机器学习: Part 2

- 无监督学习: k均值算法
- 从不完全数据中学习: 期望最大化(EM)算法

*Slides based on those of Pascal Poupart

无监督学习

- 无监督学习在没有标签的数据里发现潜在的结构
- 聚类（clustering）是一种典型的无监督学习任务
- 聚类是将输入数据通过分析划分成若干个类簇(clusters)，同一类簇中的数据之间具有某种内在的关系
- 例如，有一堆苹果，要把它们分成两类，可以按照大小分，也可以按照颜色分

聚类	分类
将输入数据根据某种内在关系划分成类簇	将输入数据划分为预先定义的类
类簇及其个数是事先未知的	类及其个数是事先已知的
没有训练数据	有训练数据

- 硬聚类：每个样例被确定性地划分到某个类簇中
- 软聚类：每个样例按照每种概率分布划分到类簇中

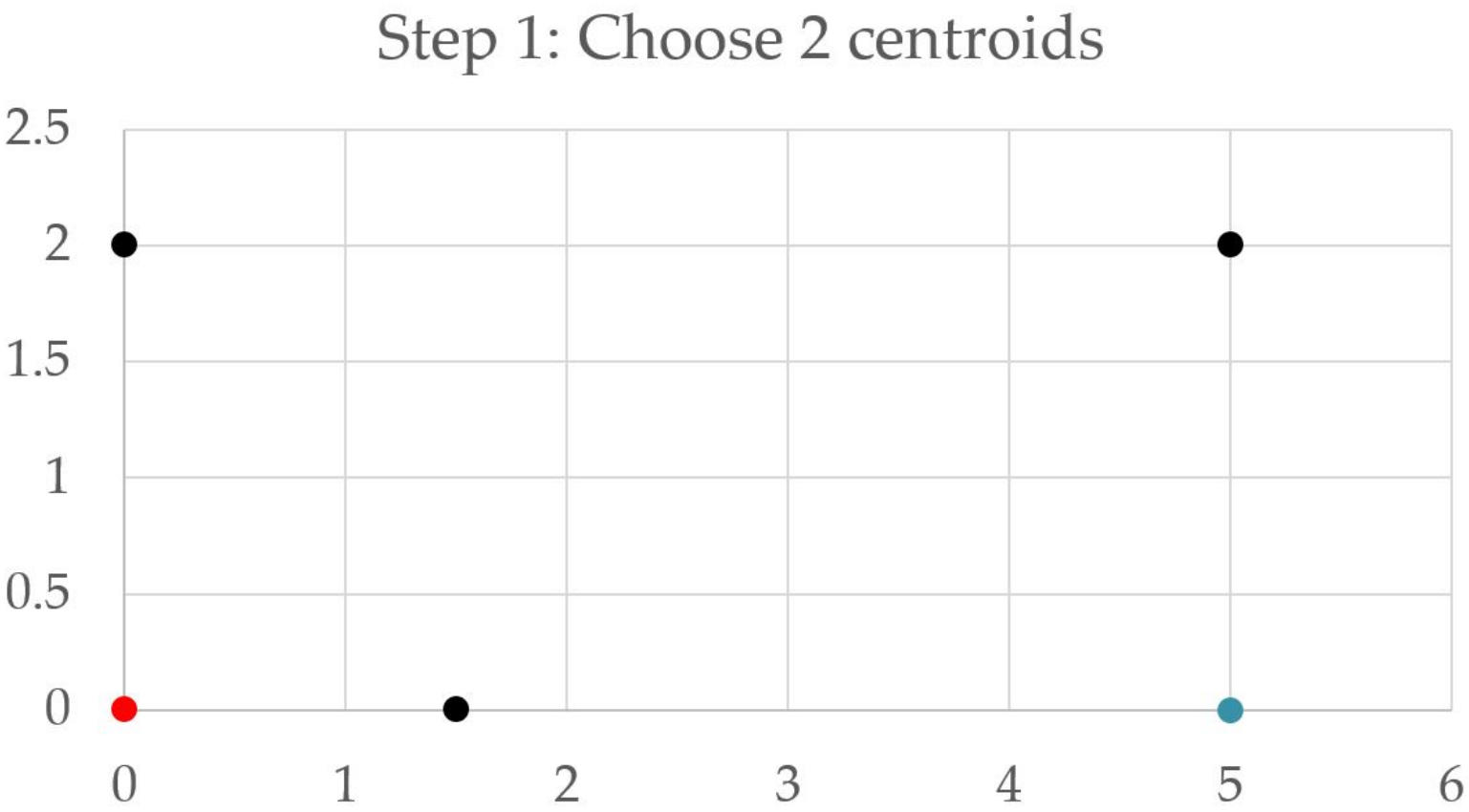
k均值算法

用于硬聚类

1. 选择k个重心（centroid）。
2. 寻找最近的重心并且更新聚类分配。将每个数据点都分配给离它最近的重心的聚类。距离的度量通常是欧式距离。
3. 将重心移动到它们的聚类的中心。每个聚类的重心的新位置是通过计算该聚类中所有数据点的平均位置得到的。
4. 重复第2和3步，直到每次迭代时重心的位置不再显著变化（即直到该算法收敛）。

