Unsupervised Learning

无监督学习

OUTLINE

- 主成分分析
- 聚类分析
- R实现

无监督学习

- 我们之前研究了回归问题和分类问题
- 区别在于响应变量是离散或连续
- 如果没有响应变量？
- 对响应变量的预测不再重要
- 我们更关心特征自身的性质

无监督学习

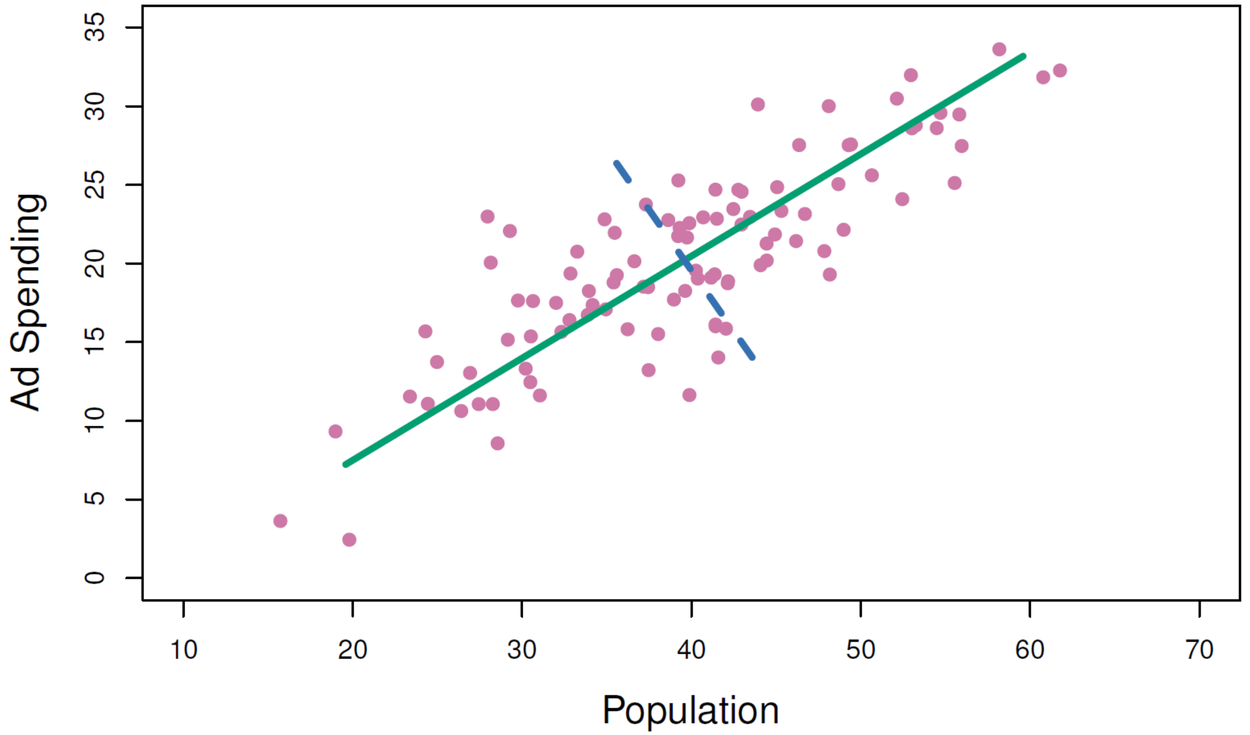
- 无监督学习更困难
- 无监督意味着没有简单的目标和标准
- 如何检验机器学习的结果？

主成分分析

- 广泛的用途
  - 有效降低数据的维度
  - 图像识别
  - 图像去噪
- 几种视角
  - 最大方差投影;
  - 最小重构误差;
  - 针对某些参数的概率模型
  - 线性流形(Manifold)对齐.

