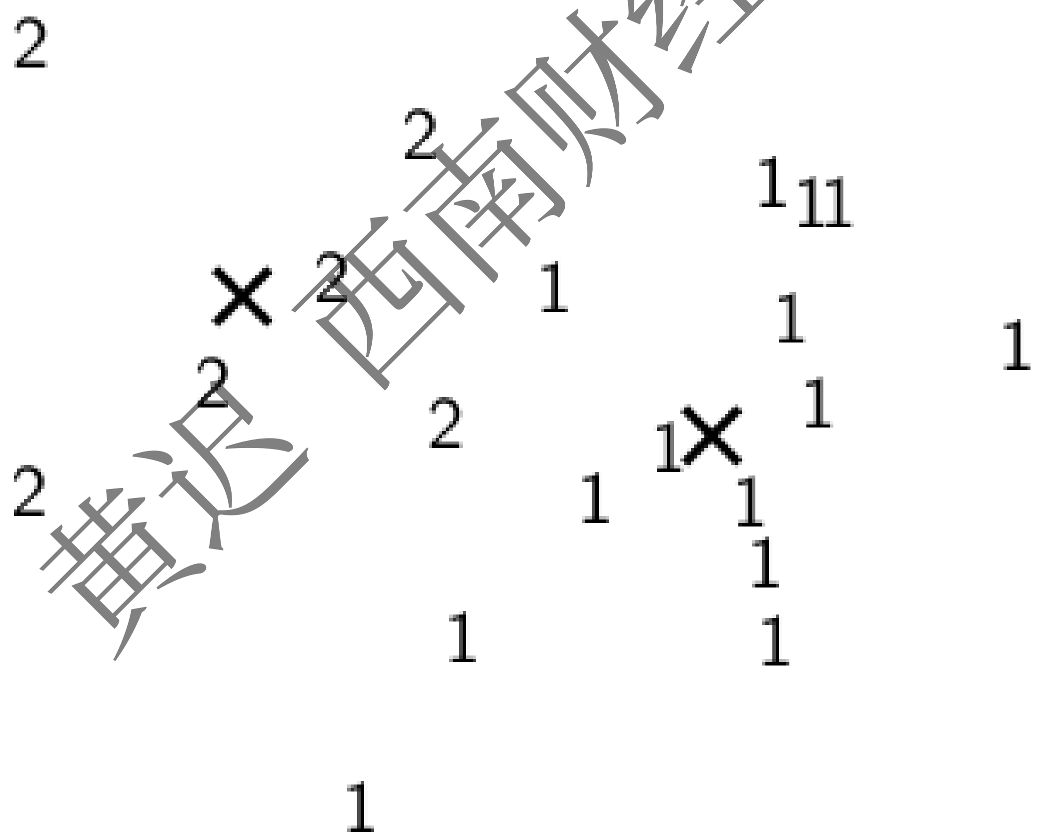


例子：重新计算类中心



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例子：类中心和分配结果最终收敛



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

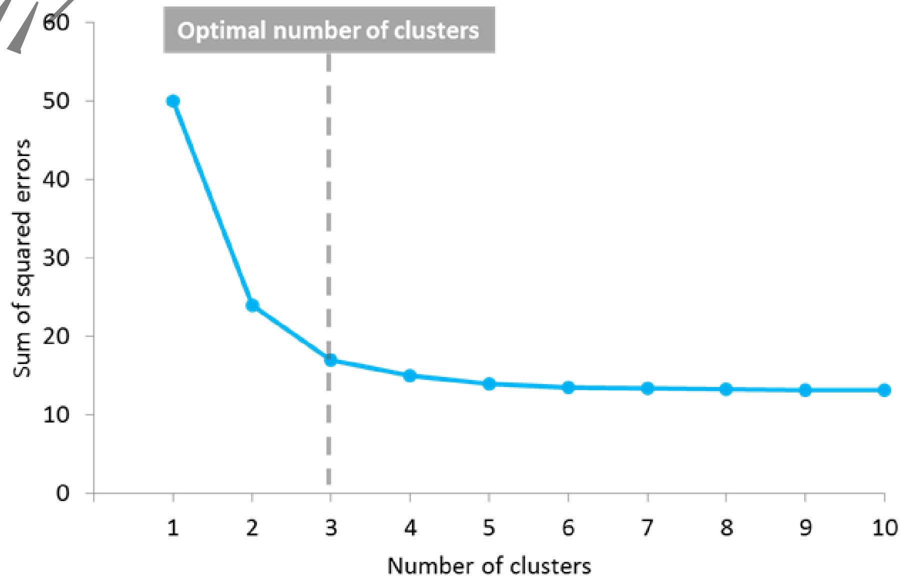
算法说明

- 对于给定的初值（样本中心），（经过有限次迭代）算法必然收敛，即聚类结果是确定的
- 算法属于启发式算法，不能保证收敛到全局最优，初值选取会直接影响聚类结果
- 可以用层次聚类得到k类，求得中心作为初值

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

类别数的选择

