#MLorDS

尝试思考下面的问题：100 名同学站成一排，其中 50 个男生，50 个女生。现在，你可以看到同学们的脸，那么你记下每个人的性别和身高，然后对男生和女生分别使用高斯函数建模。
现在，我们将问题变得困难一些：你已经知道，男生和女生的身高均值明显不同，但是你只能看到体检表上记录的身高数据，无法看到同学的性别，你要如何对男生和女生的身高进行建模呢？
在生活中，我们获得的数据集中的变量都是可观测的，然而，在可观测的数据背后，往往隐藏着深不可测的真相——例如，上面例子中，同学们的性别。这样的变量被称为**隐变量**(Hidden Variable)，EM 算法的目标是——***解决带有隐变量的数据的参数估计问题***。

Maximum Likelihood Estimator

Likelihood 是一个关于模型参数的函数。在数值上：

$$L(\theta) = P(x|\theta)$$

但是，似然是给定 x 后，关于 θ 的函数
极大似然估计的先决条件是给定数据集 x 中数据的分布！

EM——如果数据分布未知，应该怎么做

Jensen 不等式

对于凸函数

$$E[f(x)] \geq f(E[x])$$

（这个式子的几何意义就可以简单地理解为：两点函数值的平均大于两点中点处的函数值）
进一步地，如果 f 严格凸，那么取等条件：

$$P(X = E[x]) = 1$$

极大似然估计到 EM 算法

隐变量 (Hidden Variable)：是我们无法观测到的变量，例如聚类时的类别标签。我们将 (x, z) 整体看作一个完整的观测样本。
首先，我们用极大似然估计求解 θ，其中，**第 i 个样本的**可观测量被记为 x_i，不可观测的特征被记为 z_i。**注意：这里 z_i 是可以取很多值的，比如一朵鸢尾花可能被分在第一类，也有可能被分在第二类，等等**，那么，根据极大似然估计：

$$LL(\theta) = \sum_i \log p(x|\theta) = \sum_i \log \sum_{z_i} p(x_i, z_i|\theta)$$

这一步的第二个求和是因为，一个样本的 z_i 可能有多种，我们要对这个样本取所有 z_i 的概率求和，然后相加。我们现在假设 z_i 有一个分布 $Q_i(z_i)$，（例如，每朵花被分到其中一类，是有一个概率的），将 $Q_i(z_i)$ 引入方程，那么，上面的式子可以写成：

$$LL = \sum_i \log \sum_{z_i} Q_i(z_i) \frac{p(x_i, z_i|\theta)}{Q_i(z_i)}$$

现在，我们把 z_i 看成 Jensen 不等式中的 x，那么上式的第二个求和号右侧等价于

$$LL = \sum_i \log[E(\frac{p(x_i, z_i|\theta)}{Q_i(z_i)})]$$

利用 Jensen 不等式，上式可以写成：

$$LL \geq \sum_i E[\log(\frac{p(x_i, z_i|\theta)}{Q_i(z_i)})] = \sum_i \sum_{z_i} Q_i(z_i) \log \frac{p(x_i, z_i|\theta)}{Q_i(z_i)}$$

我们转换优化的目标函数：使得 LL 的下界最大即可。LL 的下界就是最右边这一部分。我们记

$$LB = \sum_i \sum_{z_i} Q_i(z_i) \log \frac{p(x_i, z_i|\theta)}{Q_i(z_i)}$$

现在，我们想让 LB 最大，这时候，LL 成为 LB 的一个上界。我们希望尽可能地使得 LB 靠近 LL，那么取等的条件是自变量全部落在它们的期望上，也就是，自变量需要成为一个恒定的值 c。

$$\frac{p(x_i, z_i|\theta)}{Q_i(z_i)} = c \Rightarrow \sum_{z_i} p(x_i. z_i|\theta) = c \sum Q_i(z_i) = c$$

把这一部分再变形，有：

$$Q_i(z_i) = \frac{p(x_i, z_i|\theta)}{c} = \frac{p(x_i, z_i|\theta)}{\sum_{z_i} p(x_i, z_i|\theta)} = \frac{p(x_i, z_i|\theta)}{p(x_i|\theta)}$$

分子上也是用了和开始时相同的手法。然后这又是一个条件概率：

$$Q_i(z_i) = p(z_i|x_i, \theta)$$

现在，我们就可以得到完整的流程了：

- 输入：独立同分布的可观测样本 x，不可观测的隐变量 z（它只是不可直接观测，而不是不能输入，就如同聚类的时候我们要首先指定一个初始类别）。
- 迭代
  - E-step: 这一步 θ 已知，计算 $Q_i(z_i)$， $Q_i(z_i) = p(z_i|x_i, \theta)$
  - M-step: 这一步 $Q_i(z_i)$ 已知，最大化目标函数以求解 θ： $\theta = \arg\max_\theta LB$

这个是在隐变量已知的情况下，取参数theta时，当前数据出现的概率

$$LL \geq \sum_i E[\log(\frac{p(x_i, z_i|\theta)}{Q_i(z_i)})] = \sum_i \sum_{z_i} Q_i(z_i) \log \frac{p(x_i, z_i|\theta)}{Q_i(z_i)}$$

这个是隐变量的分布