聚类 K-Means、线性降维 PCA

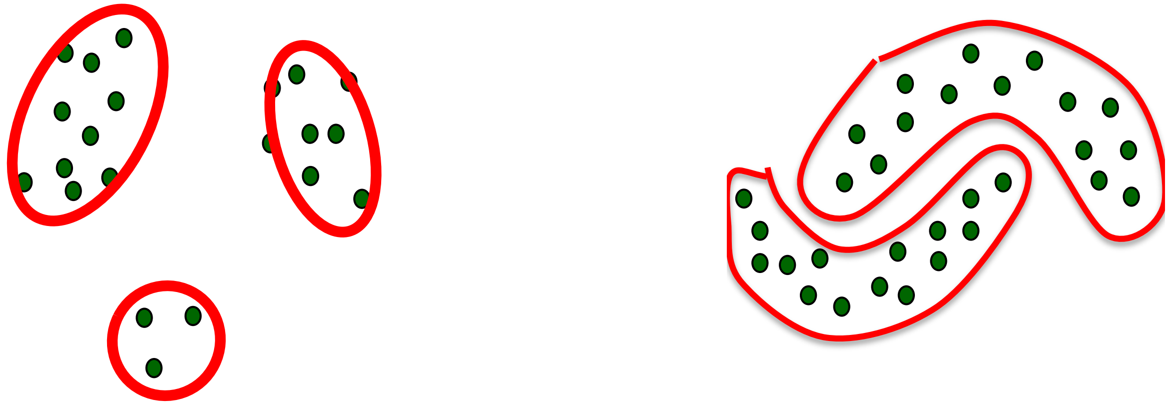
主讲人： 林倞

聚类大纲

- 聚类简介
- K-Means

什么是聚类？

- 给定一组数据实例 $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ ，聚类就是如何将它们分成不同的簇

