机器学习

5. GMM&EM算法

主要内容

- GMM聚类
- EM算法

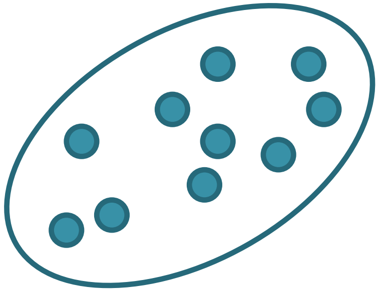
GMM聚类

➢ 密度估计

生成模型：

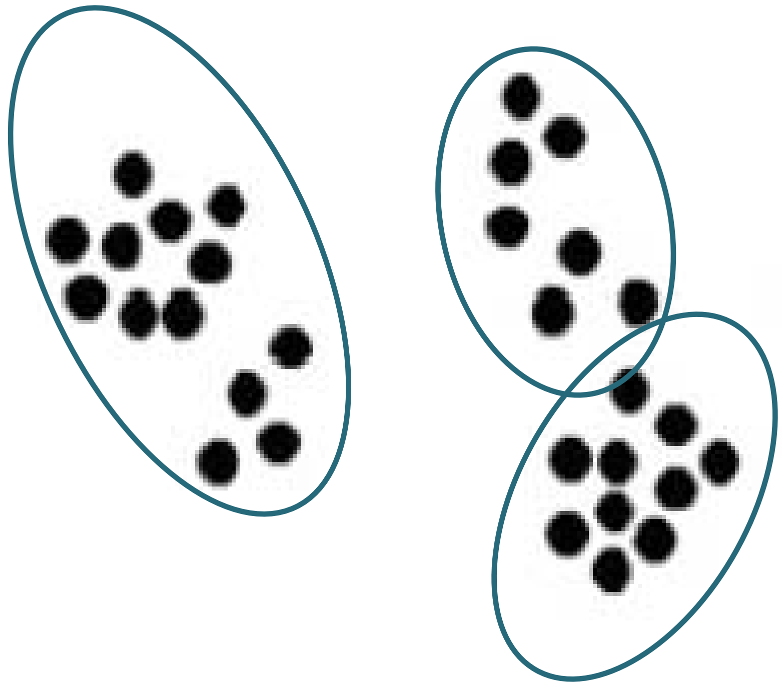
$$p(x_1, x_2, \ldots, x_n | \theta) = \prod_{i=1}^{n} p(x_i | \theta)$$

- 存在隐变量θ
- MLE 模型



GMM聚类

- 混合概率分布模型：如何表达?



- 什么混合模型更易操作?

GMM聚类

- **混合高斯分布：**
  - K 个混合分布
  - 其中第 i 个分布为高斯分布$N(\mu_i, \Sigma_i)$

Gaussian Mixture Model

- **每个数据由以下过程产生：**
  - 根据概率$\pi_i = P(y = i)$选择第 i 个混合分布
  - 根据分布$N(\mu_i, \Sigma_i)$产生数据 x

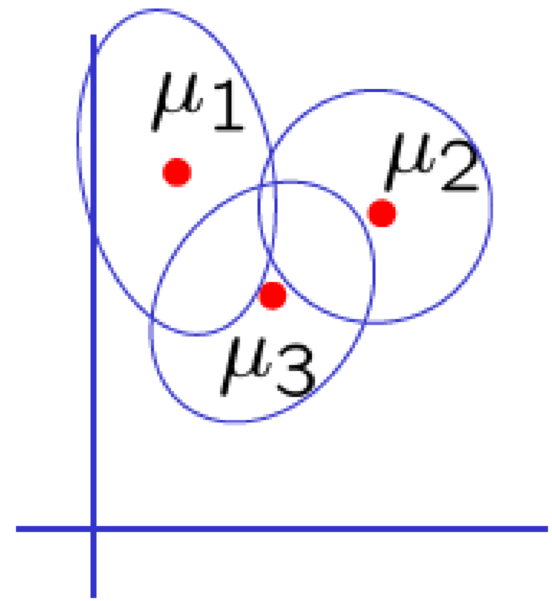
GMM聚类

- 与K均值聚类的本质差别：
  - K均值：硬判断
  - 每个样本仅属于一个类
  - 确定性模型
  - 难以预测

  - GMM：软判断
  - 每个样本以概率属于多个类
  - 生成模型
  - 容易预测，并能再生新的数据

GMM聚类

- GMM的数据分布函数



隐变量

$$p(x|y=i)=N(\mu_i,\Sigma_i)$$

$$p(x)=\sum_{i=1}^{K}p(x|y=i)P(y=i)$$

观测数据

混合成分

混合比例

GMM聚类

为简单起见，假设$\Sigma_i = \sigma^2 I$

$p(x|y = i) = N(\mu_i, \Sigma_i)$

$p(y = i) = \pi_i$

未知变量为$\mu_1, \mu_2, \ldots, \mu_K, \sigma^2, \pi_1, \pi_2, \ldots, \pi_K$

GMM聚类

未知变量为$\mu_1, \mu_2, \ldots, \mu_K, \sigma^2, \pi_1, \pi_2, \ldots, \pi_K$

最大似然估计目标函数：

$$\theta = [\mu_1, \ldots, \mu_K, \sigma^2, \pi_1, \ldots, \pi_K]$$

$$\arg\max_\theta \prod_{j=1}^{n} P(x_j|\theta)$$

$$= \arg\max_\theta \prod_{j=1}^{n} \sum_{i=1}^{K} P(y_j = i, x_j|\theta)$$

$$= \arg\max_\theta \prod_{j=1}^{n} \sum_{i=1}^{K} P(y_j = i|\theta) p(x_j|y_j = i|\theta)$$