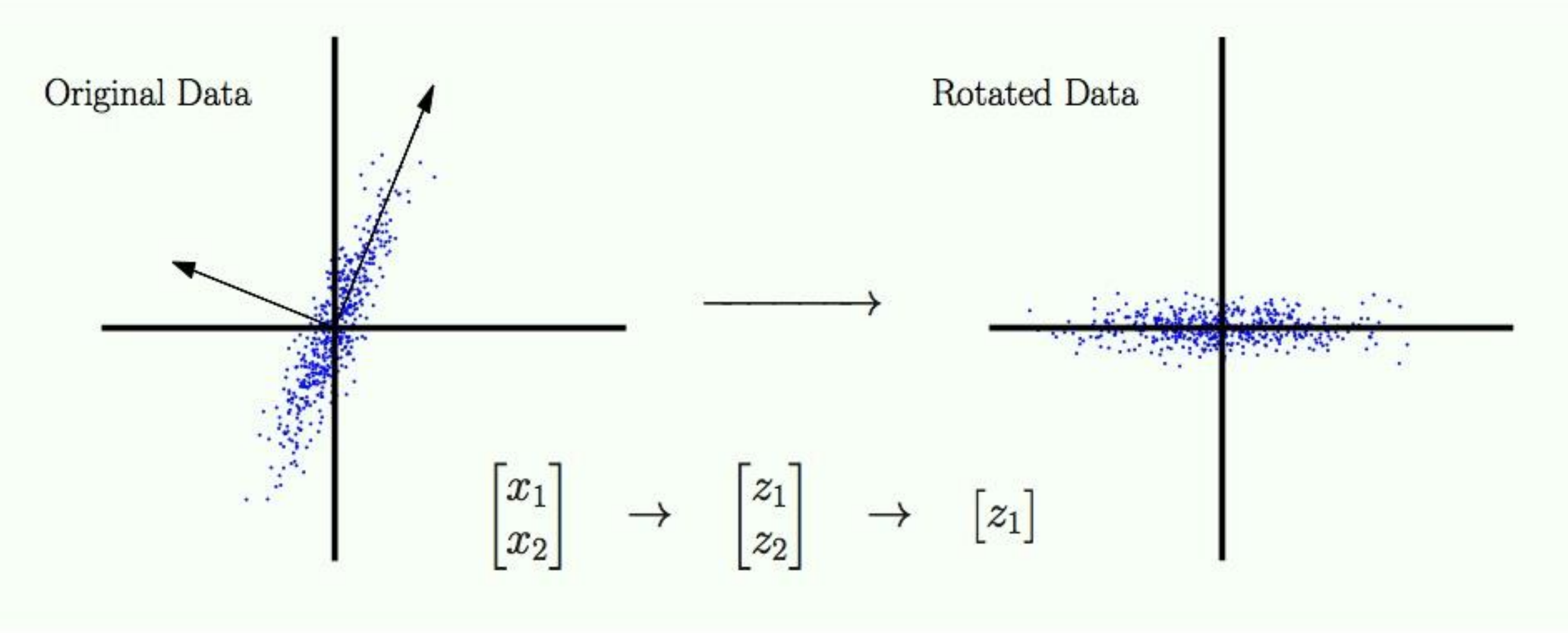
最大方差投影

- 尝试找到一个方向，能够最大程度解释方差
- 如果我们将数据投影到长轴上，投影之后方差最大
- 如果我们将数据投影到短轴上，投影之后方差最小



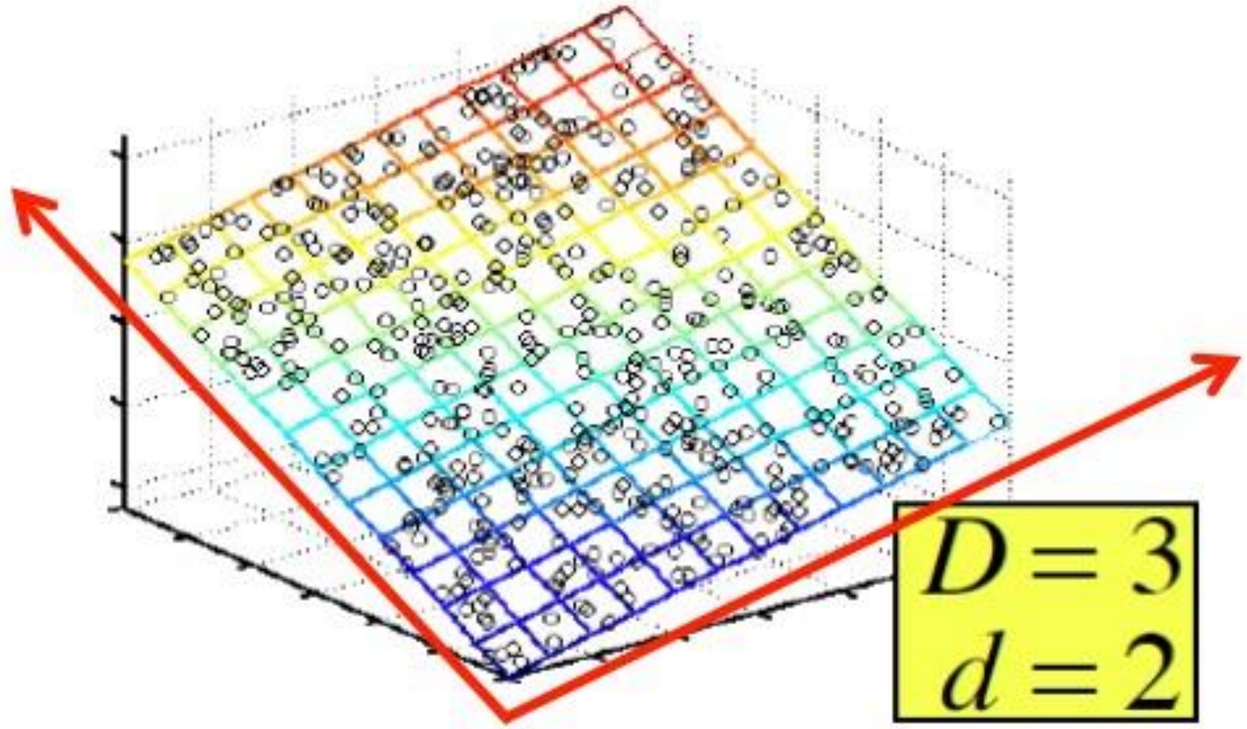
最大方差投影

- 考虑坐标变换



最大方差投影

- 3 维情况下，想找到一个平面



最大方差投影

- 给定训练样本$\{x_i\}_{i=1}^n$，其中$x_i \in \mathcal{R}^d$.
- 假设x_i已被中心化，即满足$\sum_{i=1}^n x_i = 0$
- 我们尝试找到方向v，使得原始数据在该方向上的投影$z = x^T v$方差最大。
- 显然，计算可得

  $Var(z) = \frac{1}{n}\sum_{i=1}^n \left(x_i^T v\right)^2 = \frac{1}{n}\sum_{i=1}^n v^T x_i x_i^T v = v^T S v,$

  其中$S = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$

最大方差投影

- 具体的优化问题可写为:

$$\max_{v} v^T S v \quad s.t. \ \|v\|^2 = 1.$$

- 求解带限制的优化问题，使用拉格朗日乘子法:

$$L(v, \lambda) = v^T S v + \lambda(1 - \|v\|^2)$$

- 求导得到

$$\frac{\partial L(v, \lambda)}{\partial v} = 2Sv - 2\lambda v = 0 \implies Sv = \lambda v.$$

- 因而，最优的v，为其最大的特征根所对应的特征向量.

最大方差投影

- 想一想 如何求解第二个主成分?
- $\max_{v} v^T S v \quad s.t.\ \|v\|^2 = 1, v_1^T v = 0.$
- 主成分之间正交

最小重构误差

- 给定训练样本$\{x_i\}_{i=1}^n$，其中$x_i \in \mathcal{R}^d$.
- 假设x_i已被中心化，即满足$\sum_{i=1}^n x_i = 0$
- 考虑一组正交基$u_1, \ldots, u_d \in \mathcal{R}^d$，进而有
  $x_i = \sum_{j=1}^d \alpha_{ij} u_j$，其中$\alpha_{ij} = u_j^T x_i$