- 类别数 k 是超参数，需要预先给定
  - 可以根据不同 k 值的聚类结果，选择最优值
  - 类别数小时，平均直径会增大
  - 类别数大时，平均直径会减小
  - 手肘原则



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

算法应用：图像处理

Original image (96,615 colors)

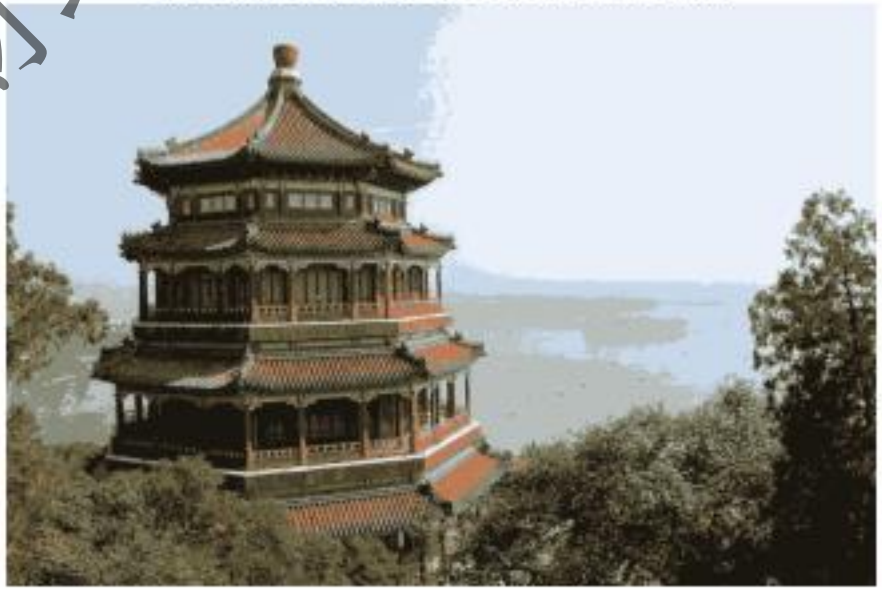
Quantized image (64 colors, K-Means)

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

算法应用：图像处理

Original image (96,615 colors)

Quantized image (8 colors, K-Means)