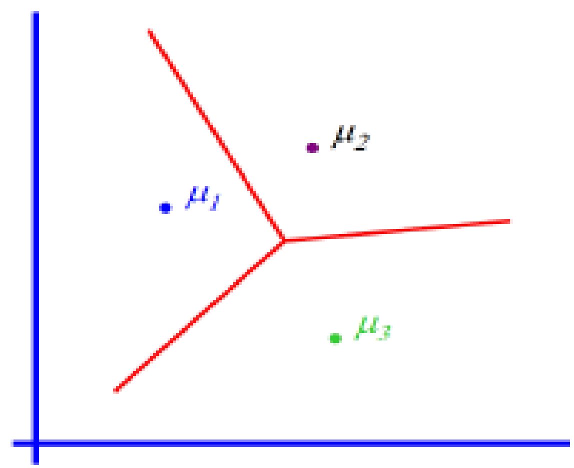
$$= \arg\max_\theta \prod_{j=1}^{n} \sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2} \|x_j - \mu_i\|^2\right)$$

GMM聚类

- 决策过程：如何判断一个点属于哪个类?
  - 基于后验信息:

$$\log \frac{P(y=i|x)}{P(y=j|x)}$$
$$= \log \frac{p(x|y=i)P(y=i)/p(x)}{p(x|y=j)P(y=j)/p(x)}$$
$$= \log \frac{p(x|y=i)\pi_i}{p(x|y=j)\pi_j} = \log \frac{\pi_i \exp(\frac{-1}{2\sigma^2}\|x-\mu_i\|^2)}{\pi_j \exp(\frac{-1}{2\sigma^2}\|x-\mu_j\|^2)} = w^T x$$

- 线性决策面!



GMM聚类

- 若约束选择为硬选择，即：

$$p(y=i)=\begin{cases}1, & 若\ i=C(j)\\ 0, & 其它\end{cases}$$

- 最大似然估计函数为：

$$\arg\max_{\theta}\prod_{j=1}^{n}P(x_j|\theta)=\arg\max_{\theta}\prod_{j=1}^{n}\sum_{i=1}^{K}\overbrace{P(y_j=i)\frac{1}{\sqrt{2\pi\sigma^2}}\exp(\frac{-1}{2\sigma^2}\|x_j-\mu_i\|^2)}^{P(y_j=i,x_j|\theta)}$$

$$=\arg\max_{\theta}\prod_{j=1}^{n}\exp(\frac{-1}{2\sigma^2}\|x_j-\mu_{C(j)}\|^2)$$

$$=\arg\min_{\mu,C}\sum_{j=1}^{n}\|x_j-\mu_{C(j)}\|^2)=\arg\min_{\mu,C}F(\mu,C)$$

➢ **近似退化为K均值！**

GMM聚类

- 一般GMM模型

$$\theta = [\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \pi_1, \ldots, \pi_K]$$

$$p(x|y=i) = N(\mu_i, \Sigma_i)$$

$$p(y=i) = \pi_i$$

- 后验决策：

$$\log \frac{P(y=i|x)}{P(y=j|x)}$$

$$= \log \frac{p(x|y=i)P(y=i)/p(x)}{p(x|y=j)P(y=j)/p(x)}$$

$$= \log \frac{p(x|y=i)\pi_i}{p(x|y=j)\pi_j} = \log \frac{\pi_i \frac{1}{\sqrt{2\pi|\Sigma_i|}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right]}{\pi_j \frac{1}{\sqrt{2\pi|\Sigma_j|}} \exp\left[-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)\right]}$$

$$= x^T W x + w^T x + c$$

GMM聚类

➢ 后验决策：

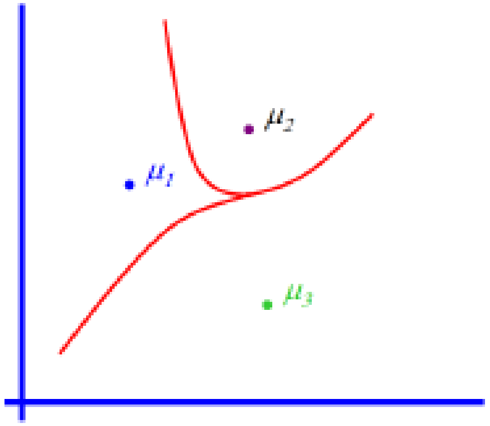
$$\log \frac{P(y=i|x)}{P(y=j|x)}$$

$$= \log \frac{p(x|y=i)P(y=i)/p(x)}{p(x|y=j)P(y=j)/p(x)}$$

$$= \log \frac{p(x|y=i)\pi_i}{p(x|y=j)\pi_j} = \log \frac{\pi_i \frac{1}{\sqrt{2\pi|\Sigma_i|}} \exp\left[-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right]}{\pi_j \frac{1}{\sqrt{2\pi|\Sigma_j|}} \exp\left[-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right]}$$

$$= x^T W x + w^T x + c$$

➢ 二次决策面：



GMM聚类

- 最大似然目标:

$$\arg\max_{\theta} \prod_{j=1}^{n} P(x_j|\theta) = \arg\max_{\theta} \prod_{j=1}^{n} \sum_{i=1}^{K} P(y_j = i, x_j|\theta)$$

$$= \arg\max_{\theta} \prod_{j=1}^{n} \sum_{i=1}^{K} P(y_j = i|\theta) p(x_j|y_j = i, \theta)$$

$$= \arg\max_{\theta} \prod_{j=1}^{n} \sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{2\pi|\Sigma_i|}} \exp\left[-\frac{1}{2}(x_j - \mu_i)^T \Sigma_i^{-1} (x_j - \mu_i)\right]$$

$$\theta = [\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \pi_1, \ldots, \pi_K]$$

- 怎样求解? 求解难度在哪里?
  - 梯度下降?
  - EM! ! !