- 我们考虑其在子空间的投影:

$$\hat{x}_i = \sum_{j=1}^{K} \alpha_{ij} u_j$$

最小重构误差

- PCA 尝试去最小化下述重构误差

$$\frac{1}{n}\sum_{i=1}^{n}\|x_i-\hat{x}_i\|^2=\frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j=1}^{d}\alpha_{ij}u_j-\sum_{j=1}^{K}\alpha_{ij}u_j\right\|^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left\|\sum_{j=K+1}^{d}\alpha_{ij}u_j\right\|^2=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=K+1}^{d}\alpha_{ij}^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=K+1}^{d}u_j^Tx_ix_i^Tu_j=\sum_{j=K+1}^{d}u_j^TSu_j$$

最小重构误差

- 其优化问题可以写为

$$\min_{\mathrm{u_j,j=1,\ldots,}K} \sum_{j=K+1}^{d} u_j^T S u_j \,, s.t. \left\| u_j \right\|^2 = 1$$

- 即等价于前K 个最大特征向量

主成分数量

- 主成分分析提供了数据的低维近似
- 在低维近似中，我们损失了多少信息？
- 统计学中用方差刻画信息

$$\underbrace{\sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n}x_{ij}^{2}}_{Var.\,of\ data}=\underbrace{\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d}v_{jk}x_{ij}\right)^{2}}_{Var.\,of\ first\ K\ PCs}+\underbrace{\frac{1}{n}\sum_{j=1}^{d}\sum_{i=1}^{n}\left(x_{ij}-\sum_{k=1}^{K}z_{ik}v_{jk}\right)^{2}}_{MSE\ of\ K-dimensional\ approximation}$$