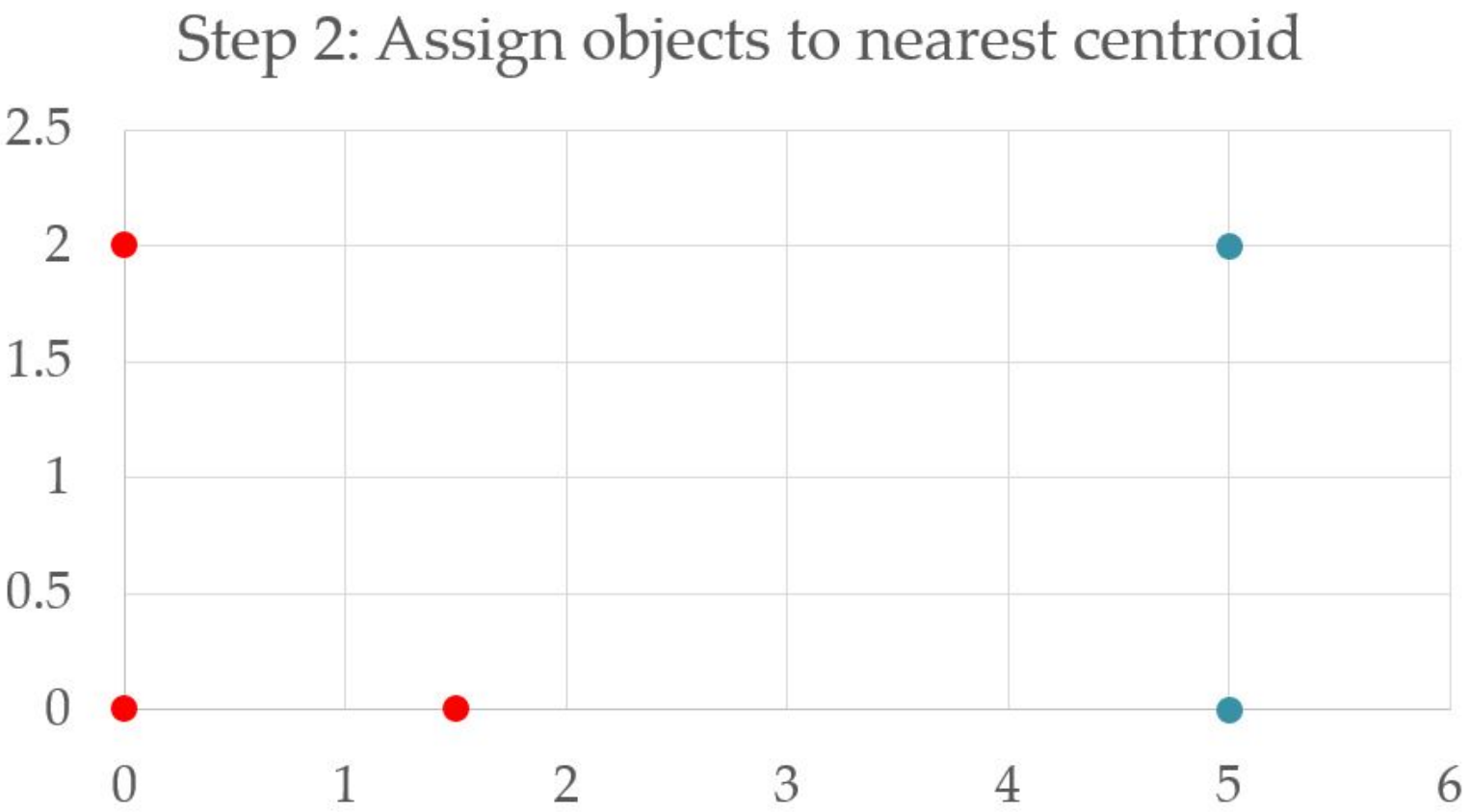
k-Means

x1 = (0, 2), x2 = (0, 0), x3 = (1.5, 0), x4 = (5, 0), x5 = (5, 2); $k = 2$