主要内容

- GMM聚类
- EM算法

EM算法

- EM算法:
  - 处理隐变量分布的一种一般、通用的方法
  - 可解释为在缺失（隐）变量数据下，最大似然估计的一种优化方法
  - 比通常采用的优化方式，如梯度下降简单的多
  - 迭代进行两个步骤：
    - ✓ E步：用均值填充隐变量(计算隐变量概率)
    - ✓ M步：在完整数据上用标准MLE/MAP估计参数

Expectation-Maximization

Majorization Minimization Algorithm

$$\min_{\mathbf{w}} \mathbf{F}(\mathbf{w})$$

Majorization Step: Substitute $\mathbf{F}(\mathbf{w})$ by a surrogate function $Q(\mathbf{w}|\mathbf{w}^k)$ such that

$$F(\mathbf{w}) \leq Q(\mathbf{w}|\mathbf{w}^k)$$

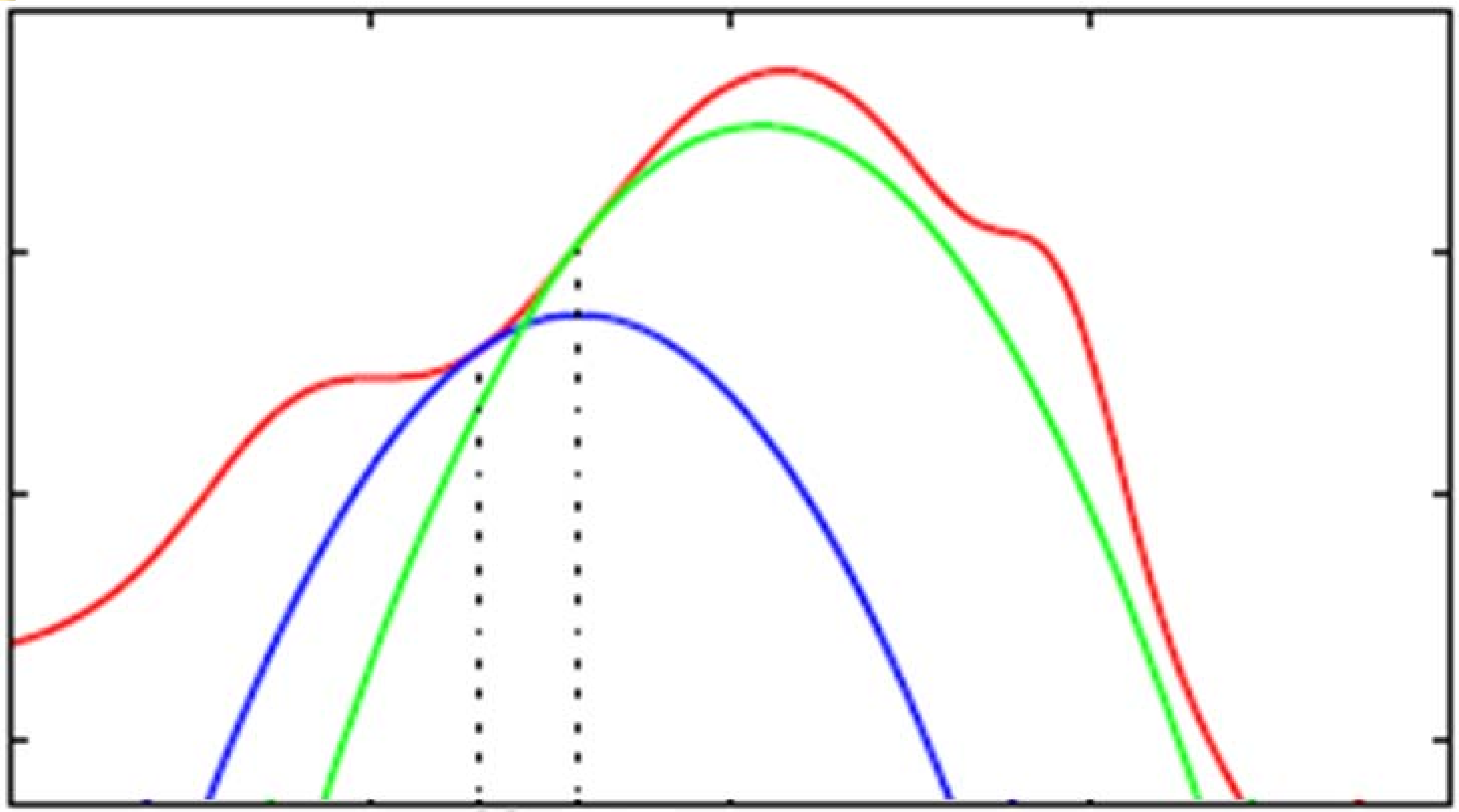
with equality holding at $\mathbf{w} = \mathbf{w}^k$.

Minimization Step: Obtain the next parameter estimate $\mathbf{w}^{k+1}$ by solving the following minimization problem:

$$\mathbf{w}^{k+1} = \arg\min_{\mathbf{w}} Q(\mathbf{w}|\mathbf{w}^k).$$

- 统计与优化领域非常常用的技术!

