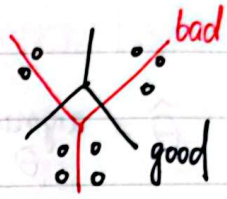
8 Expectation Maximization (EM)

Unsupervised Learning: Kmeans & Gaussian Mixture Models (GMM)

考虑下述任务：

Clustering (informal goals): Goal: data 为unlabeled, 将它们分成数组，每组中点相似。



这个任务极为常用！

问题在于：如何建模出 partitions 的好坏？

Input: unlabeled data: $D = \{x^{(i)}\}_{i=1}^{N}$, $x^{(i)} \in R^M$

⇒ cluster centers: $C = [c_1, \cdots, c_K]$ $c_j \in R^M$.

cluster assignments: $z = [z^{(1)}, z^{(2)}, \cdots, z^{(N)}]$, $z^{(i)} \in \{1, \cdots, K\}$

Objective: $\hat{C} = \underset{C}{\operatorname{argmin}} \sum_{i=1}^{N} \min_j \|x^{(i)} - c_j\|_2^2$

$= \underset{C}{\operatorname{argmin}} \sum_{i=1}^{N} \min_{z^{(i)}} \|x^{(i)} - c_{z^{(i)}}\|_2^2$

$\Leftrightarrow \hat{C}, \hat{z} = \underset{C, z}{\operatorname{argmin}} \sum_{i=1}^{N} \|x^{(i)} - c_{z^{(i)}}\|_2^2$ $\quad J(C, z)$.

Algorithm: 1) Initialize $C = \{c_1, \cdots, c_K\}$ ← * 随机K个样本作初始均值向量

2) Repeat until Convergence:

a) for i in $\{1, \cdots, N\}$.

$z^{(i)} \leftarrow \operatorname{argmin}_j (\|x^{(i)} - c_j\|_2)^2$ ← 每个 x_i 选最近 c_j

b) for j in $\{1, \cdots, k\}$.

$c_j \leftarrow \underset{\uparrow}{\operatorname{argmin}} \sum_{i: z^{(i)}=j} (\|x^{(i)} - c_j\|_2)^2$

此优化 i.e.: $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$, $C_i \leftarrow \mu_i$

* 不止随机，也可 Furthest Point Heuristic：先随机选一个，之后每一个选取尽可能离先前选的点们远！（但 outlier 处理是问题）

也可 K-means++：先随机选 c_1，则 $j = 2, \cdots, K$, pick c_j from distribution:

$P(c_j = x^{(i)}) \propto \min_{j' < j} \|x^{(i)} - c_{j'}\|^2$, where $x^{(i)}$ are not picked before