- 被解释方差的比例（PVE）随M上升

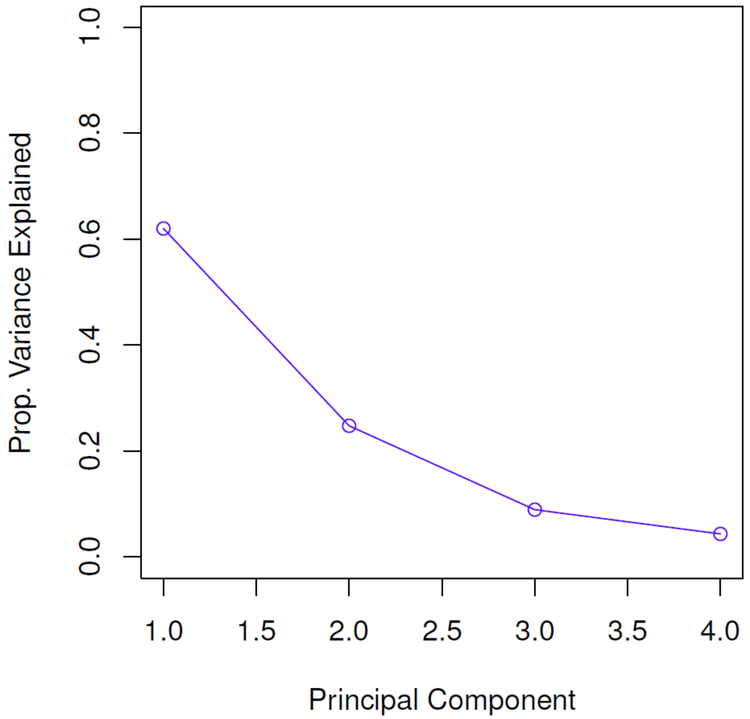
主成分数量

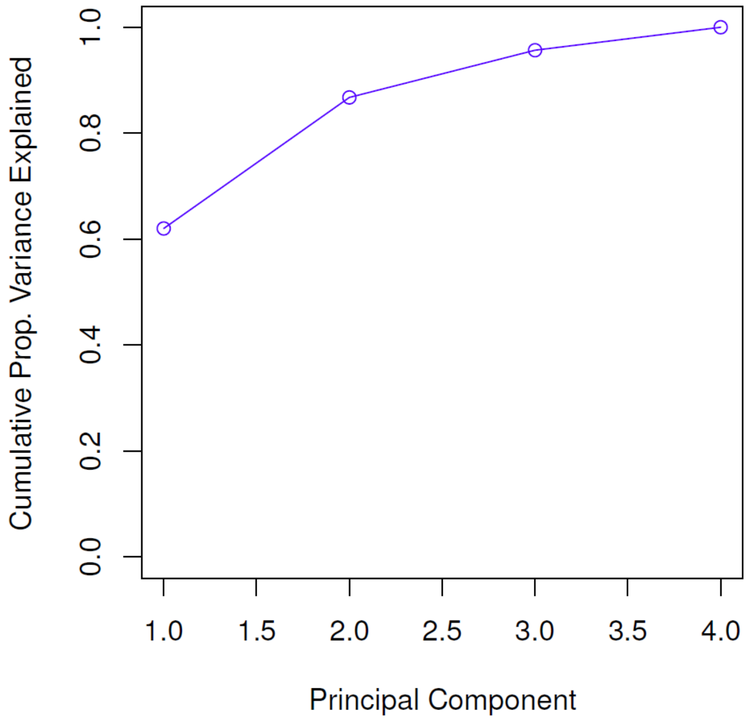
- 主成分分析提供了数据的低维近似
- 在低维近似中，我们损失了多少信息？
- 统计学中用方差刻画信息

$$\underbrace{\sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n}x_{ij}^{2}}_{Var.\,of\ data}=\underbrace{\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d}v_{jk}x_{ij}\right)^{2}}_{Var.\,of\ first\ K\ PCs}+\underbrace{\frac{1}{n}\sum_{j=1}^{d}\sum_{i=1}^{n}\left(x_{ij}-\sum_{k=1}^{K}z_{ik}v_{jk}\right)^{2}}_{MSE\ of\ K-dimensional\ approximation}$$

