

Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心



高斯混合模型及求解算法

硕士研究生 王殿元

2019年2月17日

- 基本概念
 - 高斯分布
 - 高斯混合模型
 - 最大似然估计
 - 非凸优化方法
- 算法原理
 - EM算法
- 参考文献

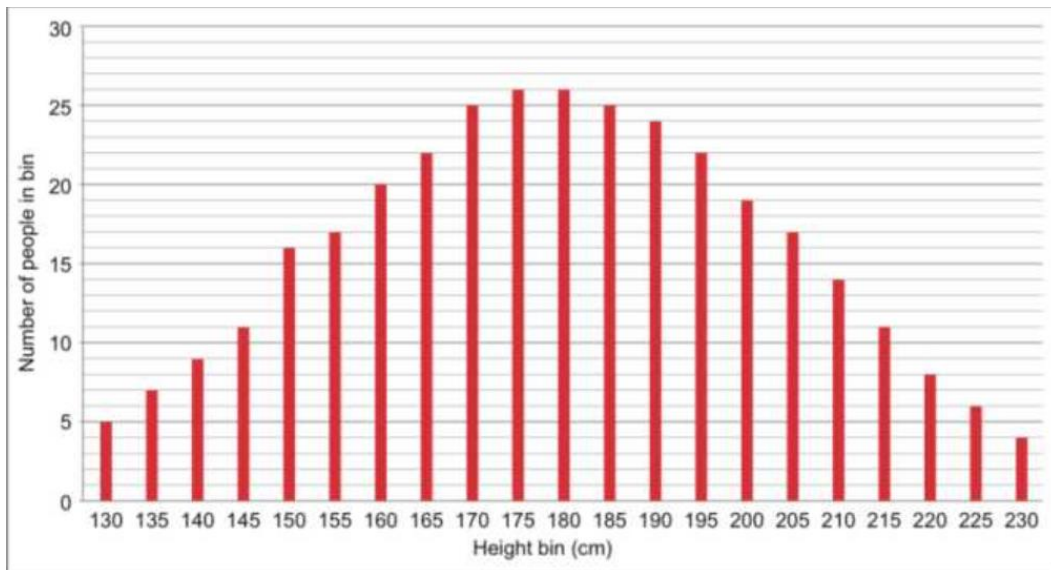
- 预期收获
 - 1. 了解高斯混合模型概念及定义
 - 2. 了解常用的非凸优化方法
 - 3. 理解EM算法原理及在GMM中的应用



基本概念

- 高斯分布也称正态分布，是一种在自然界大量的存在的、最为常见的分布形式，

概率密度函数 $p(x) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



一维身高数据随机采样

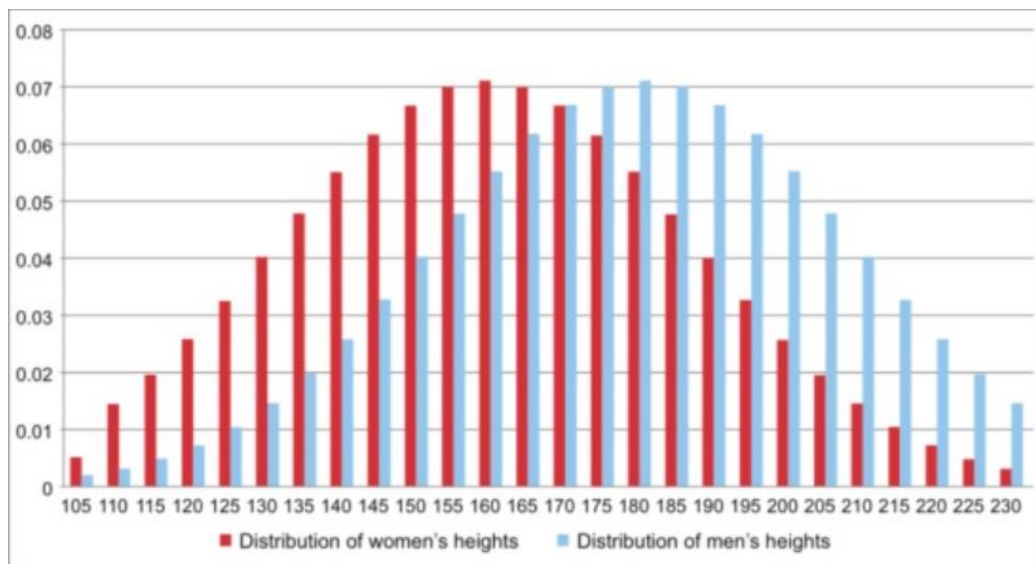
$$X = 334, \mu = 180$$

- 高斯混合模型(Gaussian mixture model, GMM)是对高斯模型进行简单的扩展，使用多个高斯分布的组合来刻画数据分布。

概率密度函数 $p(x) = \sum_{k=1}^K \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$

$$1 = \pi_1 \cdot 1 + \pi_2 \cdot 1 + \dots + \pi_k \cdot 1$$

e.g. $\sum_{k=1}^K \pi_k = 1$



$K=2$ (男、女)

$$\mu_{\text{men}} = 180, \mu_{\text{women}} = 160$$

关于参数 π_k 的意义:

X: 观测变量

Z: 隐变量 (样本属于第几个高斯分布)

Z	C_1	C_2	C_3	$\cdots$	C_{K-1}	C_K
Q	q_1	q_2	q_3	$\cdots$	q_{K-1}	q_K

$$\begin{aligned}
 p(x) &= \sum_z p(x, z) \\
 &= \sum_{k=1}^K p(x, z = C_k) \\
 &= \sum_{k=1}^K p(z = C_k) \cdot p(x|z = C_k) \\
 &= \sum_{k=1}^K q_k \cdot N(x|\mu_k, \Sigma_k)
 \end{aligned}$$

π_k 的意义就是样本属于第k的高斯分布的概率大小

$$p(x) = \sum_{k=1}^K \pi_k \frac{1}{\sqrt{2\Sigma_k^2\pi}} e^{-\frac{(x-\mu_k)^2}{2\Sigma_k^2}} \quad \text{e.g. } \sum_{k=1}^K \pi_k = 1$$

$$p(x) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = N(x|\mu, \sigma)$$

用极大似然法(MLE)估计概率密度

- 输入 $\{x_i\}_{i=1 \sim N}$
- 最大化 $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1 \sim K}$

$$\begin{aligned}\theta_{MLE} &= \arg \max_{\theta} \log[\prod_i^N p(x_i)] \\ &= \arg \max_{\theta} \sum_{i=1}^N \log[p(x_i)]\end{aligned}$$

$$= \arg \max_{\theta} \sum_{i=1}^N \log \left[\sum_{k=1}^K \pi_k \frac{1}{\sqrt{2\Sigma_k^2\pi}} e^{-\frac{(x-\mu_k)^2}{2\Sigma_k^2}} \right] \quad \text{e.g. } \sum_{k=1}^K \pi_k = 1$$

上式可以看出最大似然估计求得目标函数是和的对数形式很难展开，难求偏导

优化非凸问题时常用的方法有：

①梯度下降

沿梯度下降的方向求解极小值（也可以沿梯度上升方向求解极大值）。

无约束优化问题时的常用方法

②启发式方法

一个基于直观或经验构造的算法，在可接受的花费下给出解决优化问题的一个可行解，该可行解与最优解的偏离程度一般不能被预计。模拟退火、神经网络。

③最大期望算法（Expectation-Maximization）

用于对包含隐变量或缺失数据的概率模型进行参数估计

1.不需要调参、一定收敛
2.编程简单

④将非凸问题转化为凸问题来研究



算法原理

EM算法是基于极大似然估计理论的优化算法

$$\log p(X|\theta) = \log \prod_{i=1}^N p(X_i|\theta) = \sum_{i=1}^N \log p(X_i|\theta) = \sum_{i=1}^N \log \left[\sum_{c=1}^K p(X_i, Z_c|\theta) \right]$$

由詹森不等式

$$\log p(X|\theta) = \sum_{i=1}^N \log \left[\sum_{c=1}^K \frac{q(Z_c)}{q(Z_c)} p(X_i, Z_c|\theta) \right] \geq \sum_{i=1}^N \sum_{c=1}^K \left[q(Z_c) \log \frac{p(X_i, Z_c|\theta)}{q(Z_c)} \right] = L(\theta, q)$$

优化目标

$$\theta = \arg \max_{\theta} L(\theta, q)$$

通过最大化目标函数逼近对数似然的极大值

- E步 (E-step)

【固定前一次迭代的 $\theta^{(t-1)}$ ，求解 $q^{(t)}$ 使 $L(\theta, q)$ 取极大值】

$$q = \arg \max_q L(\theta, q) = \arg \max_q \sum_{i=1}^N \sum_{c=1}^K [q(Z_c) \log \frac{p(X_i, Z_c | \theta)}{q(Z_c)}]$$

考虑到先前的不等关系，这里对 $\log p(X|\theta) - L(\theta, q)$ 进行展开

- E步 (E-step)

【固定前一次迭代的 $\theta^{(t-1)}$ ，求解 $q^{(t)}$ 使 $L(\theta, q)$ 取极大值】

$$E - step: q^{(t)} = p(Z|X, \theta^{(t-1)})$$

$$\begin{aligned} \log p(X|\theta) - L(\theta, q) &= \sum_{i=1}^N \log \left[\sum_{c=1}^K p(X_i, Z_c|\theta) \right] - \sum_{i=1}^N \sum_{c=1}^K \left[q(Z_c) \log \frac{p(X_i, Z_c|\theta)}{q(Z_c)} \right] \\ &= \sum_{i=1}^N \left[\log p(X_i|\theta) \sum_{c=1}^K q(Z_c) - \sum_{c=1}^K q(Z_c) \log \frac{p(X_i, Z_c|\theta)}{q(Z_c)} \right] \\ &= \sum_{i=1}^N \sum_{c=1}^K q(Z_c) \log \frac{p(X_i|\theta) q(Z_c)}{p(X_i, Z_c|\theta)} \\ &= \sum_{i=1}^N \sum_{c=1}^K q(Z_c) \log \frac{q(Z_c)}{p(Z_c|X_i, \theta)} \end{aligned}$$

