

Prototype-based clustering

Chong Liu Harbin Institute of Technology

November 12, 2018

Contents

1	前言	2
2	k均值算法	2
3	EM算法	3
4	高斯混合聚类	3
5	EM算法求解高斯混合分布模型参数	4
6	参考文献	5

1 前言

原型聚类亦称“基于原型的聚类”，此类算法假设聚类结构能通过一组原型刻画，在现实聚类任务中极为常用。通常情况下，算法先对原型初始化，然后对原型进行迭代更新求解。

2 k均值算法

给定样本集 $D = \{x_1, x_2, \dots, x_m\}$ ，“k均值”(k-means)算法针对聚类所得簇划分 $C = \{C_1, C_2, \dots, C_k\}$ 最小化平方误差

$$E = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|_2^2 \quad (1)$$

其中 $\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$ 是簇 C_i 的均值向量。直观看来，式(1)在一定程度上刻画了簇内样本围绕簇均值向量的紧密程度， E 值越小则簇内样本相似度越高。

最小化式(1)并不容易，找到它的最优解需考察样本集 D 所有可能的簇划分，这是一个NP难的问题[Aloise et al., 2009]，因此，k均值算法采用了贪心策略，通过迭代优化来近似求解式(1)。算法流程如算法1所示，其中第1行对均值向量进行初始化，在第4—8行与第9—16行依次对当前簇划分及均值向量迭代更新，若迭代更新后聚类结果保持不变，则在第18行将当前簇划分结果返回。

Algorithm 1: k均值算法

Input: 样本集 $D = \{x_1, x_2, \dots, x_m\}$ 和聚类簇数 k

- 1 从 D 中随机选择 k 个样本作为初始均值向量 $\{\mu_1, \mu_2, \dots, \mu_k\}$
- 2 **repeat**
- 3 令 $C_i = \phi (1 \leq i \leq k)$
- 4 **for** $j = 1, \dots, m$ **do**
- 5 计算样本 x_j 与各均值向量 $\mu_i (1 \leq i \leq k)$ 的距离: $d_{ji} = \|x_j - \mu_i\|_2$;
- 6 根据距离最近的均值向量确定 x_j 的簇标
- 7 记: $\lambda_j = \arg \min_{i \in \{1, 2, \dots, k\}} d_{ji}$;
- 8 将样本 x 划入相应的簇: $C_{\lambda_j} \cup x_j$;
- 9 **end**
- 10 **for** $i = 1, 2, \dots, k$ **do**
- 11 计算新均值向量: $\mu'_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$;
- 12 **if** $\mu'_i \neq \mu_i$ **then**
- 13 将当前均值向量 μ_i 更新为 μ'_i
- 14 **else**
- 15 保持当前均值向量不变
- 16 **end**
- 17 **end**
- 18 **until** 当前均值向量均未更新;
- 19 **Output:** 簇划分 $C = \{C_1, C_2, \dots, C_k\}$

3 EM算法

未观测变量的学名是“隐变量”(latent variable)。令 X 表示已观测变量集， Z 表示隐变量集， Θ 表示模型参数。若欲对 Θ 做极大似然估计，则应最大化对数似然

$$LL(\Theta|X, Z) = \ln P(X, Z|\Theta) \quad (2)$$

然而由于 Z 是隐变量，上式无法直接求解。此时我们可通过对 Z 计算期望，来最大化已观测数据的对数“边际似然”(marginal likelihood)

$$LL(\Theta|X) = \ln P(X|\Theta) = \ln \sum_Z P(X, Z|\Theta) \quad (3)$$

EM(Expectation-Maximization)算法[Dempster et al.,1977]是常用的估计隐变量的利器，它是一种迭代式的方法，其基本想法是：若参数 Θ 已知，则可根据训练数据推断出最优隐变量 Z 的值(E步)；反之，若 Z 的值已知，则可以方便的对参数 Θ 做极大似然估计(M步)

进一步，若我们不是取 Z 的期望，而是基于 Θ^t 计算隐变量 Z 的概率分布 $P(Z|X, \Theta^t)$ ，则EM算法的两个步骤是：

1. E步(Expectation):以当前参数 Θ^t 推断隐变量分布 $P(Z|X, \Theta^t)$ ，并计算对数似然 $LL(\Theta|X, Z)$ 关于 Z 的期望

$$Q(\Theta|\Theta^t) = \mathbb{E}_{Z|X, \Theta^t} LL(\Theta|X, Z) \quad (4)$$

2. M步(Maximization):寻找参数最大化期望似然，即

$$\Theta^{t+1} = \arg \max_{\Theta} Q(\Theta|\Theta^t) \quad (5)$$

4 高斯混合聚类

与 k 均值用原型向量来刻画聚类结构不同，高斯混合(Mixture-of-Gaussian)聚类采用概论模型来表达聚类原型。

Definition 1 (多元) 高斯分布

对 n 维样本空间 χ 中的随机向量 x ，若 x 服从高斯分布，其概率密度函数为

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \quad (6)$$

其中 μ 是 n 维均值向量， Σ 是 $n \times n$ 的协方差矩阵。由式(2)可看出，高斯分布完全由均值向量 μ 和协方差矩阵 Σ 这两个参数确定。为了明确显示高斯分布与相应参数的依赖关系，将概率密度函数记为 $p(x|\mu, \Sigma)$ 。

我们可定义

Definition 2 高斯混合分布

$$p_M = \sum_{i=1}^k \alpha_i \cdot p(x|\mu_i, \Sigma_i) \quad (7)$$

该分布共由 k 个混合成分组成，每个混合成分对应一个高斯分布。其中 μ_i 与 Σ_i 是第 i 个高斯混合成分的参数，而 $\alpha_i > 0$ 为相应的“混合系数”(mixture coefficient)， $\sum_{i=1}^k \alpha_i = 1$ 。