Majorization Minimization Algorithm



GMM聚类

- 最大似然目标：

$$\arg\max_{\theta} \prod_{j=1}^{n} P(x_j|\theta) = \arg\max_{\theta} \prod_{j=1}^{n} \sum_{i=1}^{K} P(y_j = i, x_j|\theta)$$

$$= \arg\max_{\theta} \prod_{j=1}^{n} \sum_{i=1}^{K} P(y_j = i|\theta) p(x_j|y_j = i, \theta)$$

$$= \arg\max_{\theta} \prod_{j=1}^{n} \sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{2\pi|\Sigma_i|}} \exp\left[-\frac{1}{2}(x_j - \mu_i)^T \Sigma_i^{-1} (x_j - \mu_i)\right]$$

$$\theta = [\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \pi_1, \ldots, \pi_K]$$

EM算法

简单情况：

- 无标号数据$x_1, x_2, \ldots, x_n$
- K 个类
- $p(y = i) = \pi_i$, i=1,2,…,K
- 已知共同方差σ^2
- 求均值变量$\mu_1, \mu_2, \ldots, \mu_K$

$$
\begin{aligned}
p(x_1,\ldots,x_n|\mu_1,\ldots\mu_K) &= \prod_{j=1}^{n} p(x_j|\mu_1,\ldots,\mu_K) \\
&= \prod_{j=1}^{n} \sum_{i=1}^{K} p(x_j, y_j = i|\mu_1,\ldots,\mu_K) \\
&= \prod_{j=1}^{n} \sum_{i=1}^{K} p(x_j|y_j = i|\mu_1,\ldots,\mu_K)p(y_j = i) \\
&\propto \prod_{j=1}^{n} \sum_{i=1}^{K} \exp(-\frac{1}{2\sigma^2}\|x_j - \mu_i\|^2)\pi_i
\end{aligned}
$$

MLE
目标函数

EM算法

➢ E步骤

- 假设上一步迭代获得的参数值为：$\theta^{t-1} = [\mu_1^{t-1}, \mu_2^{t-1}, \ldots, \mu_K^{t-1}]$
- 在当前 t 步，构造以下 Q 函数：

$$Q(\theta^t|\theta^{t-1}) = \sum_{j=1}^{n} \sum_{i=1}^{K} P(y_j = i|x_j, \theta^{t-1}) \log P(x_j, y_j = i|\theta^t)$$

$$\begin{aligned} P(y_j = i|x_j, \theta^{t-1}) &= P(y_j = i|x_j, \mu_1^{t-1}, \ldots, \mu_K^{t-1}) \\ &\propto P(x_j|y_j = i, \mu_1^{t-1}, \ldots, \mu_K^{t-1}) P(y_j = i) \\ &\propto \exp(-\frac{1}{2\sigma^2}\|x_j - \mu_i^{t-1}\|^2)\pi_i \\ &= \frac{\exp(-\frac{1}{2\sigma^2}\|x_j - \mu_i^{t-1}\|^2)\pi_i}{\sum_{i=1}^{K} \exp(-\frac{1}{2\sigma^2}\|x_j - \mu_i^{t-1}\|^2)\pi_i} \end{aligned}$$

更新每个数据归类于某个聚类的概率

EM算法

- M步骤

$$Q(\theta^t|\theta^{t-1}) = \sum_{j=1}^{n}\sum_{i=1}^{K} P(y_j = i|x_j, \theta^{t-1}) \log P(x_j, y_j = i|\theta^t)$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{K} P(y_j = i|x_j, \theta^{t-1})[\log \underbrace{P(x_j|y_j = i, \theta^t)}_{\propto \exp(-\frac{1}{2\sigma^2}\|x_j - \mu_i^t\|^2)} + \log \underbrace{P(y_j = i|\theta^t)}_{\pi_i}]$$

巧妙在何处？？？

布

$$Q(\mu_i^t|\theta^{t-1}) \propto \sum_{j=1}^{n} R_{i,j}^{t-1}(-\frac{1}{2\sigma^2}\|x_j - \mu_i^t\|^2)$$

$$\frac{\partial}{\partial \mu_i^t} Q(\mu_i^t|\theta^{t-1}) = 0 \Rightarrow \sum_{j=1}^{n} R_{i,j}^{t-1}(x_n - \mu_i^t) = 0$$

$$\mu_i^t = \sum_{j=1}^{n} w_j x_j \text{ where } w_j = \frac{R_{i,j}^{t-1}}{\sum_{j=1}^{n} R_{i,j}^{t-1}} = \frac{P(y_j=i|x_j,\theta^{t-1})}{\sum_{l=1}^{n} P(y_l=i|x_l,\theta^{t-1})}$$

更新聚类中心

EM算法

- 综合EM步骤
  - E步：计算所有点归属于每类的概率：

$$P(y_j = i|x_j, \theta^{t-1}) = \frac{\exp(-\frac{1}{2\sigma^2}\|x_j - \mu_i^{t-1}\|^2)\pi_i}{\sum_{i=1}^{K}\exp(-\frac{1}{2\sigma^2}\|x_j - \mu_i^{t-1}\|^2)\pi_i}$$

  - ✓ K均值为硬分配，GMM为软分配
  - M步：计算参数最大值：

$$\mu_i^t = \sum_{j=1}^{n} w_j x_j \qquad w_j = \frac{P(y_j=i|x_j,\theta^{t-1})}{\sum_{l=1}^{n} P(y_l=i|x_l,\theta^{t-1})}$$

  - ✓ 等价于加权MLE.