- M步 (M-step)

【使用 $q^{(t)}$ 求解 $\theta^{(t)}$ 使 $L(\theta, q)$ 取极大值】

该极值问题的必要条件是 $\frac{\partial L(\theta, q)}{\partial \theta} = 0$

$$\max_{\theta} L(\theta, q) \Rightarrow \frac{\partial}{\partial \theta} [L(\theta, q)] = 0$$

E_q 表示联合似然 $p(X, Z | \theta)$
对隐分布 $q(Z)$ 的数学期望

$$\Rightarrow \frac{\partial}{\partial \theta} E_q [\log(X, Z | \theta)] = 0$$

$$M - step: \theta^{(t)} = \arg \max_{\theta} E_{q^{(t)}} [\log(X, Z | \theta)]$$

结合GMM的E步展开:

$$q^{(t)}(Z_c) = p(Z_c|X, \theta^{(t-1)}) = \frac{p(X|Z_c, \theta)p(Z_c|\theta)}{p(X|\theta)} = \frac{p(X|Z_c, \theta)p(Z_c|\theta)}{\sum_{c=1}^K p(X|Z_c, \theta)p(Z_c|\theta)}$$

$$q^{(t)}(Z_c) = p(Z_c|X, \theta^{(t-1)}) = \frac{\pi_c N(x|\mu_c, \sigma_c^2)}{\sum_{c=1}^K \pi_c N(x|\mu_c, \sigma_c^2)}$$

$$E - step: q^{(t)} = p(Z|X, \theta^{(t-1)})$$

$$p(X|\theta) = \sum_{c=1}^K \pi_c N(x|\mu_c, \sigma_c^2), \sum_{c=1}^K \pi_c = 1$$

$$\theta = \{\pi_1, \dots, \pi_K, \mu_1, \dots, \mu_K, \sigma_1^2, \dots, \sigma_K^2\}$$

结合GMM的M步展开:

$$E_{q(t)}[\log(X, Z|\theta)] = \sum_{i=1}^N \sum_{c=1}^K [q(Z_c) \log p(X_i, Z_c|\theta)] = \sum_{i=1}^N \sum_{c=1}^K q(Z_c) \log[\pi_c N(X_i|\mu_c, \sigma_c^2)]$$

$$\Rightarrow \frac{\partial}{\partial \mu_c} = 0 \Rightarrow \mu_c = \frac{\sum_{i=1}^N q(Z_c) X_i}{\sum_{i=1}^N q(Z_c)}$$

$$\Rightarrow \frac{\partial}{\partial \sigma_c} = 0 \Rightarrow \sigma_c = \frac{\sum_{i=1}^N q(Z_c) (X_i - \mu_c)^2}{\sum_{i=1}^N q(Z_c)}$$

$$\Rightarrow \frac{\partial}{\partial \pi_c} = 0 \Rightarrow \pi_c = \frac{\sum_{i=1}^N q(Z_c) X_i}{N}$$

$$M - step: \theta^{(t)} = \arg \max_{\theta} E_{q(t)}[\log(X, Z|\theta)]$$

$$p(X|\theta) = \sum_{c=1}^K \pi_c N(x|\mu_c, \sigma_c^2), \sum_{c=1}^K \pi_c = 1$$

$$\theta = \{\pi_1, \dots, \pi_K, \mu_1, \dots, \mu_K, \sigma_1^2, \dots, \sigma_K^2\}$$

- EM算法:

- EM第一步: 随机化 $\{\pi_c, \mu_c, \Sigma_c\}$ 先确定K个高斯分布的具体形式

第n个样本落在第c个高斯的概率

- EM第二步: E-step: $\gamma_{nc} = \frac{\pi_c N(X_n | \mu_c, \Sigma_c)}{\sum_{c=1}^K \pi_c N(X_n | \mu_c, \Sigma_c)}$ ($n = 1 \sim N; c = 1 \sim K$)

- EM第三步: M-step: $N_c = \sum_{n=1}^N \gamma_{nc}$

- $\pi_c^{new} = \frac{N_c}{N}$ (π_c : 第c个高斯的概率)

- $\mu_c^{new} = \frac{1}{N_c} \sum_{n=1}^N \gamma_{nc} X_n$

- $\Sigma_c^{new} = \frac{1}{N_c} \sum_{n=1}^N \gamma_{nc} (X_n - \mu_c^{new})(X_n - \mu_c^{new})^T$

得到新的 π_c^{new} 、 μ_c^{new} 、 Σ_c^{new} 代回第二步中继续计算直至收敛

- 余爱华. 基于EM算法的高斯混合模型参数估计[J]. 现代计算机: 下半月版, 2011(15).
- Polykovskiy, D. and Novikov, A., Bayesian Methods for Machine Learning . Coursera and National Research University Higher School of Economics. 2017
- Expectation Maximization, STA-663: Computational Statistics and Statistical Computing . Duke University. 2016

知人者智，自知者明。
胜人者有力，自胜者强。
知足者富，强行者有志。
不失其所者久，死而不亡者，寿。

谢谢！