假设样本的生成过程由高斯混合分布给出：首先，根据 $\alpha_1, \alpha_2, \dots, \alpha_k$ 定义的先验分布选择高斯混合成分，其中 α_i 为选择第 i 个混合成分的概率；然后，根据被选择的混合成分的概率密度函数进行采样，从而生成相应的样本。若训练集 $D = \{x_1, x_2, \dots, x_m\}$ 由上述过程生成，令随机变量 $z_j \in \{1, 2, \dots, k\}$ 表示生成样本 x_j 的高斯混合成分，其取值未知。显然， z_j 的先验概率 $P(z_j = i)$ 对应于 $\alpha_i (i = 1, 2, \dots, k)$ 。根据贝叶斯定理， z_j 的后验分布对应于

$$p_M(z_j = i|x_j) = \frac{P(z_j = i) \cdot p_M(x_j|z_j = i)}{p_M(x_j)} = \frac{\alpha_i \cdot p(x_j|\mu_i, \Sigma_i)}{\sum_{l=1}^k \alpha_l \cdot p(x_j|\mu_l, \Sigma_l)} \quad (8)$$

换言之， $p_M(z_j = i|x_j)$ 给出了样本 x_j 由第 i 个高斯混合成分生成的后验概率，为方便叙述，将其简记为 $\gamma_{ji} (i = 1, 2, \dots, k)$ 。

当高斯混合分布(7)已知时，高斯混合聚类将把样本集 D 划分为 k 个簇 $C = \{C_1, C_2, \dots, C_k\}$ ，每个样本 x_j 的簇标记 λ_j 如下确定：

$$\lambda_j = \arg \max_{i \in \{1, 2, \dots, k\}} \gamma_{ji} \quad (9)$$

因此，从原型聚类的角度来看，高斯混合聚类是采用概率模型（高斯分布）对原型进行刻画，簇划分则由原型对应后验概率确定。

5 EM算法求解高斯混合分布模型参数

那么，对于式(7)，模型参数 $\{(\alpha_i, \mu_i, \Sigma_i) | 1 \leq i \leq k\}$ 如何求解呢？显然，给定样本集 D ，可采用极大似然估计，即最大化（对数）似然

$$LL(D) = \ln\left(\prod_{j=1}^m p_M(x_j)\right) = \sum_{j=1}^m \ln\left(\sum_{i=1}^k \alpha_i \cdot p(x_j|\mu_i, \Sigma_i)\right) \quad (10)$$

常采用EM算法进行迭代优化求解。

若参数 $\{(\alpha_i, \mu_i, \Sigma_i) | 1 \leq i \leq k\}$ 能使式(10)最大化，有

1. $\mu_i = \frac{\sum_{j=1}^m \gamma_{ji} x_j}{\sum_{j=1}^m \gamma_{ji}}$
2. $\Sigma_i = \frac{\sum_{j=1}^m \gamma_{ji} (x_j - \mu_i')(x_j - \mu_i')^T}{\sum_{j=1}^m \gamma_{ji}}$
3. $\alpha_i = \frac{\sum_{j=1}^m \gamma_{ji}}{m}$

高斯混合聚类算法描述如算法2所示。算法第1行对高斯混合分布的模型参数进行初始化。然后，在第2—12行基于EM算法对模型参数进行迭代更新。若EM算法的停止条件满足，则在第14-17行根据高斯混合分布确定簇划分，在第18行返回最终结果。

Algorithm 2: 高斯混合聚类算法

Input: 样本集 $D = \{x_1, x_2, \dots, x_m\}$ 和高斯混合成分个数 k

- 1 初始化高斯混合分布的模型参数 $\{(\alpha_i, \mu_i, \Sigma_i) | 1 \leq i \leq k\}$
- 2 **repeat**
- 3 **for** $j = 1, \dots, m$ **do**
- 4 根据式(8)计算 x_j 由各混合成分生成的后验概率，
即 $\gamma_{ji} = P_M(z_j = i | x_j) (1 \leq i \leq k)$
- 5 **end**
- 6 **for** $i = 1, 2, \dots, k$ **do**
- 7 计算新均值向量: $\mu'_i = \frac{\sum_{j=1}^m \gamma_{ji} x_j}{\sum_{j=1}^m \gamma_{ji}}$;
- 8 计算新协方差矩阵: $\Sigma'_i = \frac{\sum_{j=1}^m \gamma_{ji} (x_j - \mu'_i)(x_j - \mu'_i)^T}{\sum_{j=1}^m \gamma_{ji}}$;
- 9 计算新混合系数: $\alpha'_i = \frac{\sum_{j=1}^m \gamma_{ji}}{m}$;
- 10 **end**
- 11 将模型参数 $\{(\alpha_i, \mu_i, \Sigma_i) | 1 \leq i \leq k\}$ 更新为 $\{(\alpha'_i, \mu'_i, \Sigma'_i) | 1 \leq i \leq k\}$
- 12 **until** 满足停止条件;
- 13 $C_i = \phi (1 \leq i \leq k)$
- 14 **for** $j = 1, \dots, m$ **do**
- 15 根据式(3)确定 x_j 的簇标记 λ_j ;
- 16 将 x_j 划入相应的簇: $C_{\lambda_j} = C_{\lambda_j} \cup x_j$
- 17 **end**
- Output:** 簇划分 $C = \{C_1, C_2, \dots, C_k\}$

6 参考文献

- [1] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016.