2 无监督学习(Unsupervised Learning)

2.1 简介

- 动机(Motivation)

无监督学习的目的是寻找无标签数据 $\{x^{(1)}, \ldots, x^{(m)}\}$ 中的隐藏模式。

- Jensen不等式(Jensen's inequality)

记 f 为凸函数(convex function)，X 为随机变量。我们得到如下不等式：

$$\boxed{E[f(X)] \geqslant f(E[X])}$$

2.2 聚类(Clustering)

2.2.1 最大期望算法(Expectation-Maximization)

- 隐变量(Latent variables)

隐变量是隐藏或不可观测的变量，它使得估计问题更困难，通常用 z 表示。下表是存在隐变量时最常用的设置：

设置	隐变量 z	$x\|z$	备注
k 高斯混合模型(Mixture of k Gaussians)	多项的（ϕ）	$\mathcal{N}(\mu_j, \sum_j)$	$\mu_j \in \mathbb{R}^n, \phi \in \mathbb{R}^k$
因子分析(Factor analysis)	$\mathcal{N}(0, I)$	$\mathcal{N}(\mu + \Lambda z, \psi)$	$\mu_j \in \mathbb{R}^n$

- 算法

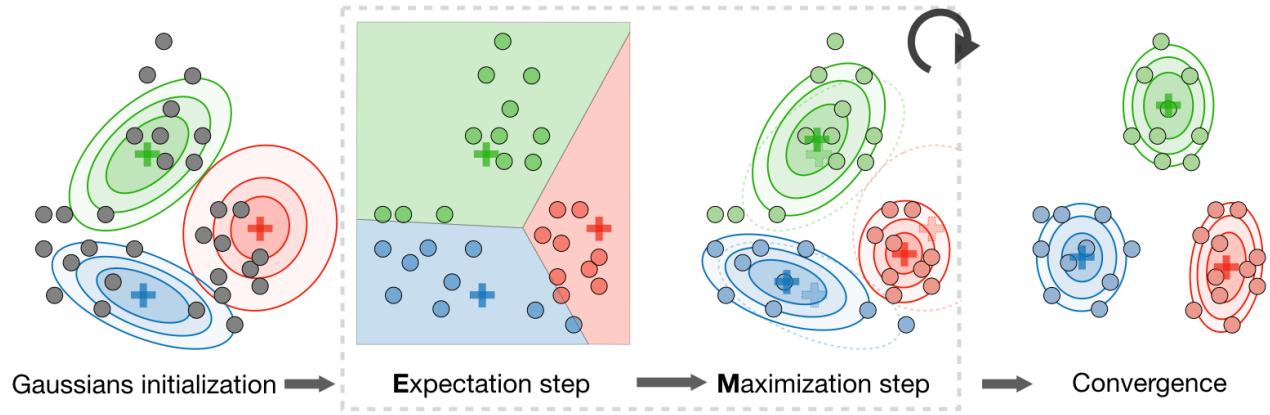
最大期望算法（Expectation-Maximization，EM）通过重复构建似然下界（E-step），并优化下界（M-step）来给出通过MLE参数 θ 的有效方法，如下:

E-step：估计簇 $z^{(i)}$ 中每个数据点 $x^{(i)}$ 的后验概率(posterior probability) $Q_i(z^{(i)})$，如下：

$$\boxed{Q_i(z^{(i)}) = P(z^{(i)}|x^{(i)}; \theta)}$$

M-step：使用后验概率 $Q_i(z^{(i)})$ 作为簇在数据点 $x^{(i)}$ 的权重，去分别重新估计每个簇，如下：

$$\theta_i = argmax_\theta \sum_i \int_{z^{(i)}} Q_i(z^{(i)}) \log\left(\frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}\right) dz^{(i)}$$



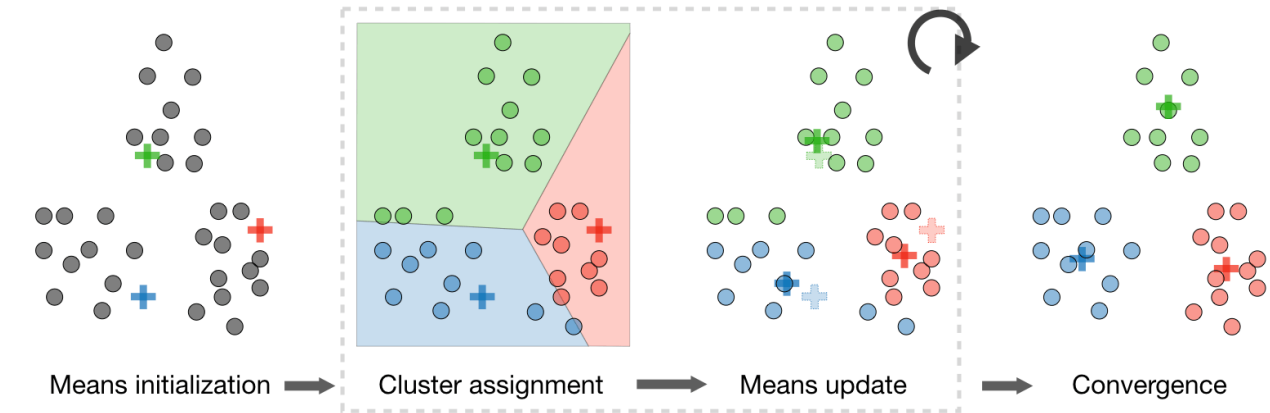
2.2.2 k-均值聚类(k-means clustering)