- 被解释方差的比例（PVE）随M上升

主成分分析





- 左图为每个主成分的PVE，右图为累积PVE

主成分分析

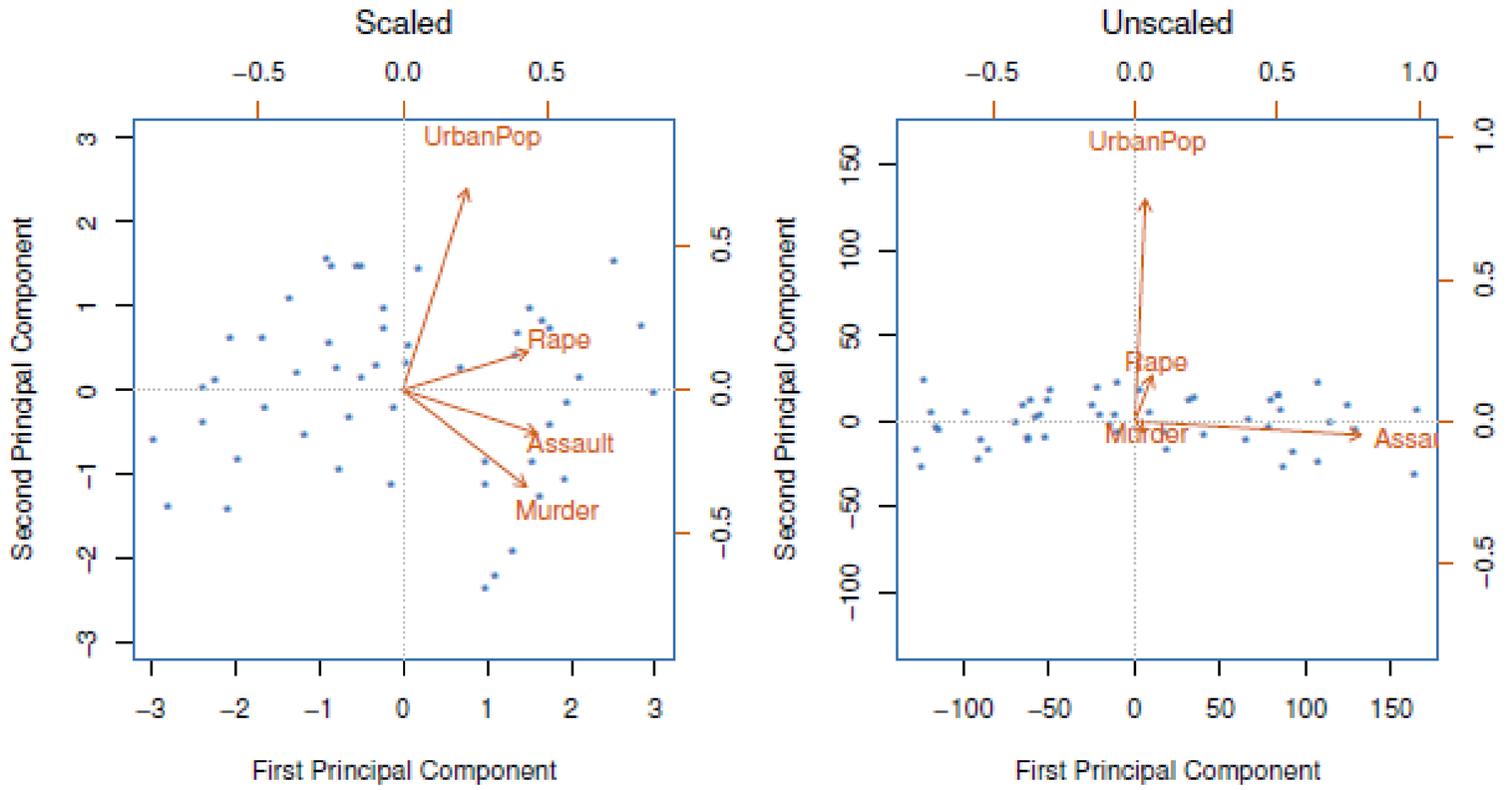


FIGURE 12.4. *Two principal component biplots for the* `USArrests` *data.* Left: *the same as Figure 12.1, with the variables scaled to have unit standard deviations.* Right: *principal components using unscaled data.* `Assault` *has by far the largest loading on the first principal component because it has the highest variance among the four variables. In general, scaling the variables to have standard deviation one is recommended.*

聚类分析

- 聚类分析的目标是在一定准则下将观测数据划分为几类
- 每个数据只能属于其中一类，不能属于多个类，各类之间的数据没有交集
- 设$C_1, \dots, C_K$是表示数据指标的K 个集合，对应数据的K 个类别。假设$C_1 \cup \cdots \cup C_K = \{1, \dots, n\}, C_j \cap C_k = \phi$
- 每个数据所属的集合用潜变量$z_i \in \{1, \dots, K\}$表示

K-均值

- 最常见的聚类分析方法之一，可作为比较基准或其它方法初值。
- 目标函数

$$\min \sum_{k=1}^{K} \sum_{i,i' \in C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 = \min \sum_{k=1}^{K} \sum_{i,i' \in C_k} \|x_i - x_i'\|^2$$

- 等价形式

$$\min \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} \left(x_{ij} - \mu_{kj}\right)^2 = \min \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2$$

- μ_k是第k个聚类的均值

K-均值

1. 初始化：随机分配K个数作为初始均值$\mu_1^{(0)},\ldots,\mu_k^{(0)}$
2. 迭代直到目标函数收敛：
   a）将每个样本点分配到距离最近的聚类

$$z_i^{(t+1)} = argmin_k \left\| x_i - \mu_k^{(t)} \right\|^2$$

   b）计算聚类均值

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} x_i I\left(z_i^{(t+1)} = k\right)}{\sum_{i=1}^{n} I\left(z_i^{(t+1)} = k\right)}$$