k-Means

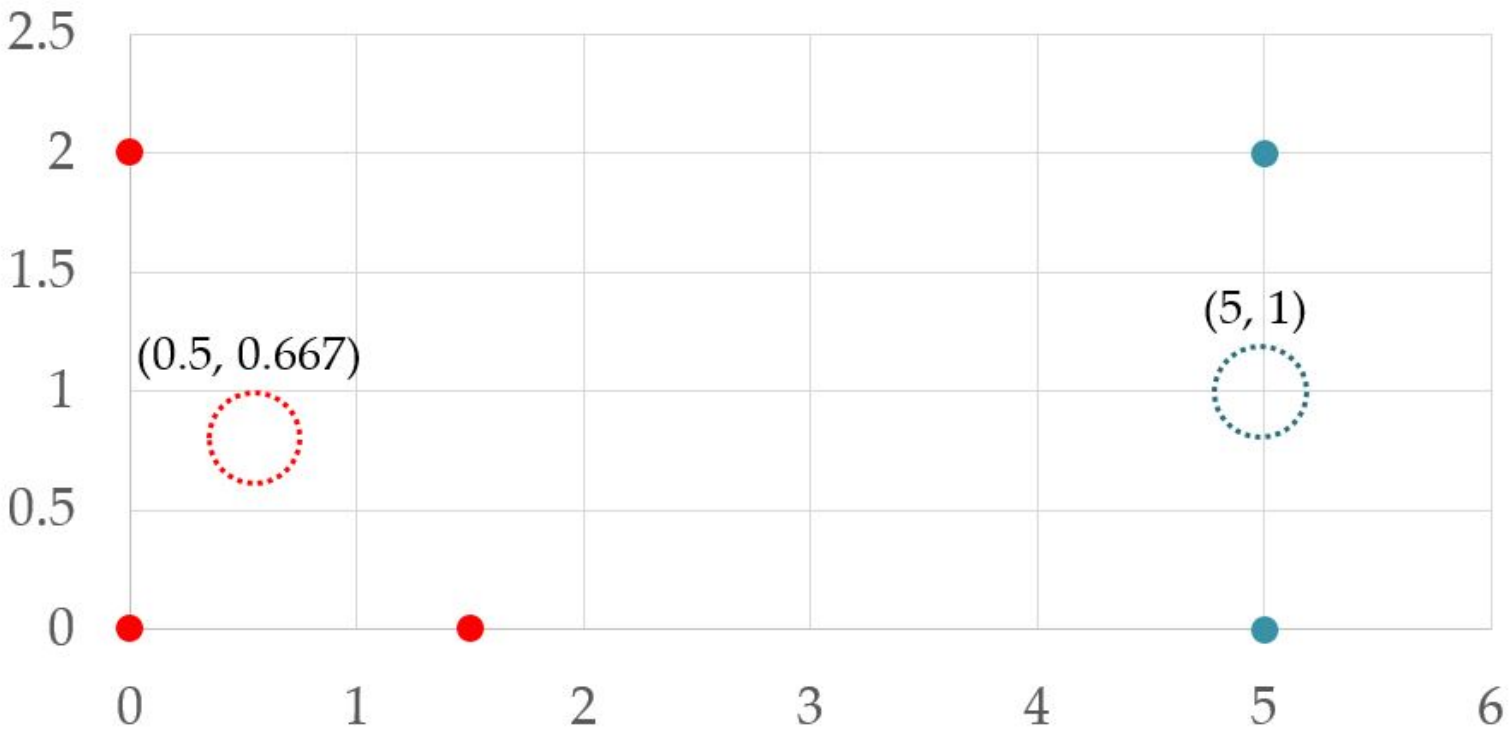
x1 = (0, 2), x2 = (0, 0), x3 = (1.5, 0), x4 = (5, 0), x5 = (5, 2); $k = 2$