EM算法

- 一般GMM情形：
  - 无标号数据$x_1, x_2, \ldots, x_n$
  - K 个类
  - $p(y = i) = \pi_i$, i=1,2,…,K
  - 已知共同方差σ^2
  - 求μ_i,，π_i，Σ_i, i=1,2,…,K

EM算法

- 一般GMM情形：

- 需要学习：$\theta = \{\mu_i, , \pi_i, \Sigma_i, i=1,2,\ldots,K\}$
- 假设在 t-1 步估计值为θ^{t-1}
- 在 t 步，首先建立 Q 函数(E 步)，然后最大化得到θ^t(M 步)

$$Q(\theta^t|\theta^{t-1}) = \sum_{j=1}^{n}\sum_{i=1}^{K} P(y_j = i|x_j, \theta^{t-1}) \log P(x_j, y_j = i|\theta^t)$$

EM算法

$$Q(\theta^t|\theta^{t-1}) = \sum_{j=1}^{n}\sum_{i=1}^{K} P(y_j = i|x_j, \theta^{t-1}) \log P(x_j, y_j = i|\theta^t)$$

- E步：计算每个数据隶属概率

$$R_{i,j}^{t-1} = P(y_j = i|x_j, \theta^{t-1}) = \frac{\exp(-\frac{1}{2\sigma^2}\|x_j - \mu_i^{t-1}\|^2)\pi_i^{t-1}}{\sum_{i=1}^{K}\exp(-\frac{1}{2\sigma^2}\|x_j - \mu_i^{t-1}\|^2)\pi_i^{t-1}}$$

- M步：计算加权MLE最大参数值：

$$\frac{\partial}{\partial\theta^t} Q(\theta^t|\theta^{t-1}) = 0$$

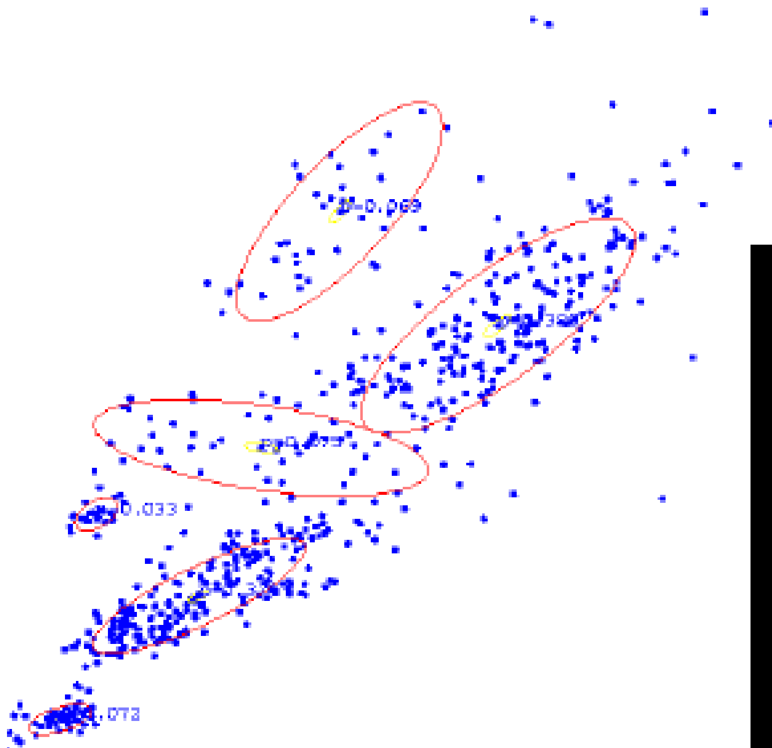
$$\mu_i^t = \sum_{j=1}^{n} w_j x_j \quad \text{where } w_j = \frac{R_{i,j}^{t-1}}{\sum_{j=1}^{n} R_{i,j}^{t-1}}$$

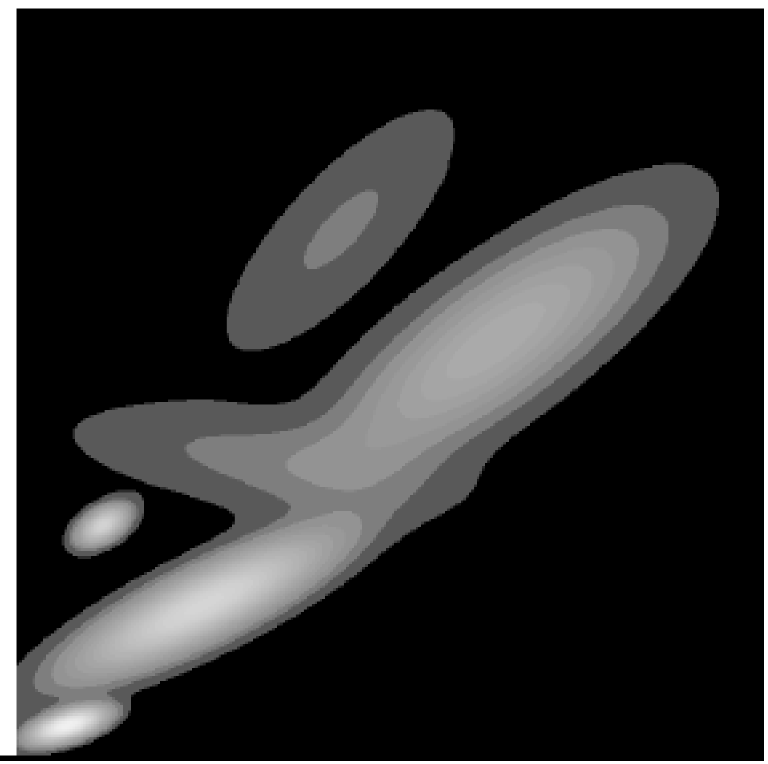
$$\Sigma_i^t = \sum_{j=1}^{n} w_j (x_j - \mu_i^t)^T (x_j - \mu_i^t)$$

$$\pi_i^t = \frac{1}{n}\sum_{j=1}^{n} R_{i,j}^{t-1}$$

EM算法

- 示例





EM算法

- 为什么EM能够有效?
- 在一般情况下EM还能类似操作吗?