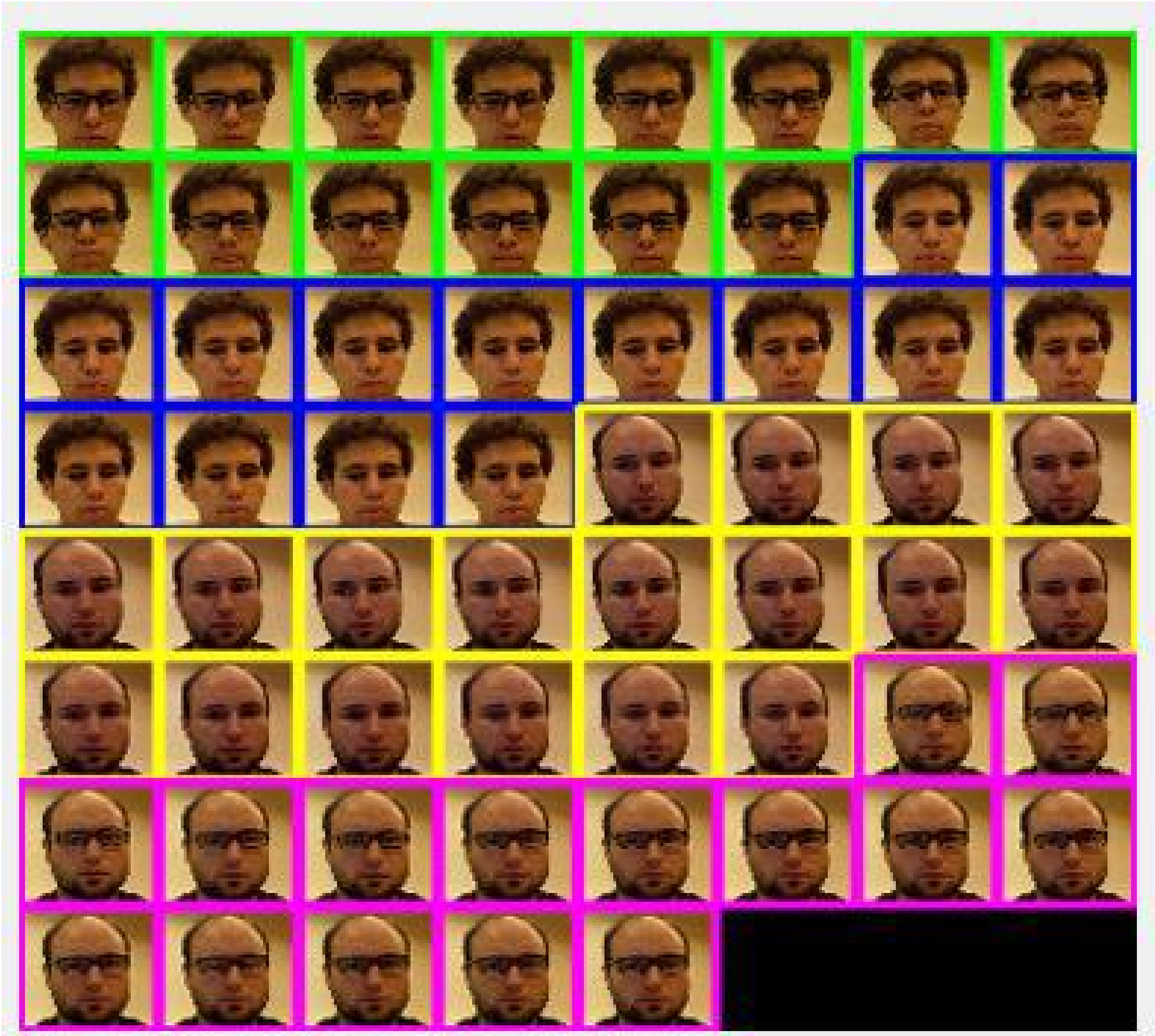


- 目标：
  - 簇内实例具有高相似度
  - 簇间实例具有低相似度

相似性准则很重要

- 不同的相似性准则可能导致不同的结果

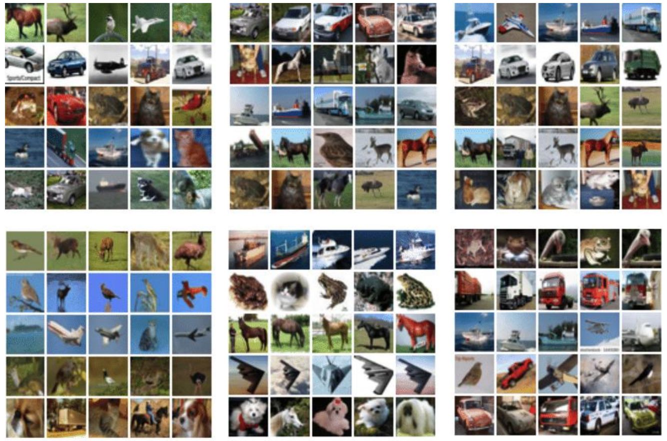
相似还是不相似？



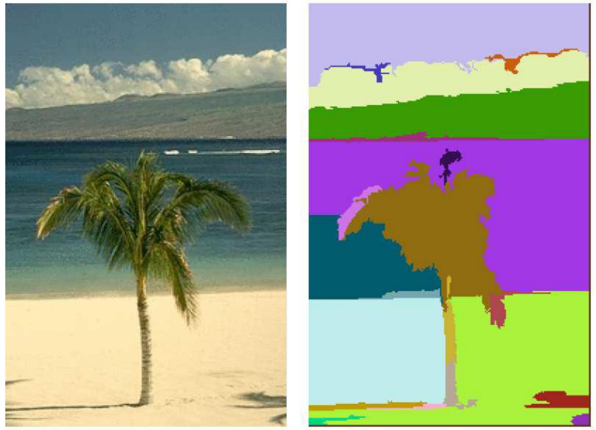
准则1：身份　　　　　　准则2：眼镜

实际应用

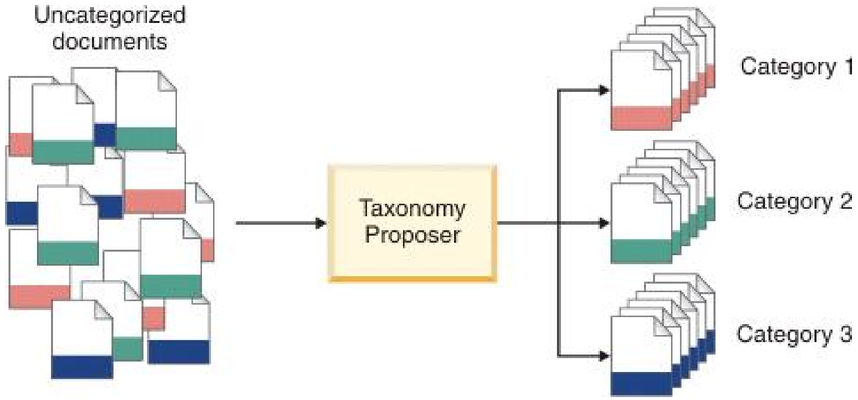
- 图像分组


(a) (b) (c)