k-Means

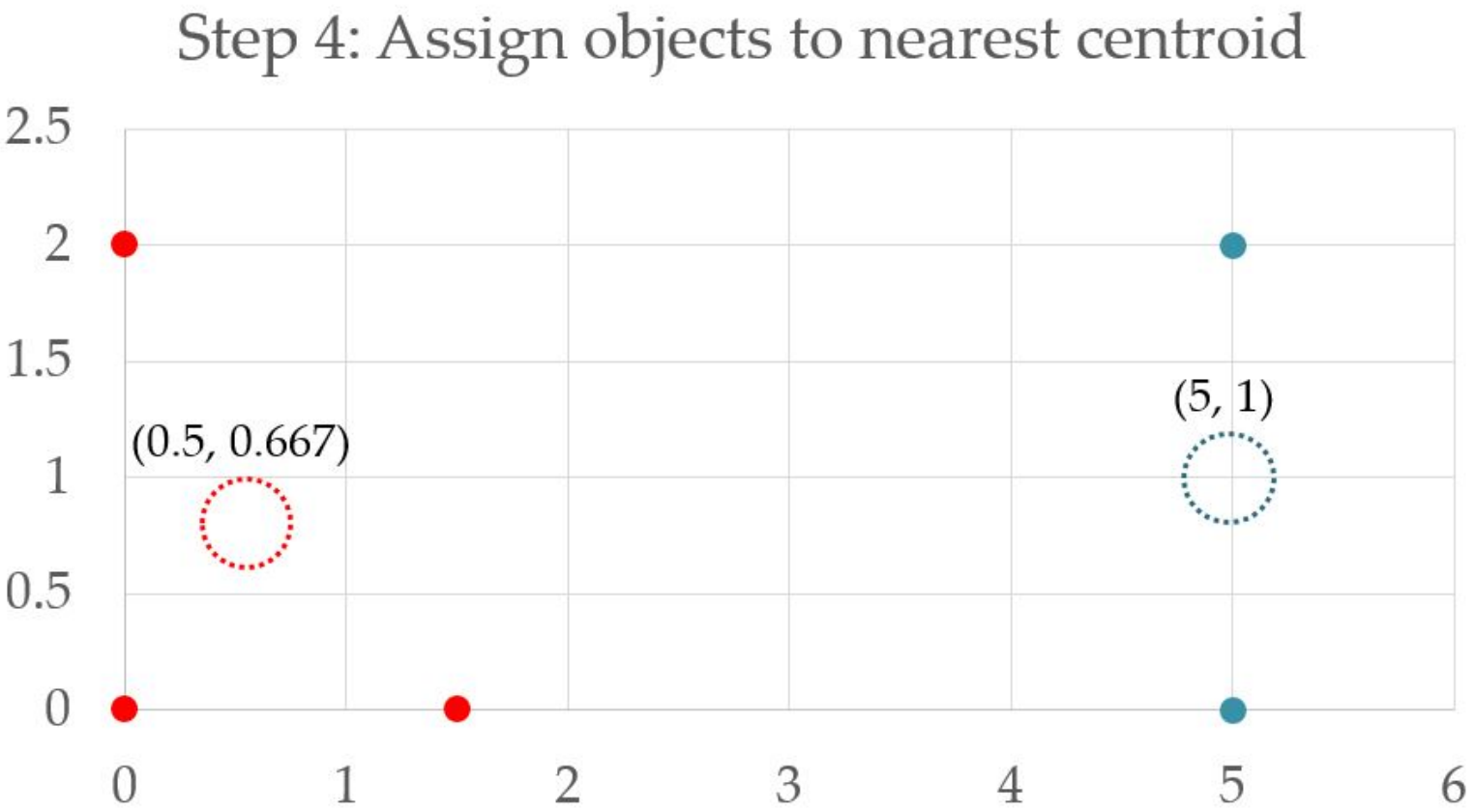
$x1 = (0, 2), x2 = (0, 0), x3 = (1.5, 0), x4 = (5, 0), x5 = (5, 2); \quad k = 2$