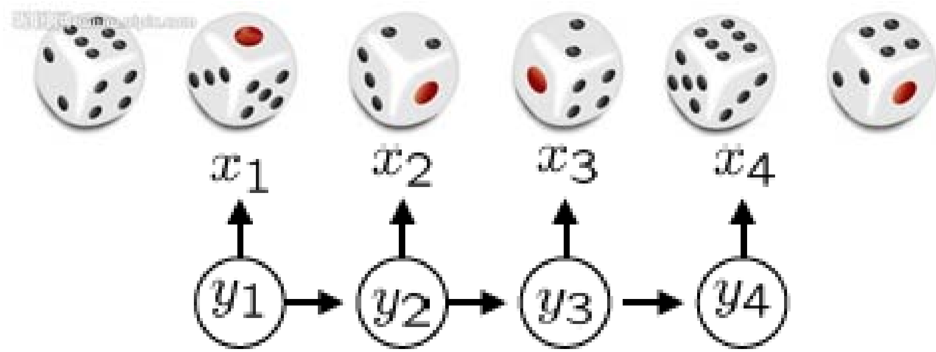
EM算法

- 问题:

  - 观察数据：$D = \{x_1, x_2, \dots, x_n\}$
  - 隐变量：y
  - 参数：θ
  - 目标：$\theta_n = \arg\max_\theta \log P(D|\theta)$

EM算法

- 例子：隐马尔科夫模型：



- 观察数据：$D = \{x_1, x_2, \dots, x_n\}$
- 隐变量：$y = y_1, y_2, \dots, y_n$
- 参数：$\theta = [\pi_i, A, B]$

  初始概率：$P(x_1 = i) = \pi_i$

  转换概率：$P(y_{t+1} = j | y_t = i) = A_{ij}$

  掷色子概率：$P(x_t = l | y_t = i) = B_{il}$
- 目标：$\theta_n = \arg\max_\theta \log P(D|\theta)$

EM算法

➢ 目标： $\arg\max_{\theta} \log P(D|\theta)$

$$\log P(D|\theta^t) = \int dy\, q(y) log P(D|\theta^t)$$

$$= \int dy\, q(y) log \left[\frac{P(y, D|\theta^t)}{P(y|D, \theta^t)} \frac{q(y)}{q(y)}\right]$$

$$= \underbrace{\int dy\, q(y) log P(y, D|\theta^t) - \underbrace{\int dy\, q(y) \log q(y)}_{H(q)}}_{F_{\theta^t}(q(\cdot), D)} + \underbrace{\int dy\, q(y) \log \frac{q(y)}{P(y|D, \theta^t)}}_{KL(q(y) \| P(y|D, \theta^t))}$$

➢ E步： $Q(\theta^t|\theta^{t-1}) = \mathbb{E}_y[\log P(y, D|\theta^t)|D, \theta^{t-1}]$

$$= \int dy\, P(y|D, \theta^{t-1}) \log P(y, D|\theta^t)$$

➢ M步： $\theta^t = \arg\max_{\theta} Q(\theta|\theta^{t-1})$

EM算法

$$\log P(D|\theta^t) = \underbrace{\int dy\, q(y) log P(y,D|\theta^t) \overbrace{- \int dy\, q(y)\log q(y)}^{H(q)}}_{F_{\theta^t}(q(\cdot),D)} + \underbrace{\int dy\, q(y)\log\frac{q(y)}{P(y|D,\theta^t)}}_{KL(q(y)\|P(y|D,\theta^t))}$$

- E步：$Q(\theta^{t+1}|\theta^t) = \int dy\, P(y|D,\theta^t)\log P(y,D|\theta^{t+1})$

$$q(y) = P(y|D,\theta^t)$$

$$\Rightarrow KL(q(y)\|P(y|D,\theta^t)) = 0$$

$$\Rightarrow \log P(D|\theta^t) = F_{\theta^t}(P(y|D,\theta^t),D)$$

$$= \int dy\, P(y|D,\theta^t) log P(y,D|\theta^t) - \int dy\, P(y|D,\theta^t)\log P(y|D,\theta^t)$$

- M步：$\le \int dy\, P(y|D,\theta^t) log P(y,D|\theta^{t+1}) - \int dy\, P(y|D,\theta^t)\log P(y|D,\theta^t)$

EM算法

$$\log P(D|\theta^t) = \underbrace{\int dy\, q(y) log P(y, D|\theta^t) \underbrace{- \int dy\, q(y) \log q(y)}_{H(q)}}_{F_{\theta^t}(q(\cdot), D)} + \underbrace{\int dy\, q(y) \log \frac{q(y)}{P(y|D, \theta^t)}}_{KL(q(y)\|P(y|D,\theta^t))}$$

- 定理： $P(D|\theta^t) \leq P(D|\theta^{t+1})$

$$\begin{aligned}
\log P(D|\theta^t) &= F_{\theta^t}(P(y|D,\theta^t), D) \\
&\leq \int dy\, P(y|D,\theta^t) log P(y, D|\theta^{t+1}) - \int dy\, P(y|D,\theta^t) \log P(y|D,\theta^t) \\
&= F_{\theta^{t+1}}(P(y|D,\theta^t), D) \\
&= \log P(D|\theta^{t+1}) - KL(P(y|D,\theta^t)\|P(y|D,\theta^{t+1})) \\
&\leq \log P(D|\theta^{t+1})
\end{aligned}$$