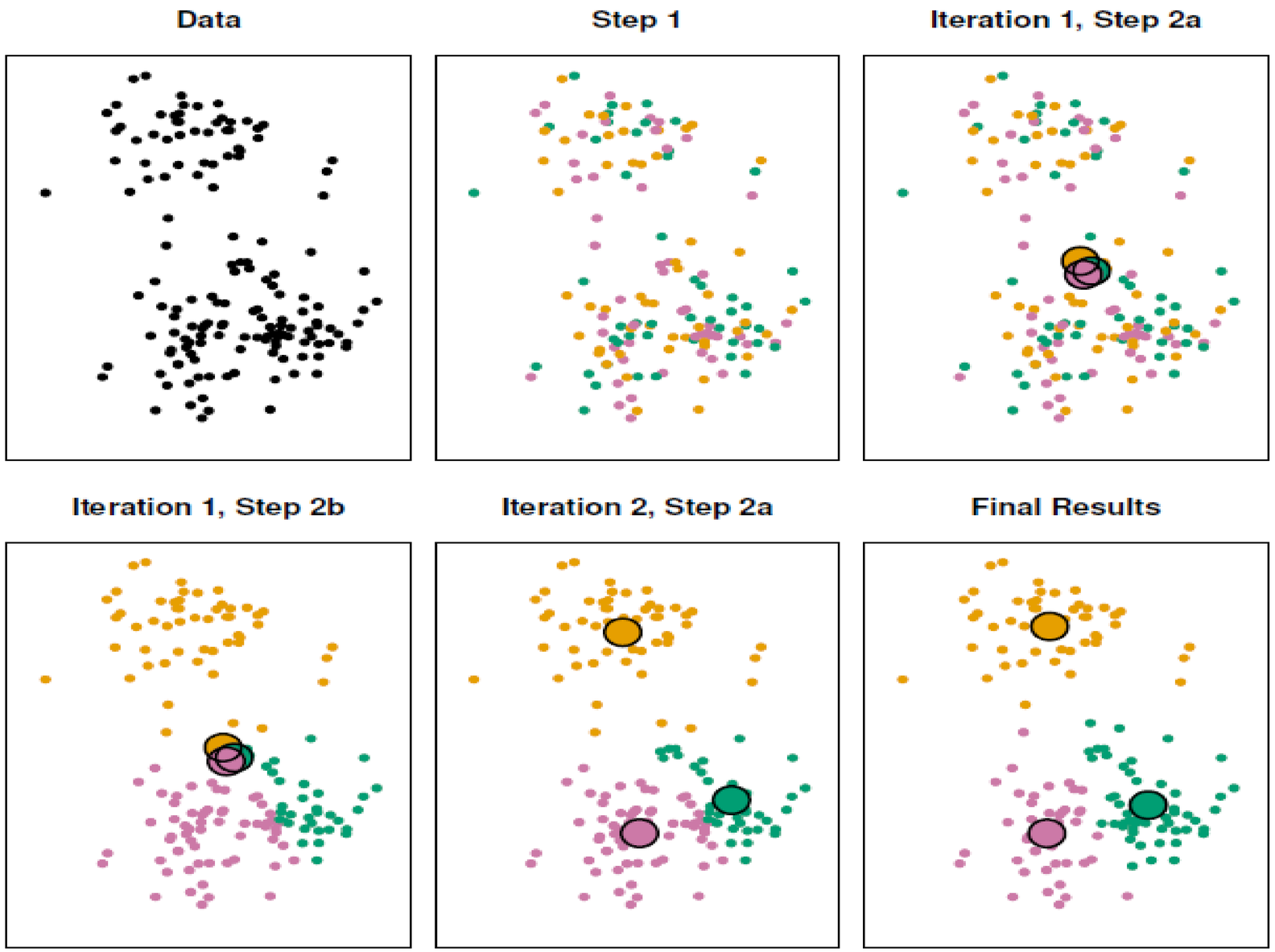
K-均值

- 除了初始化聚类均值，也可以初始化潜变量
- 给每个样本点随机分配1到K中的一个数$z_1^{(0)},\ldots,z_n^{(0)}$
- 计算初始化聚类均值

$$\mu_k^{(0)}=\frac{\sum_{i=1}^{n}x_i I\left(z_i^{(0)}=k\right)}{\sum_{i=1}^{n}I\left(z_i^{(0)}=k\right)}$$