Step 3: Re-compute centroids



k-Means

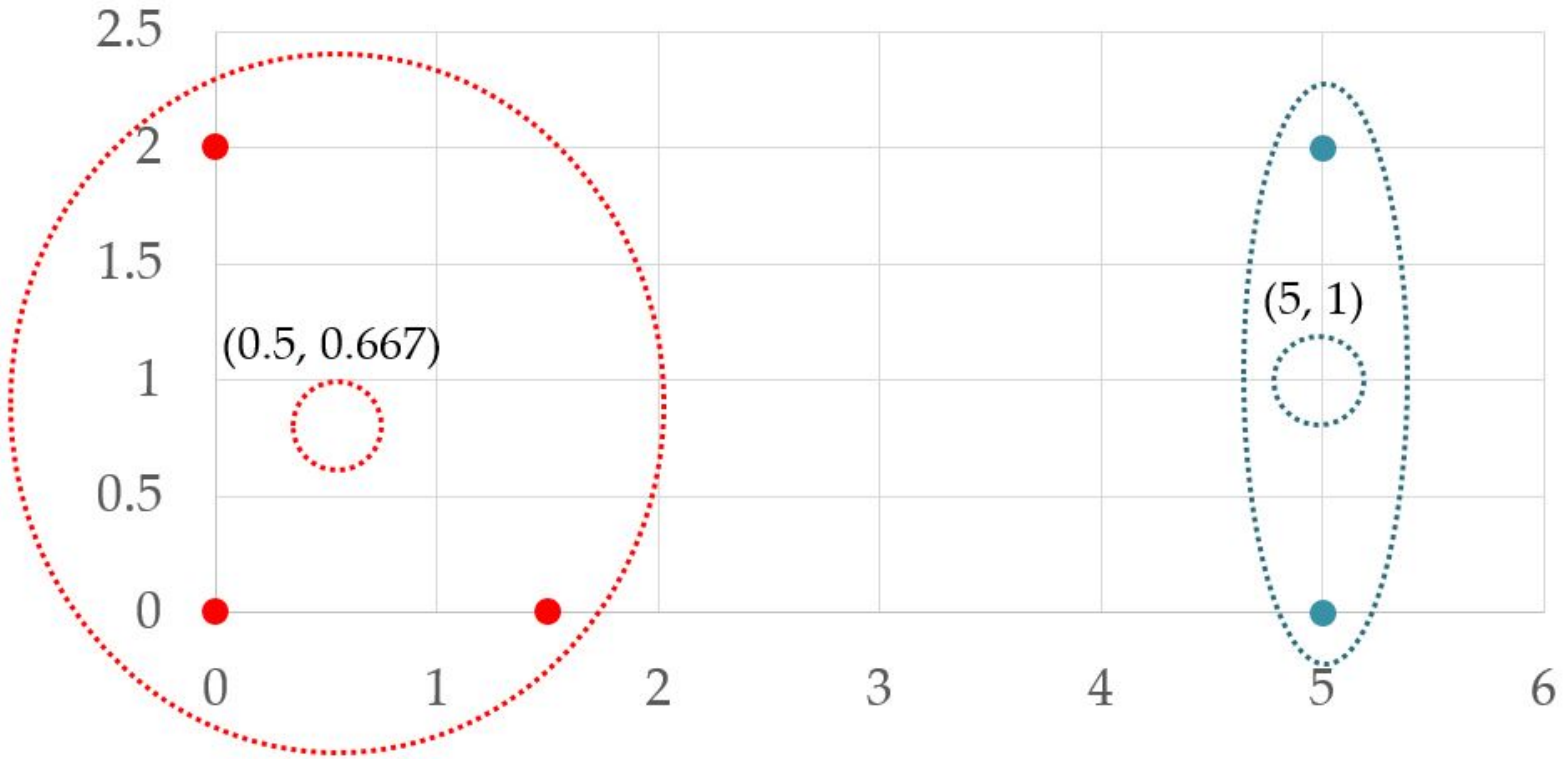
x1 = (0, 2), x2 = (0, 0), x3 = (1.5, 0), x4 = (5, 0), x5 = (5, 2); $k = 2$