EM算法

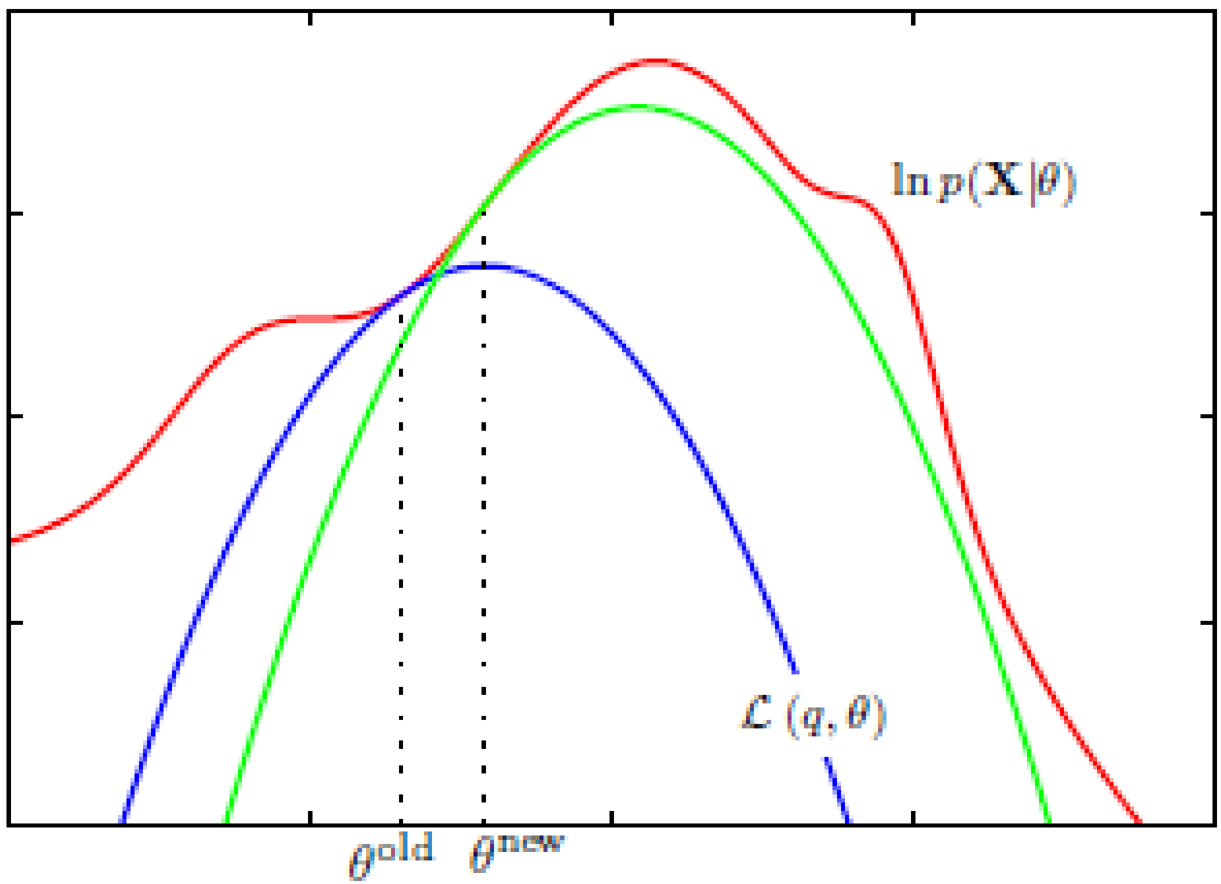
- 目标： $\arg\max_{\theta} \log P(D|\theta)$
- E步： $$Q(\theta^t|\theta^{t-1}) = \mathbb{E}_y[\log P(y, D|\theta^t)|D, \theta^{t-1}]$$ $$= \int dy\, P(y|D, \theta^{t-1}) \log P(y, D|\theta^t)$$
- M步： $$\theta^t = \arg\max_{\theta} Q(\theta|\theta^{t-1})$$