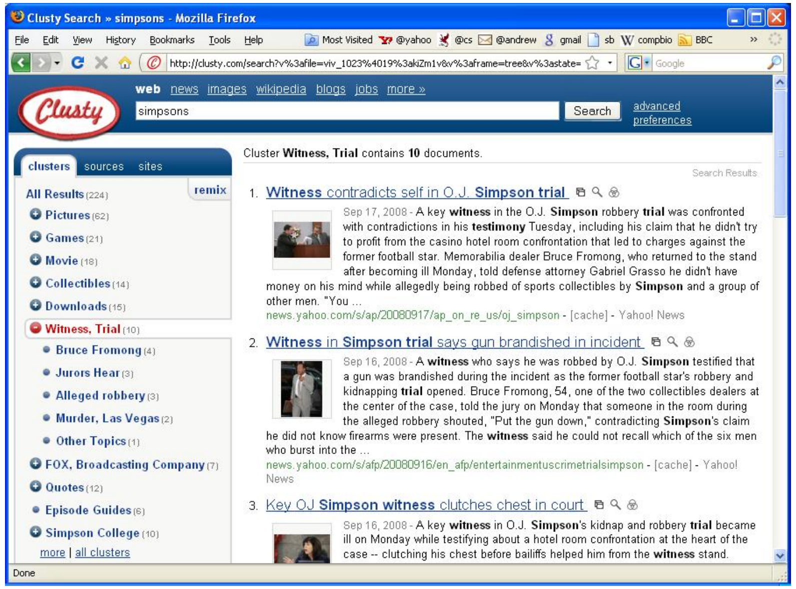
- 图像分割



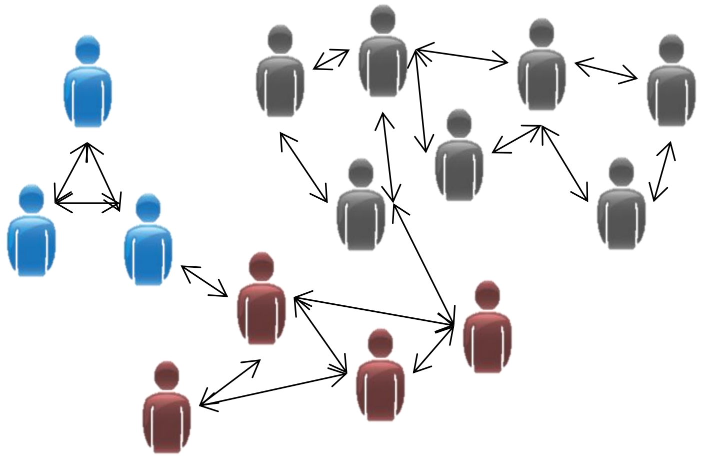
- 自动将语义相似的文档分组在一起



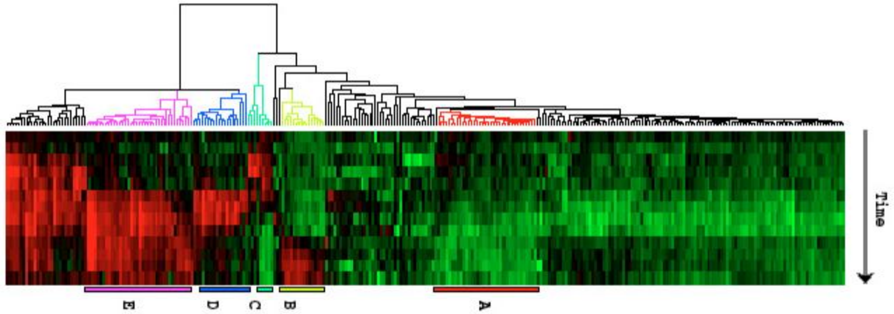
- 网络搜索结果聚类



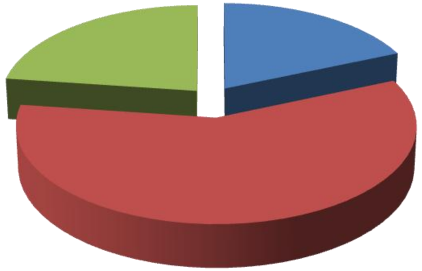
- 社交网络分析



- 基因表达数据聚类



- 市场细分

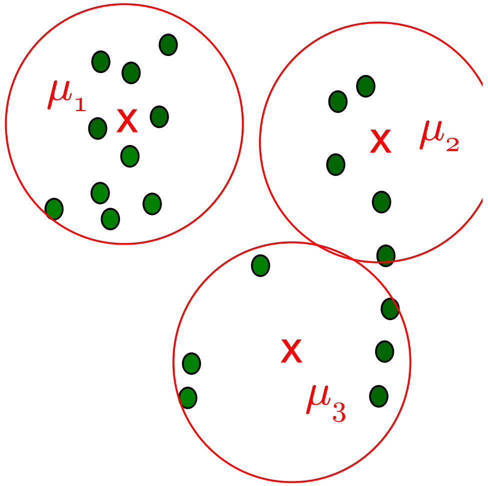


聚类大纲

- 聚类简介
- K-Means

K-均值算法

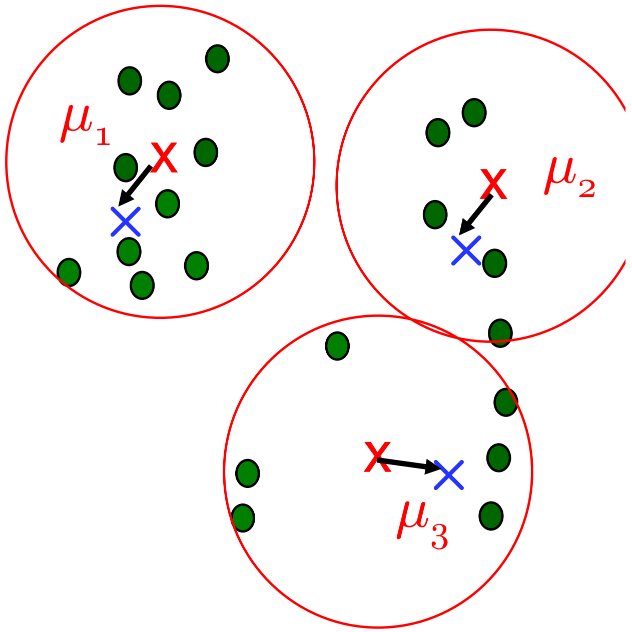
- 指定 K 个中心 $\boldsymbol{\mu}_k$，其中 $k=1,\cdots,K$，然后评估每个数据点 $\boldsymbol{x}^{(n)}$ 与所有中心 $\boldsymbol{\mu}_k$ 之间的距离



- 数据点 $\boldsymbol{x}^{(n)}$ 被分配给距离其最近的簇 k

$$r_{nk}=\begin{cases}1, & if\ k=\arg\min_j \left\|\boldsymbol{x}^{(n)}-\boldsymbol{\mu}_j\right\|^2\\ 0, & otherwise\end{cases}$$

- 使用簇内样本的均值来更新中心



两个问题：

1) 这个算法到底在做什么？

2) 这个算法能保证收敛吗？

$$\boldsymbol{\mu}_k \leftarrow \frac{\sum_{n=1}^{N} r_{nk}\, \boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}}$$

- 重复上述分配和中心更新步骤

收敛性保证

- 定义一个目标函数，它是所有数据实例与其对应中心之间距离的总和：

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k \right\|^2$$

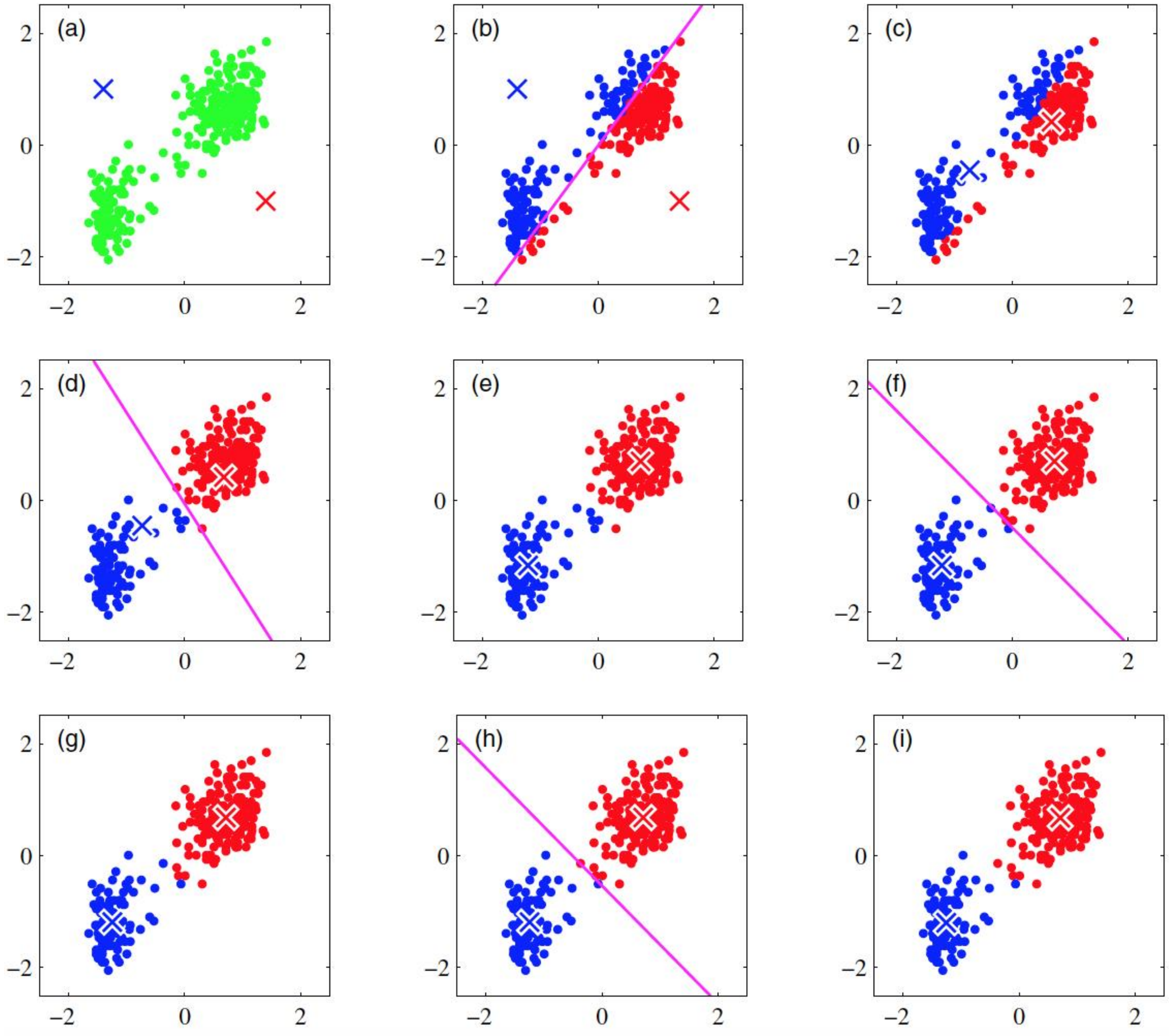
- K-均值可以通过以*交替方式*更新 $\boldsymbol{r}_n$ 和 $\boldsymbol{\mu}_k$ 的以下优化问题得到：

$$\min_{\boldsymbol{r}_n, \boldsymbol{\mu}_k} J$$

$$s.t.: \quad \boldsymbol{r}_n \in \text{onehot vector} \quad \forall\, n \,\&\, k$$