k-Means

$x1 = (0, 2), x2 = (0, 0), x3 = (1.5, 0), x4 = (5, 0), x5 = (5, 2);\ \ k = 2$

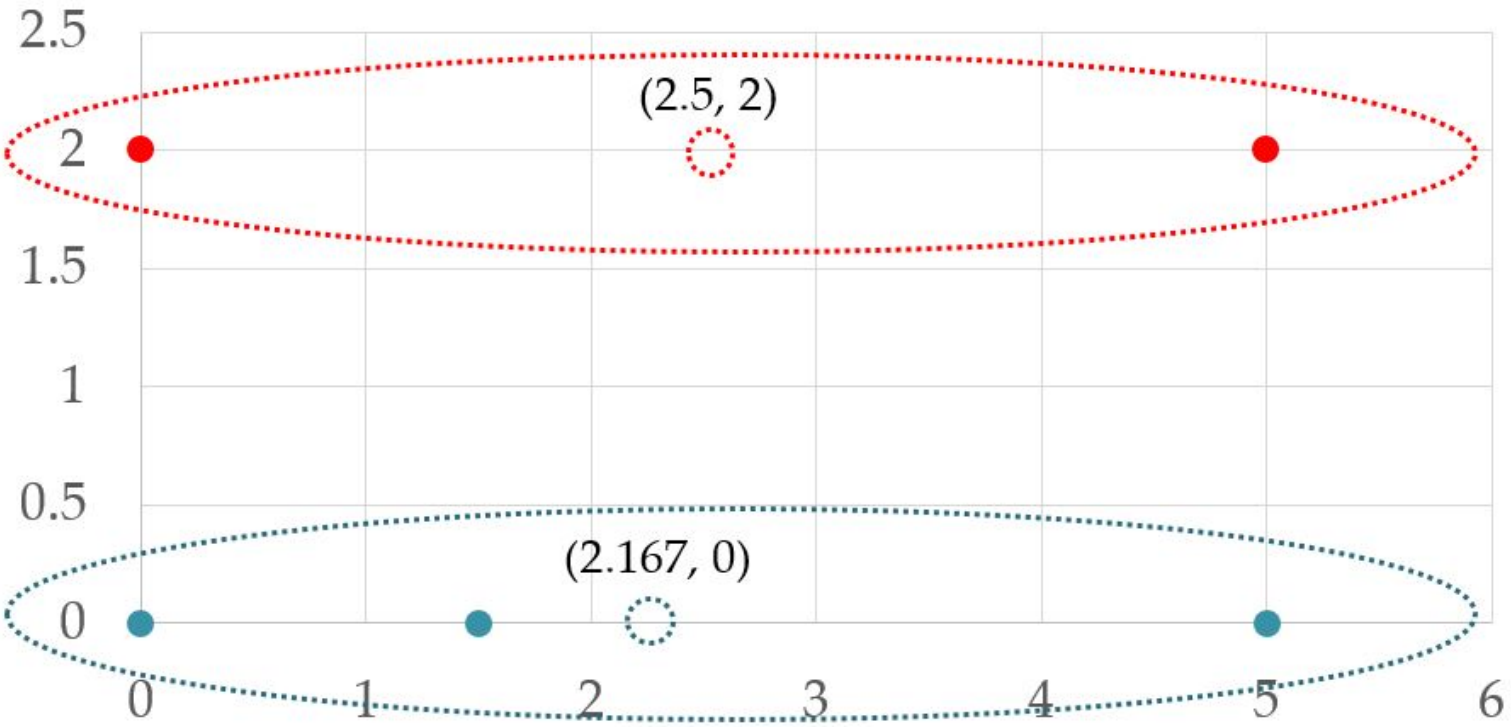
Step 5: Converged



k-Means

x1 = (0, 2), x2 = (0, 0), x3 = (1.5, 0), x4 = (5, 0), x5 = (5, 2); $k = 2$

Another converged solution



如何选择初始重心：k-Means^{++}

初始簇中心的选择会严重影响到聚类算法的最终结果

基本思想：初始的聚类中心之间的相互距离要尽可能的远

1. 从数据集$\mathcal{X}$中随机（均匀分布）选取一个样本点作为第一个初始聚类中心