记 $c^{(i)}$ 为簇内第 i 个点，μ_j 是簇 j 的中心点。

- 算法

随机初始化簇形心 $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}$，k均值算法重复以下步骤直到收敛：

$$c^{(i)} = argmin_j \|x^{(i)} - \mu_j\|^2 \text{ 和 } \mu_j = \frac{\sum_{i=1}^{m} 1_{\{c^{(i)}-j\}} x^{(i)}}{\sum_{i-1}^{m} 1_{\{c^{(i)}-j\}}}$$



- 失真函数(Distortion function)

为了查看算法是否收敛，定义如下的失真函数：

$$J(c, \mu) = \sum_{i=1} \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

2.2.3 层次聚类(Hierarchical clustering)

- 算法

它是一种聚类算法，采用聚合分层方法，以连续方式构建嵌套的聚类。

- 类型

为了优化不同的目标函数，有不同种类的层次聚类算法，汇总见下表：

Ward linkage	Average linkage	Complete linkage
簇内距离最小	簇之间平均距离最小	簇之间最大距离最小

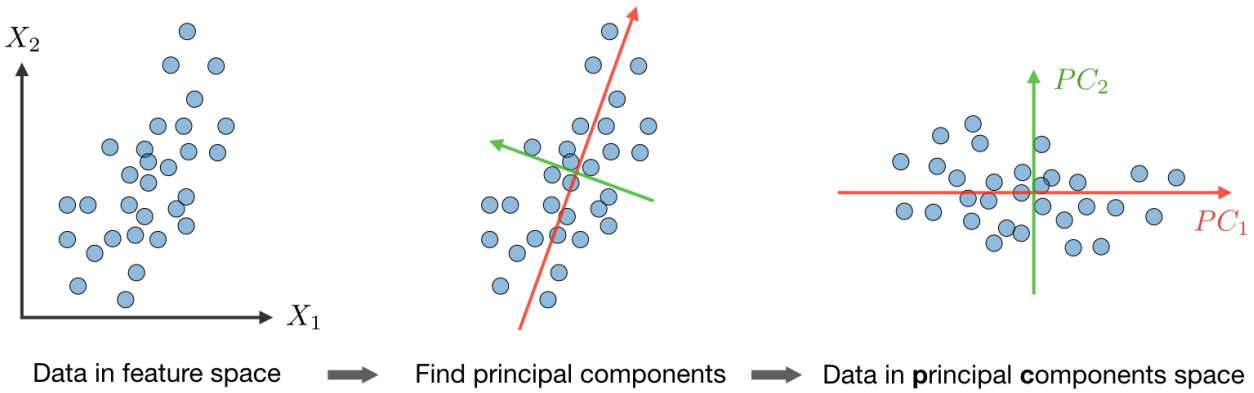
2.2.4 聚类评估指标

在无监督的学习环境中，通常很难评估模型的性能，因为没有像监督学习环境中那样的ground-truth标签。

- 轮廓系数(Silhouette coefficient)

记 a 表示一个样本和同一类中其他点的平均距离，b 表示一个样本和距离最近的其他类的平均距离，一个样本的轮廓系数可表示为：

$$\boxed{s = \frac{b - a}{\max(a, b)}}$$



- Calinski-Harabaz指数(Calinski-Harabaz index)

记 k 为簇的个数，类间、类内的散布阵(dispersion matrices) B_k 和 W_k 定义如下：

$$B_k = \sum_{j=1}^{k} n_{c^{(i)}} (\mu_{c^{(i)}} - \mu)((\mu_{c^{(i)}} - \mu))^T, W_k = \sum_{j=1}^{m} (x^{(i)} - \mu_{c^{(i)}})(x^{(i)} - \mu_{c^{(i)}})^T,$$

Calinski-Harabaz指数 $s(k)$ 表明了聚类模型对聚类的定义的好坏，得分越高，聚类就越密集，分离得也越好。定义如下:

$$s(k) = \frac{T_r(B_k)}{T_r(W_k)} \times \frac{N - k}{k - 1}$$

2.3 降维(Dimension reduction)

2.3.1 主成分分析(Principal component analysis)