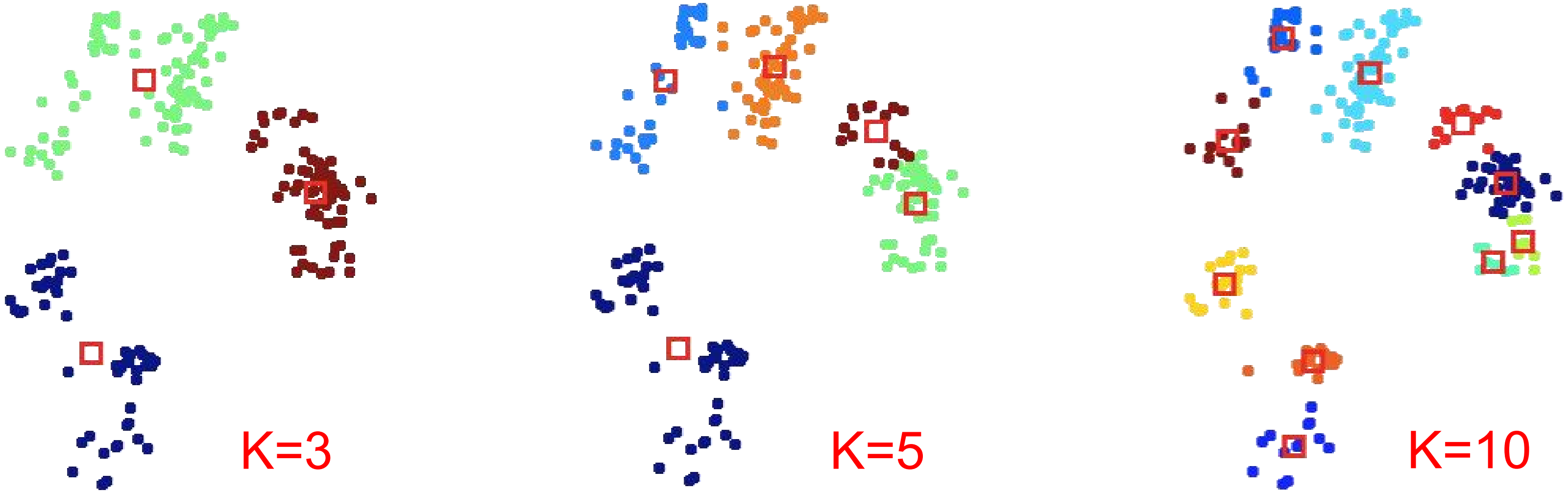
其中，$\boldsymbol{r}_n \triangleq [r_{n1}, r_{n2}, \cdots, r_{nK}]$ 需为独热向量