2. 计算每个样本x与当前已有聚类中心之间的最短距离$D(x)$，然后计算每个样本点被选为下一个聚类中心的概率$P(x)$

$$P(x) = \frac{D(x)^2}{\sum_{x \in \mathcal{X}} D(x)^2}$$

3. 以概率$P(x)$选择x作为一个新的簇中心；
4. 重复第2-3步，直到选择出k个聚类中心

误差平方和SSE(sum of the squared errors)

- 如何衡量聚类效果的好坏？

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} |x - m_i|^2,$$

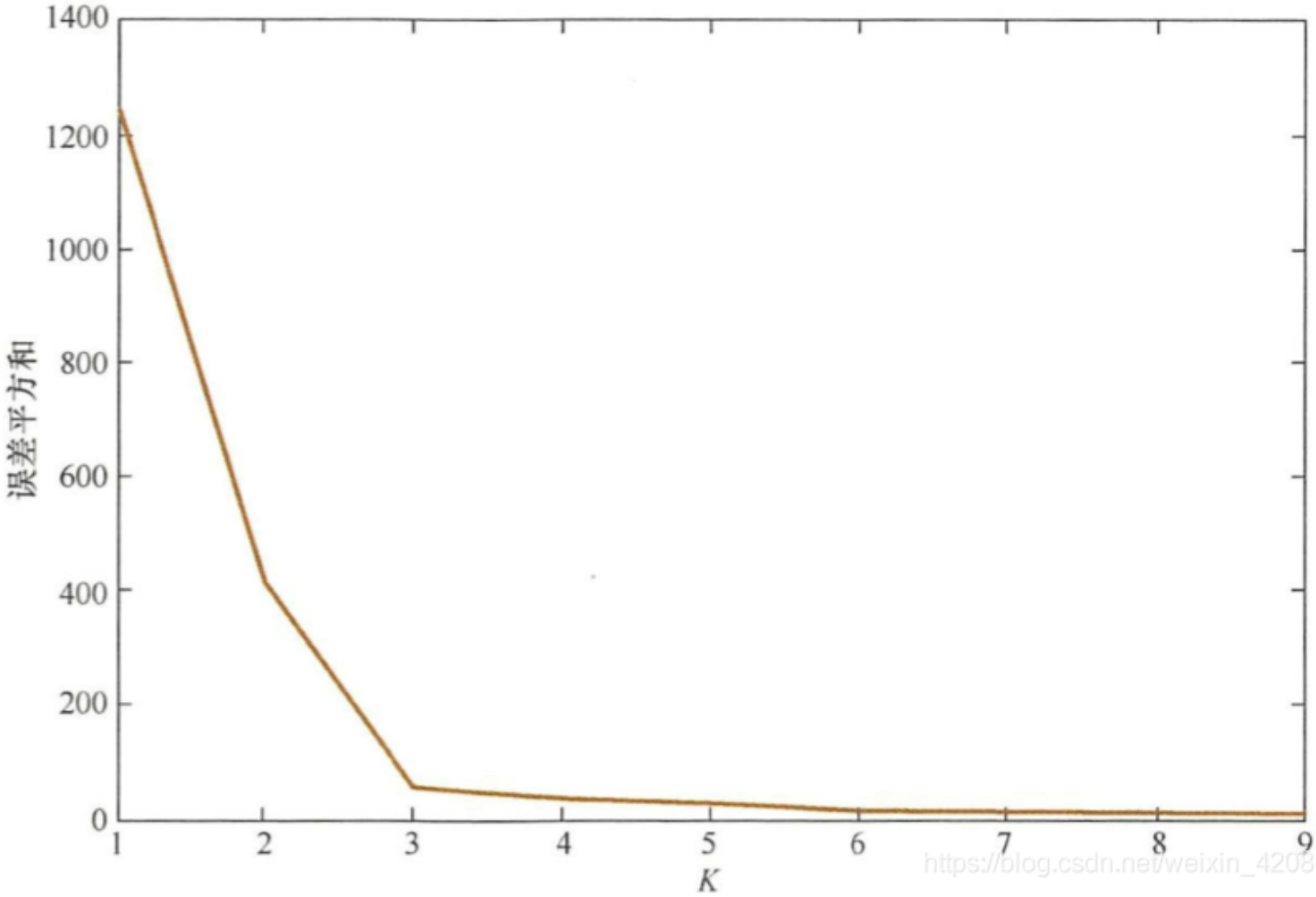
其中C_i是第i个簇，m_i是C_i的重心，$|x - m_i|$是x与m_i的距离