$$P(D|\theta^t) \leq P(D|\theta^{t+1})$$

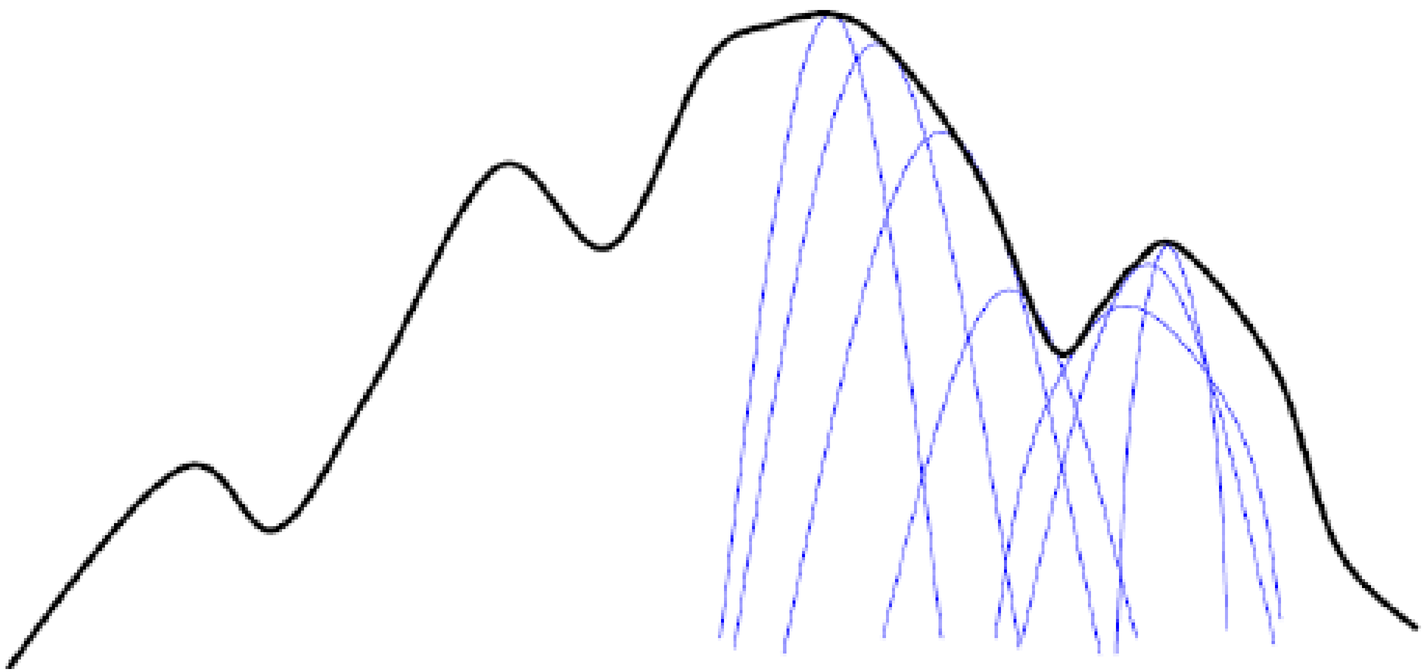
EM算法

- 收敛原理图

$$\log P(D|\theta^t) = \int dy\, q(y) log P(y, D|\theta^t) - \int dy\, q(y) \log q(y) + \int dy\, q(y) \log \frac{q(y)}{P(y|D, \theta^t)}$$

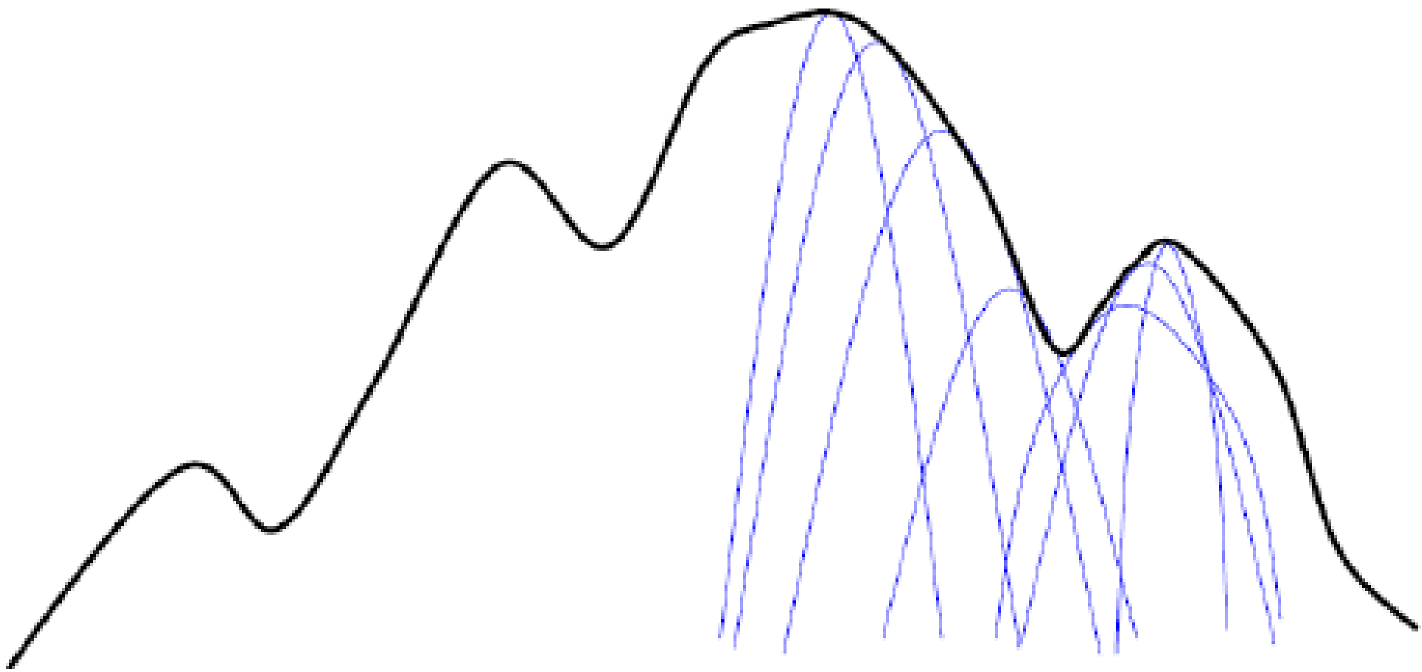


- M步是否一定要找到最大?
- 是否存在局部极优问题?
- 如何尝试解决?



EM算法

- 局部极优问题:



- 如何尝试克服?

要求

1. GMM方法的基本原理
2. EM方法的基本思想

阅读：

[1] Pattern Recognition and Machine Learning, Christopher , M. Bishop, Springer, 2006. 9. Mixture Models and EM