- 总距离 J *单调递减*，因此可以保证 K-均值算法收敛

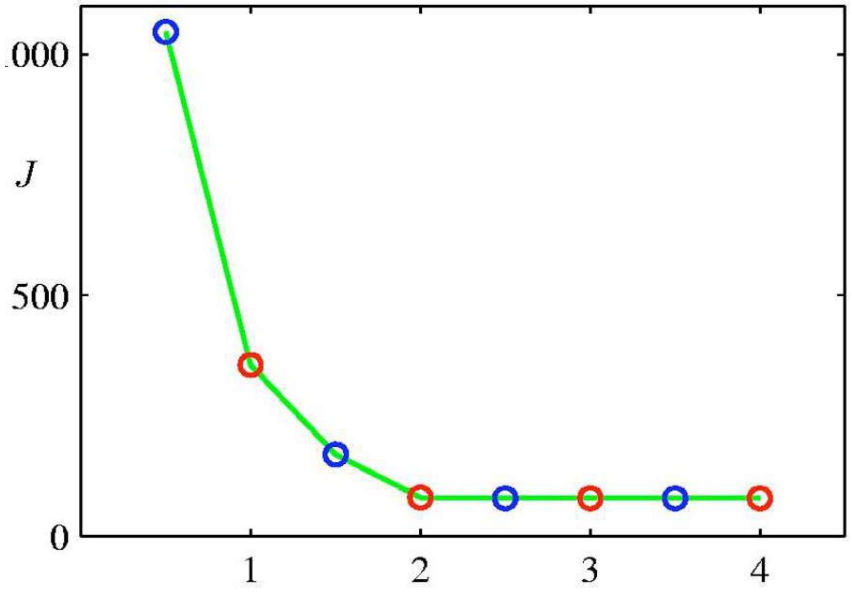

(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
2
0
−2
−2
0
2

问题：簇的数量

- 如何设置 K 的值对于最终的聚类结果至关重要



- 随着簇数量 K 的增加，距离 J 被确定为递减。因此，K 不能通过最小化 J 来确定

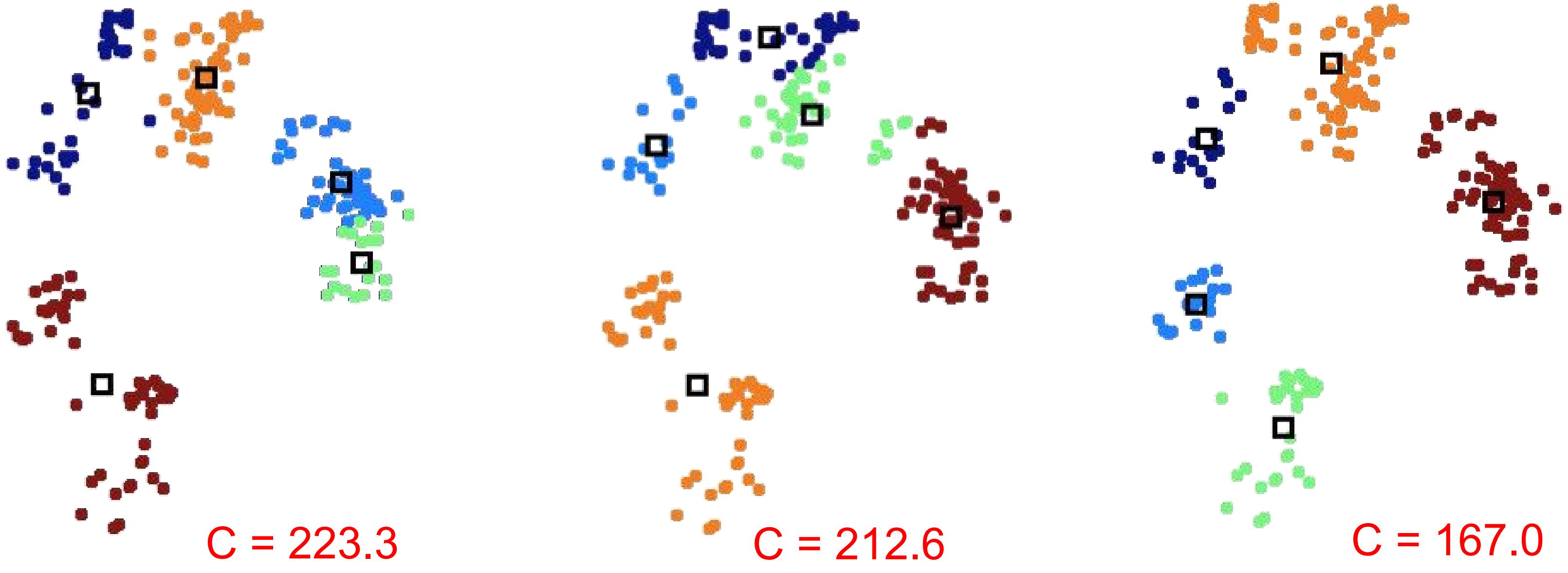


1) 一种可能的方法是选择拐点（这里 K=2）

2) 另一种可能的方法是根据下游应用的性能来确定最佳的 K 值

问题：初始化

- K-均值的性能也高度依赖于初始中心点



1) 随机方法

- 随机选择数据实例作为初始化
- 问题：可能会选择相邻的实例

2) 基于距离的方法

- 从一个随机数据实例开始
- 选择距离现有中心最远的点
- 问题：可能会选择离群值

3) 随机 + 距离方法

- 从一个随机数据实例开始
- 从其余距离现有中心较远的实例中随机选择下一个中心

问题：硬分配

- 硬分配

  一个数据实例要么属于某个簇，要么不属于，这种分配方式是确定性的，即 $\boldsymbol{r}_n$ 必须是一个独热向量

- 软 K-均值

  软 K-均值不是确定性地将 $\boldsymbol{x}^{(n)}$ 分配给某个簇，而是以软性方式进行分配

β 控制了分布的尖锐程度

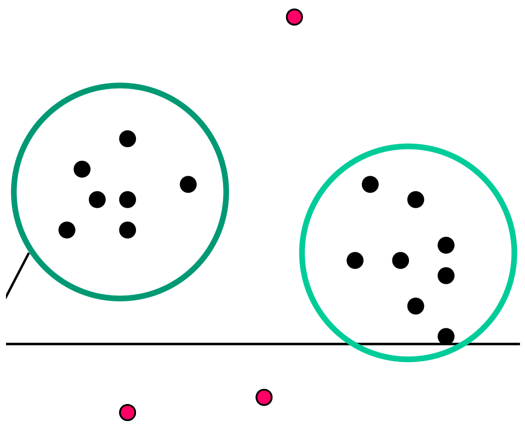
$$r_{nk} = \frac{e^{-\beta\left\|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_k\right\|^2}}{\sum_{i=1}^{K} e^{-\beta\left\|\boldsymbol{x}^{(n)} - \boldsymbol{\mu}_i\right\|^2}}$$

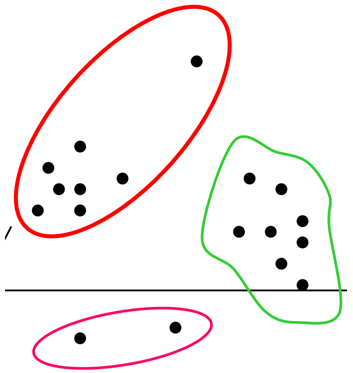
$$\boldsymbol{\mu}_k \leftarrow \frac{\sum_{n=1}^{N} r_{nk}\,\boldsymbol{x}_n}{\sum_{n=1}^{N} r_{nk}}$$

r_{nk}可以被解释为数据点 $\boldsymbol{x}^{(n)}$ 属于簇 k 的概率

其他问题

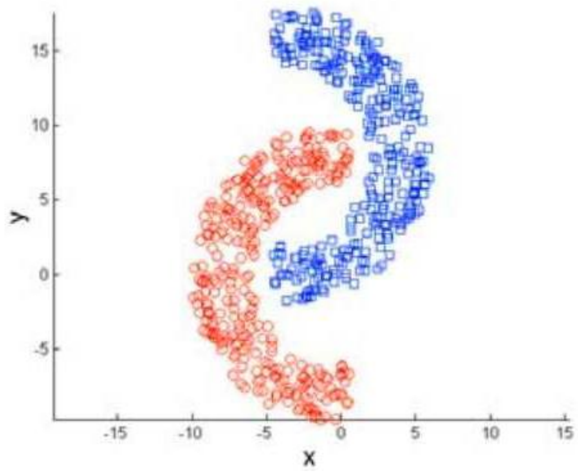
- 对离群值敏感





- 圆形形状

欧氏距离意味着边界只能是球形的。当簇具有不规则形状时，性能会很差



课堂小结

- 聚类的目标——簇内高相似、簇间低相似
- K-Means的算法
  - 算法过程
  - 等价的优化目标函数
  - 存在的问题，如：簇数量的确定、初始化、硬分配等
- 软K-Means

线性降维PCA大纲

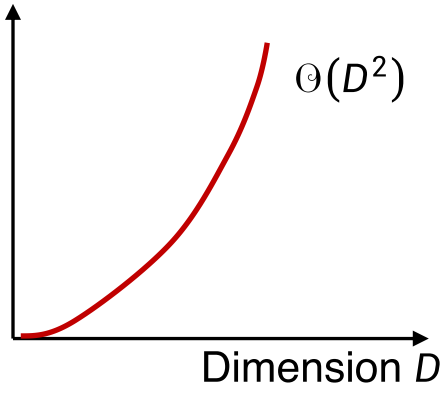
- 动机
- 视角 1: 最小化重构误差
- 视角 2: 最大化方差
- 视角 3: 奇异值分解（SVD）
- PCA的其他应用

动机

- 许多类型数据的维度非常高，例如下面图像的维度高达

$$256 \times 256 = 65536$$

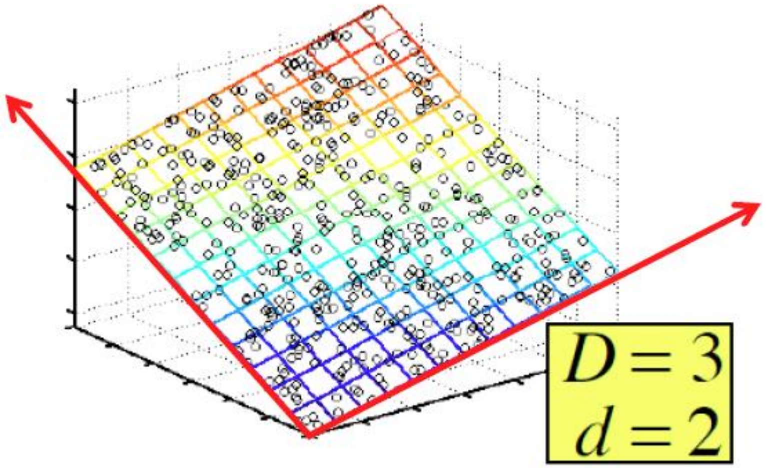
- 如果我们直接在原始数据上工作，后续任务（例如，分类）的复杂性可能会非常高。



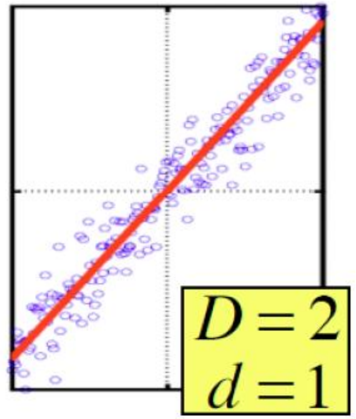
逻辑回归的复杂度

降低数据的维度为什么可行？

- 高维数据通常近似存在于一个低维的内在空间上



3维数据近似存在于一个2维平面上

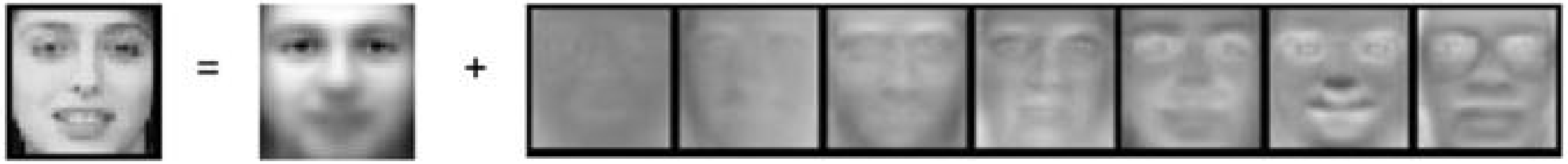


2维数据近似存在于一个1维直线上

关键在于如何找到主方向，从而使得数据样本的表示误差尽可能小

- 对于真实世界的数据，也可以用低维来表示

例如：如果找到合适的“方向”，人脸可以用很少的几个值很好地表示



$$\boldsymbol{x} \approx \boldsymbol{\mu}_0 + a_1\boldsymbol{\mu}_1 + \cdots + a_7\boldsymbol{\mu}_7$$