- 重复K-均值算法中的迭代

K–均值

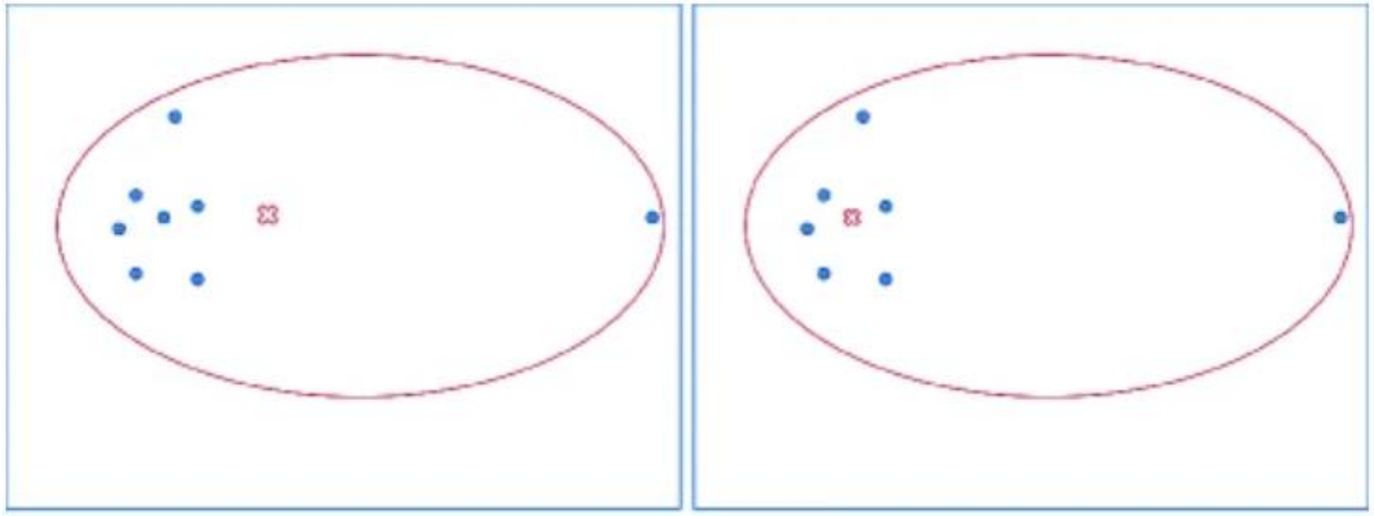

Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

K-均值

- 计算简单高效，每步迭代中目标函数取值不会上升。
- 需要提前给出K的取值
- 不能保证找到全局最优解
- K-均值的结果与初始化相关
- 考虑多个初始值下聚类的比较

K-中心点

- K-均值算法对异常值敏感
- 对异常值敏感是均值的性质和缺点
- 如果希望得到对异常值的稳健性，需要考虑将均值替换为中心点



(a) Mean　　(b) Medoid

- 均值受右端的异常值影响，中心点则几乎不受影响

K-中心点

- K-中心点算法是K-均值算法的推广。
- 使用中心点替代均值，等价于将目标函数中的L_2范数替换为L_1范数

$$\min \sum_{k=1}^{K} \sum_{i,i' \in C_k} \sum_{j=1}^{p} |x_{ij} - x_{i'j}| = \min \sum_{k=1}^{K} \sum_{i,i' \in C_k} \|x_i - x_i'\|_1$$

- 等价形式

$$\min \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} |x_{ij} - \mathrm{m}_{kj}| = \min \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - m_k\|_1$$

- m_k是第k个聚类的中心点

K-中心点

1. 初始化：随机选择数据中K个点作为初始中心点$m_1^{(0)},\dots,m_k^{(0)}$

2. 将每个非中心样本点分配到距离最近的聚类

$$z_i^{(t+1)} = argmin_k \left\| x_i - m_k^{(t)} \right\|_1$$

3. 计算目标函数

$$\sum_{k=1}^{K} \sum_{i \in C_k} \left\| x_i - m_k^{(t)} \right\|_1$$

4. 随机交换非中心样本点x_i与中心点$m_k^{(t)}$，重复步骤2, 3，如果目标函数增加，撤销交换，如果目标函数减小，继续随机交换直到目标函数收敛

K-中心点

- 相比K-均值算法，对异常值点更稳健
- 不适用于非球形分布的聚类分析
- 不能保证找到全局最优解
- K-中心点的结果与初始化相关
- 对大规模数据的表现一般

基于模型的聚类方法

- 假设给定潜变量z_i时，x_i服从特定的分布
- 一般假设$z_i \sim Multinomial(\pi_1, \ldots, \pi_k)$
- 假设给定$z_i = k$时，$x_i \sim N(\mu_k, \Sigma_k)$
- 这给出了高斯混合模型

$$x_i \sim \sum_{k=1}^{K} \pi_k N(\mu_k, \Sigma_k)$$

- 我们感兴趣的参数为$\theta = (\pi_k, \mu_k, \Sigma_k), k = 1, \ldots, K$

高斯混合模型

- 首先，给出似然函数

$$\prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k N(\mu_k,\Sigma_k) \propto \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k \frac{1}{|\Sigma_k|^{1/2}}\exp\left\{-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right\}$$

- 取对数得到

$$\mathrm{l}(\theta|x) = \sum_{i=1}^{n}\log\sum_{k=1}^{K}\pi_k \frac{1}{|\Sigma_k|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right\} + constant$$

- 无法进一步化简从而给出显示解

高斯混合模型

- 引入潜变量z_i，得到完全似然函数

$$\prod_{i=1}^{n}\sum_{k=1}^{K}\left(\pi_k N(\mu_k,\Sigma_k)\right)^{I(z_i=k)}$$

- 取对数得到

$$\mathrm{l}(\theta|x)=\sum_{i=1}^{n}(\log\pi_k+\log N(\mu_k,\Sigma_k))I(z_i=k)+constant$$