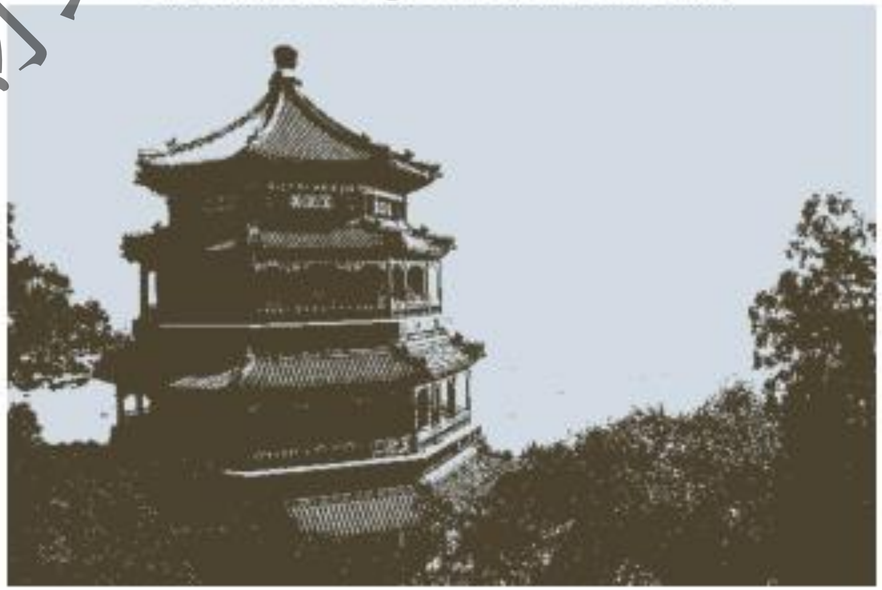


黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

算法应用：图像处理

Original image (96,615 colors)

Quantized image (2 colors, K-Means)



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

算法优缺点

- 优点：简单，易于理解和实现；收敛快，一般仅需5-10次迭代
- 缺点：对 k 值选取对结果有很大的影响

  对初值敏感，不同的初始中心得到的聚类结果可能完全不同

  对于不凸的数据集效果不好

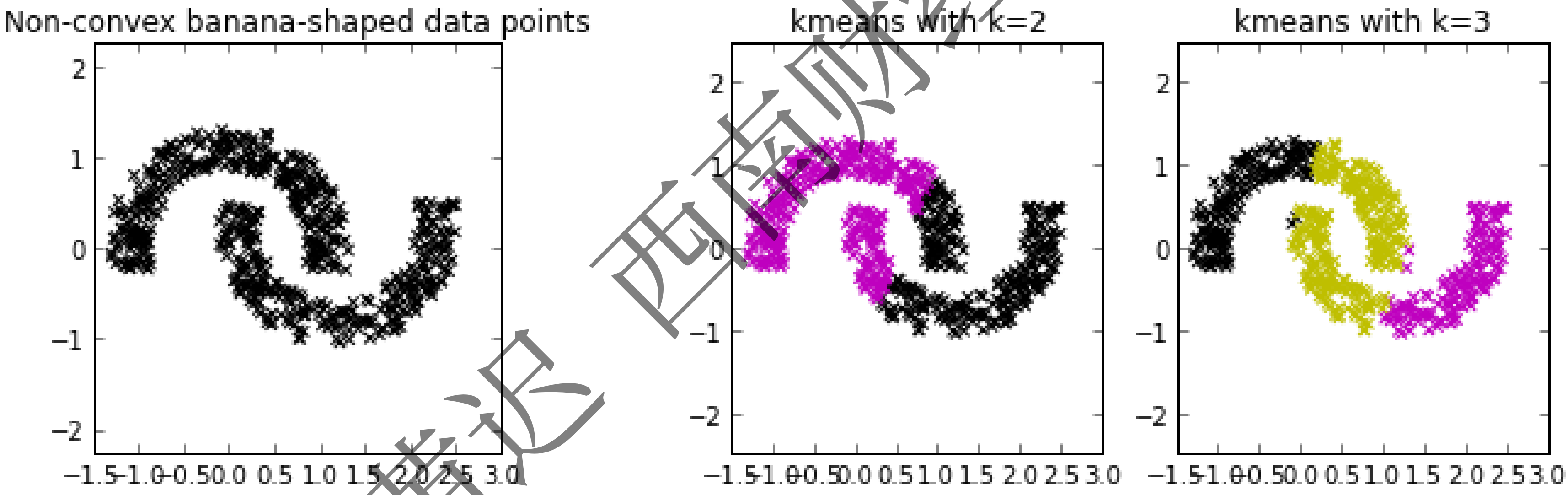
  对噪点过于敏感，因为算法基于均值

  结果不一定是全局最优，只能保证局部最优

  对非球型簇、不同尺寸、不同密度的簇分组效果不好

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

算法在非凸数据集的表现



黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

解决办法：谱聚类