Mixtures of Gaussians and the EM algorithm

- 本节使用期望最大值算法(EM algorithm)来进行密度估计(density estimation)
- 混合高斯模型(mixture of Gaussians)建模
  - 给定联合分布(joint distribution) $p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$
  - zi服从多项式分布，是隐含类别标签 $z^{(i)} \sim \text{Multinomial}(\phi)$
    - $\phi_j \geq 0,\ \sum_{j=1}^{k} \phi_j = 1$
    - $\text{parameter } \phi_j \text{ gives } p(z^{(i)} = j)$
    - k表示zi能取到的个数
  - xi的条件分布是高斯分布 $x^{(i)}|z^{(i)} = j \sim \mathcal{N}(\mu_j, \Sigma_j)$
  - 假设每个xi都是从{1,...,k}中随机选取zi来生成的，xi服从来自zi的k个高斯分布中的一个分布
  - zi是隐含随机变量(latent random variables)
  - 模型的参数为 ϕ, μ, Σ
- 似然函数求解
  - $$\begin{aligned} \ell(\phi, \mu, \Sigma) &= \sum_{i=1}^{m} \log p(x^{(i)}; \phi, \mu, \Sigma) \\ &= \sum_{i=1}^{m} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)}|z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi). \end{aligned}$$
  - 如果我们用偏导数=0来求解各个参数，发现根本不可能以闭合形式(closed form)来找到这些参数的最大似然估计
  - zi表示xi来自k个高斯分布的概率，**假设zi已知**
  - 则似然函数可以写成 $$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)}|z^{(i)}; \mu, \Sigma) + \log p(z^{(i)}; \phi)$$
  - 求解得到 $$\begin{aligned} \phi_j &= \frac{1}{m} \sum_{i=1}^{m} 1\{z^{(i)} = j\}, \\ \mu_j &= \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}, \\ \Sigma_j &= \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}. \end{aligned}$$
    - 与几节课前讲过的高斯判别模型相比，把{yi=1},{yi=0}换成了{zi=j}
  - 也就是说，如果zi已知，那么这个最大似然估计几乎等同于高斯判别模型，只不过这里zi扮演了高斯判别模型中分类标签的角色。
  - 可是zi实际上是未知的，怎么办呢？由此引出了EM算法
- 期望最大值算法(Expectation-Maximization)
  - 是一个迭代算法，主要分为两步：E步、M步
  - E步去“猜测”zi的值，也就是计算zi的后验概率，M步基于猜测更新参数
  - 算法
    - Repeat until convergence: {

      (E-step) For each i, j, set
      $$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$
      (M-step) Update the parameters:
      $$\begin{aligned} \phi_j &:= \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)}, \\ \mu_j &:= \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}, \\ \Sigma_j &:= \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}} \end{aligned}$$
      }
    - E步中 $$p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)}|z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)}|z^{(i)} = l; \mu, \Sigma) p(z^{(i)} = l; \phi)}$$
  - 一些注意点
    - 与高斯判别模型相比，这里的均方差每个都不一样，而且这里的zi是多项式分布
    - 与zi已知相比，这里的wj(i)代替了前面的1{zi=j}，由简单的0/1变成了概率值
    - 与k-means相比，这里的分配使用了“软”指定，k-means中给数据的所属簇进行了强赋值，而这里给wj(i)进行了弱赋值
      - 弱是指概率，从[0,1]区间内选定一个值
        强是单次最佳猜测，例如从集合{0,1,...,k}中选出一个值
    - EM算法容易导致局部最优