具有65536个值的原始图像 $\boldsymbol{x}$ 可以仅用7个值$a_1, \cdots a_7$表示

线性降维PCA大纲

- 动机
- 视角 1：最小化重构误差
- 视角 2：最大化方差
- 视角 3：奇异值分解（SVD）
- PCA的其他应用

新方向下的样本重表示

- 如何在高维空间中表示正交方向？

一组满足以下条件的向量 $\boldsymbol{u}_i$

$$\boldsymbol{u}_i^T \boldsymbol{u}_j = \delta_{ij}$$

如果$i = j$，$\delta_{ij} = 1$; 否则 $\delta_{ij} = 0$

定理： 在给定 M 个正交方向 $\{\boldsymbol{u}_i\}_{i=1}^M$ 的情况下，对数据样本 $\boldsymbol{x}$ 的最佳近似为

$$\tilde{\boldsymbol{x}} = a_1 \boldsymbol{u}_1 + a_2 \boldsymbol{u}_2 + \cdots + a_M \boldsymbol{u}_M$$

其中 a_i 等于

$$a_i = \boldsymbol{u}_i^T \boldsymbol{x}$$

证明：

$$\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|^2=\left\|\boldsymbol{x}-\sum_{i=1}^{M}a_i\boldsymbol{u}_i\right\|^2$$

$$=\|\boldsymbol{x}\|^2-2\sum_{i=1}^{M}a_i\boldsymbol{u}_i^T\boldsymbol{x}+\sum_{i=1}^{M}a_i^2$$

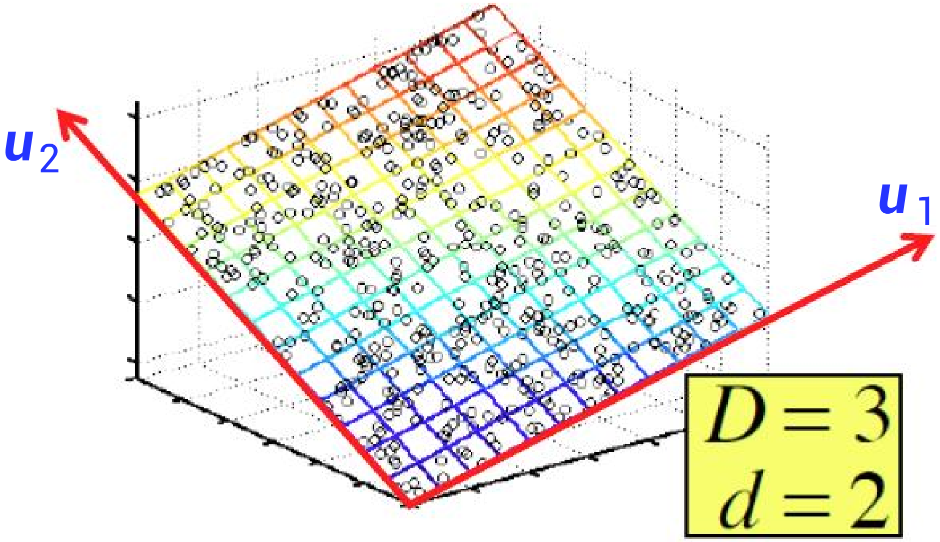
若$i\neq j$， $\boldsymbol{u}_i^T\boldsymbol{u}_j=0$； 若$i=j$， $\boldsymbol{u}_i^T\boldsymbol{u}_j=1$

这是一个二次函数，当 $a_i=\boldsymbol{u}_i^T\boldsymbol{x}$ 时可以最小化

给定方向 $\{\boldsymbol{u}_i\}_{i=1}^{M}$ ，最佳系数为 $a_i=\boldsymbol{u}_i^T\boldsymbol{x}$。但是如何找到最佳方向仍然未知

寻找最佳方向

- **目标：** 给定来自 $\mathbb{R}^D$ 的数据样本 $\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$ ，找到 M 个正交方向 $\boldsymbol{u}_i$ ，使得原始数据可以在这些方向下得到最佳表示



$$\boldsymbol{x}^{(n)} \approx \sum_{i=1}^{M} a_i^{(n)} \boldsymbol{u}_i$$

- 如果假设给定了最佳方向 $\{\boldsymbol{u}_i\}_{i=1}^{M}$，那么最佳系数 $a_i^{(n)}$ 是什么？

$$a_i^{(n)} = \boldsymbol{u}_i^T \boldsymbol{x}^{(n)}$$

不直接表示数据 $\boldsymbol{x}^{(n)}$，我们首先将数据中心化到原点，即从每个数据点 $\boldsymbol{x}^{(n)}$ 中减去其均值

$\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}}$， 其中 $\overline{\boldsymbol{x}}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}^{(n)}$

- 现在的目标可以描述为——最小化$\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}}$与其在方向$\{\boldsymbol{u}_i\}_{i=1}^{M}$下的最佳表示 $\sum_{i=1}^{M} a_i^{(n)}\boldsymbol{u}_i$之间的误差

$$E \triangleq \frac{1}{N}\sum_{n=1}^{N}\left\|\left(\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}}\right)-\sum_{i=1}^{M} a_i^{(n)}\boldsymbol{u}_i\right\|^2$$

即：解如下优化问题

$$Min_{a_i^{(n)},\,\boldsymbol{u}_i}\ E,$$

其中系数 $a_i^{(n)}$ 的最佳取值已经知道等于

$$a_i^{(n)}=\boldsymbol{u}_i^T\left(\boldsymbol{x}^{(n)}-\overline{\boldsymbol{x}}\right)$$

