K-means 是 GMM 的一种特例

发表于 2017-10-23 | 分类于 ML | 您是第 位阅读该文章的人类

从一个 GMM 的特定生成模型讨论了 K-means 与 EM 算法的关系。

对一个特定 GMM 模型的讨论

考虑到在非监督学习 K-means 中，那些不存在的所谓的“标签”也可以被当成“不可见因素”的想法。

本文通过一个特定的 GMM（高斯混合模型，Gaussian Mixture Model）模型讨论了 K-means 算法在该特定模型下可以看做一个 EM 算法。因此可以认为 K-means 算法与 EM 算法在某种先验条件下存在交集。

首先给出以下两个问题的描述：

聚类问题：给定数据点 $x_1, x_2, \cdots, x_N \in \mathbb{R}^m$，给定分类数目 K，求出 K 个类中心 $\mu_1, \mu_2, \cdots, \mu_K$，使得所有点到距离该点最近的类中心的距离的平方和 $\sum_{i=1}^{N} \min_{1 \le k \le K} \|x_i - \mu_k\|_2^2$ 最小。

含隐变量的最大似然问题：给定数据点 $x_1, x_2, \cdots, x_N \in \mathbb{R}^m$，给定分类数目 K，考虑如下生成模型，

$$p(x, z \mid \mu_1, \mu_2, \cdots, \mu_K) \propto \begin{cases} \exp\left(-\|x - \mu_z\|_2^2\right) & \|x - \mu_z\|_2 = \min\limits_{1 \le k \le K} \|x - \mu_k\|_2 \\ 0 & \|x - \mu_z\|_2 > \min\limits_{1 \le k \le K} \|x - \mu_k\|_2 \end{cases}$$

模型中 $z \in \{1, 2, \cdots, K\}$ 为隐变量。

这个模型的意义是：先验的假设真的存在那么 K 个中心，只不过看不到而已。而所有的数据确实也是根据这 K 个中心生成的，且这个生成概率是以其最近的“隐形中心”为中心呈高斯分布，并且不认为（零概率）该数据点是由离其较远的“隐形中心”生成的。（因此这个模型的条件还是很苛刻的。）

结论：用 EM 算法解这个含隐变量的最大似然问题等价于用 K-means 算法解聚类问题。

E 步骤：

1. 计算

$$p(z_i \mid x_i, \{\mu_k\}^{(t)}) \propto \begin{cases} 1 & \|x_i - \mu_{z_i}^{(t)}\|_2 = \min\limits_{1 \le k \le K} \|x_i - \mu_k^{(t)}\|_2 \\ 0 & \|x_i - \mu_{z_i}^{(t)}\|_2 > \min\limits_{1 \le k \le K} \|x_i - \mu_k^{(t)}\|_2 \end{cases}$$

以及

$$p(Z \mid X, \{\mu_k\}^{(t)}) = \prod_{i=1}^{N} p(z_i \mid x_i, \{\mu_k\}^{(t)})$$

这里使用正比于符号是考虑离数据点最近的中心可能不止一个。

下面为了简单只讨论只有一个中心，且中心编号为 y_i 的情况。

2. 计算目标函数

$$\begin{aligned} Q(\{\mu_k\} \mid \{\mu_k\}^{(t)}) &= E_{Z \mid X, \{\mu_k\}^{(t)}} [\log p(X, Z \mid \{\mu_k\})] \\ &= \sum_{i=1}^{N} E_{z_i \mid x_i, \{\mu_k\}^{(t)}} [\log p(x_i, z_i \mid \{\mu_k\})] \\ &= \sum_{i=1}^{N} \log p(x_i, z_i = y_i \mid \{\mu_k\}) \\ &= \text{const} - \sum_{i=1}^{N} \|x_i - \mu_{y_i}\|_2^2 \end{aligned}$$

所以最大化目标函数就等价于最小化 $Q'(\{\mu_k\} \mid \{\mu_k\}^{(t)}) = \sum_{i=1}^{N} \|x_i - \mu_{y_i}\|_2^2$。

M 步骤：

找寻 $\{\mu_k\}$ 使得 Q' 最小。

将指标集 $\{1, 2, \cdots, K\}$ 分割为 K 个，$I_k = \{i \mid y_i = k\}$，则

$$Q'(\{\mu_k\} \mid \{\mu_k\}^{(t)}) = \sum_{i=1}^{K} \sum_{i \in I_k} \|x_i - \mu_k\|_2^2$$

由于求和的每一项都是非负的，所以每一个内层求和 $\sum_{i \in I_k} \|x_i - \mu_k\|_2^2$ 都达到最小时，总和最小。

这和 K-means 最小化二范数平方和是一样的。