Prototype-based clustering

Chong Liu Harbin Institute of Technology

November 10, 2018

Contents

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

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \left\| x - \mu_i \right\|_2^2 \tag{1}$$

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$D = \{x_1, x_2, \ldots, x_m\}$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$\{\mu_1, \mu_2, \ldots, \mu_k\}$
2 **repeat**
3 令$C_i = \phi (1 \le i \le k)$
4 **for** $j = 1, , \ldots, m$ **do**
5 计算样本x_j与各均值向量$\mu_i (1 \le i \le k)$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
记:$\lambda_j = arg \quad min_{i \in \{1,2,\ldots,k\}} \quad d_{ji}$;
7 将样本x划入相应的簇：$C_{\lambda_j} \bigcup x_j$;
8 **end**
9 **for** $i = 1, 2, \ldots, k$ **do**
10 计算新均值向量:$\mu_i' = \frac{1}{|C_i|} \sum_{x \in C_i} x$;
11 **if** $\mu_i' \neq \mu_i$ **then**
12 将当前均值向量μ_i更新为μ_i'
13 **else**
14 保持当前均值向量不变
15 **end**
16 **end**
17 **until** 当前均值向量均未更新;

Output: 簇划分$C = \{C_1, C_2, \ldots, C_k\}$

3 EM算法

未观测变量的学名式“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

$$LL(\Theta|X,Z) = lnP(X,Z|\Theta) \tag{2}$$

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)

$$LL(\Theta|X) = lnP(X|\Theta) = ln\sum_{Z} P(X,Z|\Theta) \tag{3}$$

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

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

2. M步(Maximization):寻找参数最大化期望似然，即

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

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

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

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

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

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$\alpha_1, \alpha_2, \ldots, \alpha_k$定义的先验分布选择高斯混合成分，其中$\alpha_i$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$D = \{x_1, x_2, \ldots, x_m\}$由上述过程生成，令随机变量$z_j \in \{1, 2, \ldots, k\}$表示生成样本$x_i$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$\alpha_i (i = 1, 2, \ldots, k)$。根据贝叶斯定理，$z_j$的后验分布对应于

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

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$\gamma_{ji} (i = 1, 2, \ldots, k)$。

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$C = \{C_1, C_2, \ldots, C_k\}$，每个样本$x_j$的簇标记$\lambda_j$如下确定:

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

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

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

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$D = \{x_1, x_2, \ldots, x_m\}$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** $j = 1, , \ldots, m$ **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** $i = 1, 2, \ldots, k$ **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** $j = 1, , \ldots, m$ **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：$C_{\lambda_j} = C_{\lambda_j} \bigcup x_j$
17 **end**

Output: 簇划分$C = \{C_1, C_2, \ldots, C_k\}$

6 参考文献

[1] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016.