- 化简重建误差 E

a) 注意到$\boldsymbol{u}_i^T \boldsymbol{u}_j = \delta_{ij}$ ， 重建误差$E = \frac{1}{N}\sum_{n=1}^{N} \left\| (\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}) - \sum_{i=1}^{M} a_i^{(n)} \boldsymbol{u}_i \right\|^2$可表示成

$$E = \frac{1}{N}\left(\sum_{n=1}^{N} \left\|\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\right\|^2 - 2\sum_{n=1}^{N}\sum_{i=1}^{M} a_i^{(n)} \left(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\right)^T \boldsymbol{u}_i + \sum_{n=1}^{N}\sum_{i=1}^{M} \left(a_i^{(n)}\right)^2\right)$$

b) 代入$a_i^{(n)} = \boldsymbol{u}_i^T \left(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\right)$，可得

$$E = \frac{1}{N}\sum_{n=1}^{N} \left\|\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\right\|^2 - \sum_{i=1}^{M} \boldsymbol{u}_i^T \underbrace{\frac{1}{N}\sum_{n=1}^{N} \left(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\right)^T}_{\triangleq \boldsymbol{S}} \boldsymbol{u}_i$$

c) 将其写成矩阵形式给出

$$E=\frac{1}{N}\|\boldsymbol{X}-\overline{\boldsymbol{X}}\|_F^2-\sum_{i=1}^{M}\boldsymbol{u}_i^T\boldsymbol{S}\boldsymbol{u}_i$$

其中$\boldsymbol{X}\triangleq\left[\boldsymbol{x}^{(1)},\boldsymbol{x}^{(2)},\cdots,\boldsymbol{x}^{(N)}\right]$，$\|\cdot\|_F$ 表示Frobenius 范数

- 在约束 $\boldsymbol{u}_i^T\boldsymbol{u}_j=\delta_{ij}$下，最小化 $E=\frac{1}{N}\|\boldsymbol{X}-\overline{\boldsymbol{X}}\|_F^2-\frac{1}{\mathrm{N}}\sum_{i=1}^{M}\boldsymbol{u}_i^T\boldsymbol{S}\boldsymbol{u}_i$可等价表示为

$$\boxed{\begin{aligned}&\max_{\boldsymbol{u}_1\cdots\boldsymbol{u}_M}\sum_{i=1}^{M}\boldsymbol{u}_i^T\boldsymbol{S}\boldsymbol{u}_i\\&\qquad\qquad s.t.:\boldsymbol{u}_i^T\boldsymbol{u}_j=\delta_{ij}\end{aligned}}$$

- 考虑 $M = 1$ 的简单情况。问题简化为：

$$\max_{\boldsymbol{u}_1} \boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_1$$

$$s.t.: \boldsymbol{u}_1^T \boldsymbol{u}_1 = 1$$

  - 这等价于最大化

$$\boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_1 - \lambda(\boldsymbol{u}_1^T \boldsymbol{u}_1 - 1)$$

  - 对 $\boldsymbol{u}_1$ 求导并将其设置为 0 给出

$$\boldsymbol{S}\boldsymbol{u}_1 = \lambda \boldsymbol{u}_1,$$

从中我们可以看到$\boldsymbol{u}_1$应该是**S的特征向量**

  - 可以进一步验证，对应于最大特征值的特征向量可以最大化 $\boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_1$

- 对于 $M = 2$ 的情况，问题变为

$$\max_{\boldsymbol{u}_1, \boldsymbol{u}_2} \ \boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_1 + \boldsymbol{u}_2^T \boldsymbol{S} \boldsymbol{u}_2$$

$$s.t.: \boldsymbol{u}_1^T \boldsymbol{u}_1 = 1, \ \boldsymbol{u}_2^T \boldsymbol{u}_2 = 1, \ \boldsymbol{u}_1^T \boldsymbol{u}_2 = 0$$

  - 这等价于最大化

$$\boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_1 - \lambda_1 (\boldsymbol{u}_1^T \boldsymbol{u}_1 - 1) + \boldsymbol{u}_2^T \boldsymbol{S} \boldsymbol{u}_2 - \lambda_2 (\boldsymbol{u}_2^T \boldsymbol{u}_2 - 1)$$

约束为 $\boldsymbol{u}_1^T \boldsymbol{u}_2 = 0$

  - 对 $\boldsymbol{u}_1$ 和 $\boldsymbol{u}_2$ 求导并将其设置为 0 给出

$$\boldsymbol{S} \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1, \qquad \boldsymbol{S} \boldsymbol{u}_2 = \lambda_2 \boldsymbol{u}_2,$$

$\Rightarrow$ $\boldsymbol{u}_1$ 和 $\boldsymbol{u}_2$ 必须是S的特征向量

$\Rightarrow$ 可以看出，为了最大化 $\boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_1 + \boldsymbol{u}_2^T \boldsymbol{S} \boldsymbol{u}_2$， $\boldsymbol{u}_1$ 和 $\boldsymbol{u}_2$ 必须是对应于两个最大特征值的特征向量

对于 $M>1$ 的情况，方向 $\boldsymbol{u}_i$ 是 $\boldsymbol{S}$ 对应于最大的 M 个特征值的特征向量

$\boldsymbol{A}=\boldsymbol{X}-\overline{\boldsymbol{X}}$

问题：对于 $i \neq j$ ，$\boldsymbol{S}$ 的特征向量 $\boldsymbol{u}_i$ 是否满足 $\boldsymbol{u}_i^T \boldsymbol{u}_j=0$?

- 对于任何 $D \times D$ 实对称矩阵（例如 $\boldsymbol{S} \triangleq \boldsymbol{A}\boldsymbol{A}^T$ ），它都有 D 个彼此正交的特征向量

- 对于每个$\boldsymbol{S} \triangleq \boldsymbol{A}\boldsymbol{A}^T$ ，它可以分解为

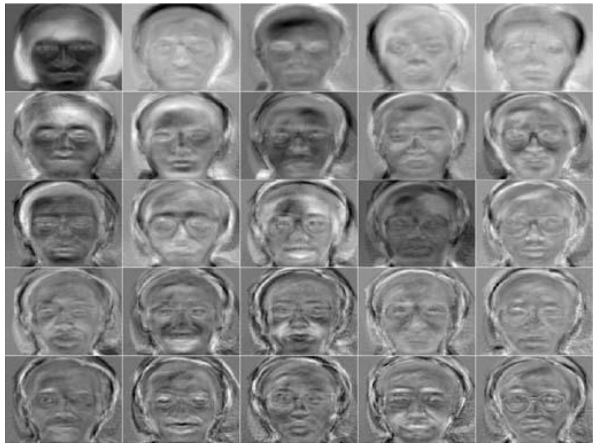
$$\boldsymbol{S}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

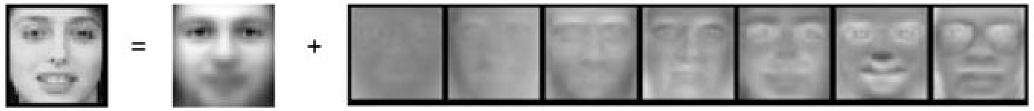
其中 $\boldsymbol{U}$ 由 $\boldsymbol{S}$ 的特征向量组成，且 $\boldsymbol{U}\boldsymbol{U}^T=I$; $\boldsymbol{\Lambda}$ 是由 $\boldsymbol{S}$ 的特征值组成的对角矩阵