- 通过EM算法得到参数估计

高斯混合模型

- 计算E-step：给定当前迭代参数$\theta^{(t)}$

$$
\begin{aligned}
& E_{\mathbf{Z}}[\ell(\boldsymbol{\theta}|\mathbf{x}_i, \mathbf{Z})|\mathbf{X}, \boldsymbol{\theta}^{(t)}] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} E\left[\mathcal{I}(z_i = k)|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}\right]\left(\log \pi_k + \log N(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} P\left(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}\right)\left(\log \pi_k + \log N(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \frac{P\left(\mathbf{x}_i|z_i = k, \boldsymbol{\theta}^{(t)}\right)P\left(z_i = k|\boldsymbol{\theta}^{(t)}\right)}{P\left(\mathbf{x}_i|\boldsymbol{\theta}^{(t)}\right)}\left(\log \pi_k + \log N(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \frac{N(\mathbf{x}_i|\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})\pi_k^{(t)}}{\sum_{j=1}^{K} N(\mathbf{x}_i|\boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})\pi_j^{(t)}}\left(\log \pi_k + \log N(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik}^{(t)}\left(\log \pi_k + \log N(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).
\end{aligned}
$$

高斯混合模型

- 计算M-step:

$$\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

$$= \operatorname{argmax} \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik}^{(t)} \Big(\log \pi_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2} tr\big((\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}\big)\Big)$$

- 对μ_k, Σ_k求导得到 $\dfrac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\mu}_k} = \boldsymbol{\Sigma}_k^{-1}\Big(\sum_{i=1}^{n} \gamma_{ik}^{(t)}(\mathbf{x}_i - \boldsymbol{\mu}_k)\Big) = 0 \implies \boldsymbol{\mu}_k^{(t+1)} = \dfrac{\sum_{i=1}^{n} \gamma_{ik}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{n} \gamma_{ik}^{(t)}}$

$$\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\Sigma}_k^{-1}} = \sum_{i=1}^{n} \gamma_{ik}^{(t)} \boldsymbol{\Sigma}_k - \sum_{i=1}^{n} \gamma_{ik}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})^T = 0$$

$$\implies \boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})^T}{\sum_{i=1}^{n} \gamma_{ik}^{(t)}}$$

高斯混合模型

- 由于$(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)$是标量，因此

$$(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)=tr\left((x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right)$$
$$=tr\left((x_i-\mu_k)(x_i-\mu_k)^T\Sigma_k^{-1}\right)$$

- 求解μ时利用了Σ为正定矩阵

- 求解Σ利用了$\frac{\partial}{\partial\Sigma^{-1}}\log|\Sigma^{-1}|=\Sigma$和

$$\frac{\partial}{\partial\Sigma^{-1}}tr\left((x_i-\mu_k)(x_i-\mu_k)^T\Sigma_k^{-1}\right)=(x_i-\mu_k)(x_i-\mu_k)^T$$

高斯混合模型

- 注意到$\sum_{k=1}^{K}\pi_k = 1$，求解带约束条件的优化问题

$$\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \pi_k} = \sum_{i=1}^{n}\left(\gamma_{ik}^{(t)}\pi_k^{-1} - \gamma_{iK}^{(t)}\pi_K^{-1}\right) = 0$$

$$\Longrightarrow \pi_k^{(t+1)} = \frac{\pi_K^{(t+1)}\sum_{i=1}^{n}\gamma_{ik}^{(t)}}{\sum_{i=1}^{n}\gamma_{iK}^{(t)}}$$

将上式从1到K相加，我们得到

$$\pi_K^{(t+1)} = \frac{\sum_{i=1}^{n}\gamma_{iK}^{(t)}}{n}.$$

同理，有 $\pi_k^{(t+1)} = \frac{\sum_{i=1}^{n}\gamma_{ik}^{(t)}}{n}$.

高斯混合模型

- 容易看出，如果令$\pi_k = \frac{1}{K}, \Sigma_k = \epsilon I, I$是单位矩阵，我们不需要更新$\pi_k$和$\Sigma_k$，则E步中

$$\gamma_{ik} = \frac{N(x_i|\mu_k, \epsilon I)}{\sum_{j=1}^{K} N(x_i|\mu_j, \epsilon I)} = \frac{\exp\left(-\frac{\|x_i - \mu_k\|^2}{2\epsilon}\right)}{\sum_{j=1}^{K} \exp\left(-\frac{\|x_i - \mu_j\|^2}{2\epsilon}\right)}$$

- 当$\epsilon \to 0$，$\gamma_{ik} = 1$当且仅当$k = argmin_j \|x_i - \mu_j\|^2$

- 此时高斯混合模型的EM 算法退化为K 均值算法

本章小节

- 主成分分析
- 聚类分析
- ……