- SSE是所有样本的聚类误差

k值的选择：手肘法

- 随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。
- 当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，
- 而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，
- 也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的真实聚类数。

k值的选择：手肘法



EM算法概览

- 用于软聚类
- 假设数据是由混合高斯模型生成的
- 随机初始化模型参数，重复以下两步直到收敛:
  - E步基于给定的模型参数对数据进行软分类
  - M步从数据中作极大似然学习以更新模型参数

极大似然参数学习: 连续模型

- 考虑一个非常简单的情形：单变量高斯分布

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

- 对数似然性为

$$L = \sum_{j=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_j-\mu)^2}{2\sigma^2}} = N(-\log\sqrt{2\pi} - \log\sigma) - \sum_{j=1}^{N} \frac{(x_j-\mu)^2}{2\sigma^2} .$$

- 令偏导为0，得到

$$\frac{\partial L}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{j=1}^{N}(x_j-\mu) = 0 \quad \Rightarrow \quad \mu = \frac{\sum_j x_j}{N}$$

$$\frac{\partial L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{j=1}^{N}(x_j-\mu)^2 = 0 \quad \Rightarrow \quad \sigma = \sqrt{\frac{\sum_j (x_j-\mu)^2}{N}} .$$

- 因而均值的极大似然值为样本均值，标准差的极大似然值为样本标准差
- 结果与“常识”一致

高斯混合模型

- 假设数据是由一个混合分布P生成的.
- 这个分布有k个成分，其中每个成分本身是一个分布
- 一个数据点是如下生成的：首先选择一个成分，然后生成那个成分的一个样本
- 混合分布的定义如下：

$$P(\mathbf{x}) = \sum_{i=1}^{k} P(C = i)P(\mathbf{x}|C = i),$$

其中$\mathbf{x}$是一个数据点的属性值.

- 对于连续数据，多变量高斯分布是成分分布的一个自然选择

高斯混合模型

- 高斯混合模型的参数是:
  - $w_i = P(C = i)$ (每个成分的权重),
  - μ_i (每个成分的均值),
  - Σ_i (每个成分的协方差).

EM算法

随机初始化模型参数，重复以下两步直到收敛:

1. E步
   - 计算数据$\mathbf{x}_j$是由成分i生成的概率

$$p_{ij} = P(C = i|\mathbf{x}_j) = \alpha P(\mathbf{x}_j|C = i)P(C = i),$$

   其中$P(\mathbf{x}_j|C = i)$是第i个高斯分布，$P(C = i) = w_i$
   - 令$n_i = \sum_j p_{ij}$，即当前分配到成分i的数据点的期望数量
2. M步：计算新的均值，协方差，和权重

$$\boldsymbol{\mu}_i \leftarrow \sum_j p_{ij}\mathbf{x}_j/n_i$$

$$\boldsymbol{\Sigma}_i \leftarrow \sum_j p_{ij}(\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^\top/n_i$$

$$w_i \leftarrow n_i/N$$

其中N 是数据点的总数量.

EM步

- E步，或期望步，可以看作是计算隐变量Z_{ij}的期望值p_{ij}，其中Z_{ij}为1如果数据$\mathbf{x}_j$是由成分i生成的否则为0
- M步，或极大步，在给定隐变量的期望值的情况下，计算新的参数值以极大化数据的对数似然性