主成分分析是一种统计方法。通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。

- 特征值、特征向量

给定一个矩阵 $A \in \mathbb{R}^{n \times n}$，如果存在一个向量 $z \in \mathbb{R}^n$，$lambda$ 叫做 A 的特征值:

$$\boxed{Az = \lambda z}$$

- 谱定理

令 $A \in \mathbb{R}^{n \times n}$。如果 A 是对称的，那么 A 可以通过正交矩阵 $U \in \mathbb{R}^{n \times n}$ 对角化。记 $Lambda = diag(\lambda 1, \ldots, \lambda n)$，我们得到：

$$\boxed{\exists \Lambda \; diagnoal, A = U \Lambda U^T \in \mathbb{R}^{n \times n}}$$

> 备注：和最大特征值关联的特征向量，被称作矩阵 A 的主特征向量(principal eigenvector)。

- 算法

主成分分析（PCA）是一种降维方法，通过使数据的方差最大化，在k维上投影数据，方法如下：

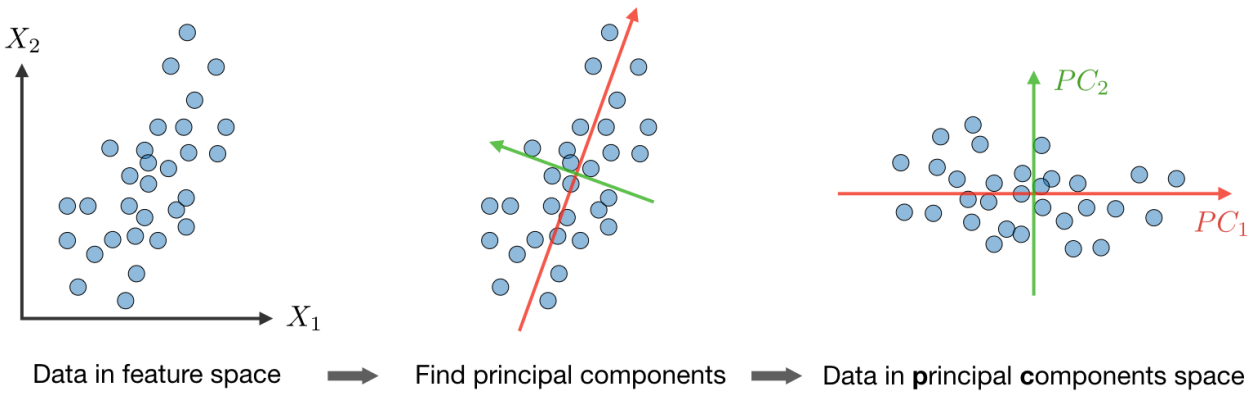
第1步：将数据标准化，使其均值为0，标准差为1。

$$\boxed{x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j}} \text{ 其中 } \boxed{\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}} \text{ 且 } \boxed{\sigma_j^2 = \frac{1}{m} \sum_{i=1}^{m} (x_j^{(i)} - \mu_j)^2}$$

第2步：计算 $\sum = \frac{1}{m} \sum_{i=1}^{m} x^{(i)} x^{(i)T}$，它与实特征值对阵。

第3步：计算 $\sum$ 的 k 个正交主特征向量 $u_1, \ldots, u_k \in \mathbb{R}^n$，即k个最大特征值的正交特征向量。

第4步：在 $span_{\mathbb{R}}(u_1, \ldots, u_k)$ 上投射数据。这个会产生 k 维空间的最大方差。



2.3.2 独立成分分析(Independent component analysis)

这是一种寻找潜在生成源的技术。

- 假设

我们假设数据x是通过混合和非奇异(non-singular)矩阵 A，由 n 维源向量 $s = (s_1, \ldots, s_n)$ 生成的（其中，s_i 是独立的随机变量）,如下：

$$\boxed{x = As}$$

目的是找到解混矩阵(unmixing matrix) $W = A^{-1}$。

- Bell和Sejnasi-ICA算法(Bell and Sejnowski ICA algorithm)

该算法通过以下步骤找到解混矩阵W：

将 $x = As = W^{-1}s$ 的概率表示为：

$$p(x) = \prod_{i=1}^{n} p_s(\omega_i^T x) \cdot |W|$$

记 g 为激活函数，给定我们的训练集 $\{x^{(i)}, i \in [1, m]\}$，则 对数似然函数可表示为：

$$l(W) = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \log\Big(g'(\omega_j^T x^{(i)}) \Big) + \log |W| \right)$$

因此，随机梯度上升学习规则是对于每个训练样本 $x(i)$，我们更新 W 如下：

$$\boxed{W \leftarrow W + \alpha \left(\begin{pmatrix} 1 - 2g(w_1^T x^{(i)}) \\ 1 - 2g(w_2^T x^{(i)}) \\ \cdot \\ \cdot \\ \cdot \\ 1 - 2g(w_n^T x^{(i)}) \end{pmatrix} {x^{(i)}}^T + (W^T)^{-1} \right)}$$