示例

输入数据：每张人脸
图像是一个数据点

前 25 个主方向





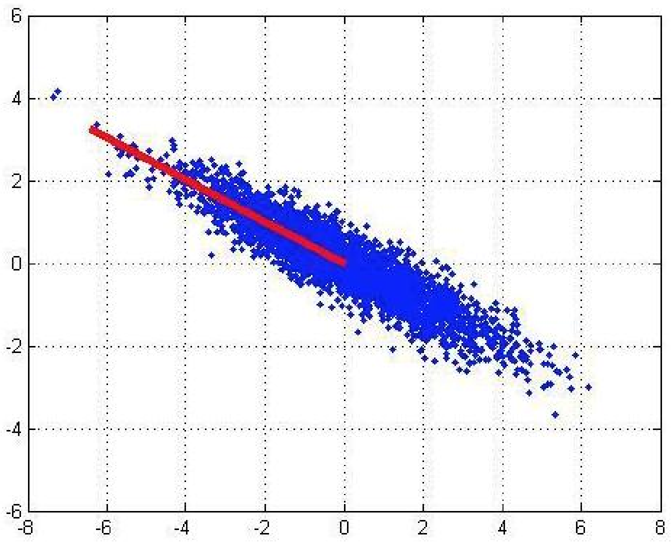
$$\boldsymbol{x} \approx \overline{\boldsymbol{x}} + a_1 \boldsymbol{u}_1 + \cdots + a_7 \boldsymbol{u}_7$$

线性降维PCA大纲

- 动机
- 视角1：最小化重构误差
- 视角2：最大化方差
- 视角3：奇异值分解（SVD）
- 主成分分析（PCA）的其他应用

问题公式化

- **目标:** 给定一个数据集 $\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$，找到正交方向 $\{\boldsymbol{u}_k\}_{k=1}^{M}$，使得数据在这些方向上的投影的方差最大化



最大化方差相当于尽可能多地保留原始数据的信息

- 对于第一个方向 $\boldsymbol{u}_1$，我们希望沿方向 $\boldsymbol{u}_1$ 投影的数据的方差最大化，即 $\{\boldsymbol{u}_1^T\boldsymbol{x}^{(n)}\}_{n=1}^N$最大化

  - 方差表达式

$$var = \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{u}_1^T(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}})\right)^2$$

$$= \boldsymbol{u}_1^T\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}})(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}})^T\boldsymbol{u}_1$$

$$= \boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1$$

  - 如先前证明的，在满足 $\boldsymbol{u}_1^T\boldsymbol{u}_1 = 1$的约束条件下，当 $\boldsymbol{u}_1$是 **S** 对应于最大特征值的特征向量时，方差最大化

- 对于第二个方向 $\boldsymbol{u}_2$，它也应该最大化方差

$$var = \boldsymbol{u}_2^T \boldsymbol{S} \boldsymbol{u}_2,$$

但应满足约束条件 $\boldsymbol{u}_i^T \boldsymbol{u}_j = \delta_{ij}$，即：

$$\boldsymbol{u}_2^T \boldsymbol{u}_2 = 1 \qquad \boldsymbol{u}_1^T \boldsymbol{u}_2 = 0$$

- 由于 $\boldsymbol{u}_1$ 是对应于最大特征值的特征向量，可以证明 $\boldsymbol{u}_2$ 是 $\boldsymbol{S}$ 对应于第二大特征值的特征向量

$\boldsymbol{u}_i$ 是 $\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}})(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}})^T$ 的对应于第 i 大特征值的特征向量

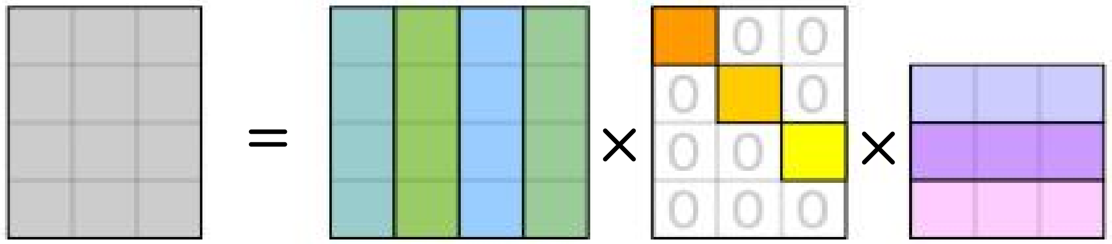
线性降维PCA大纲

- 动机
- 角度 1：最小化重构误差
- 角度 2：最大化方差
- 视角 3：奇异值分解（SVD）
- PCA的其他应用

奇异值分解 (SVD)

- 对于任何 $M \times N$ 矩阵 $\boldsymbol{A}$，它总是可以分解为

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$



$\boldsymbol{A}$	$\boldsymbol{U}$	$\boldsymbol{\Sigma}$	$\boldsymbol{V}^T$
$M \times N$	$M \times M$	$M \times N$	$N \times N$

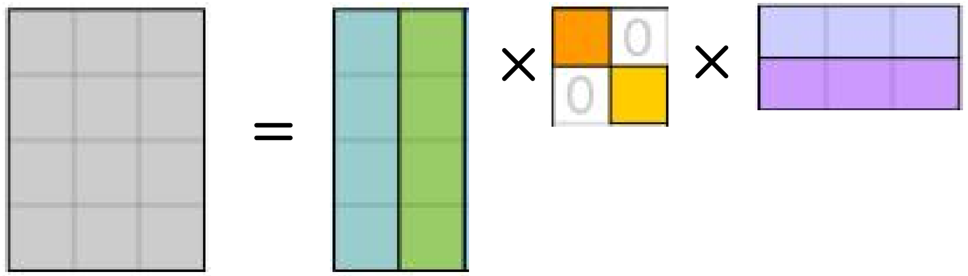
  - $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_M]$、$\boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N]$，其中 $\boldsymbol{u}_i$和 $\boldsymbol{v}_i$ 分别是 $\boldsymbol{A}\boldsymbol{A}^T$ 和 $\boldsymbol{A}^T\boldsymbol{A}$ 的第 i 个特征向量，并且 $\boldsymbol{u}_i^T\boldsymbol{u}_j = \delta_{ij}$,$\boldsymbol{v}_i^T\boldsymbol{v}_j = \delta_{ij}$
  - $\boldsymbol{\Sigma}$ 仅在对角线上有非零值，这些值是 $\boldsymbol{A}\boldsymbol{A}^T$ 或 $\boldsymbol{A}^T\boldsymbol{A}$ 的特征值的平方根（ $\boldsymbol{A}\boldsymbol{A}^T$ 和 $\boldsymbol{A}^T\boldsymbol{A}$ 的非零特征值相同）

    对角元素 $\boldsymbol{\Sigma}_{ii}$ 称为奇异值，并按降序排列

- 因为 $\boldsymbol{\Sigma}$ 仅在对角线上有非零值，所以 $\boldsymbol{A}$ 可以表示为

$$\boldsymbol{A} = \sum_{i=1}^{r} \Sigma_{ii} \boldsymbol{u}_i \boldsymbol{v}_i^T = \boldsymbol{U}'\boldsymbol{\Sigma}'\boldsymbol{V}'^T$$

其中 $\boldsymbol{u}_i$ 和 $\boldsymbol{v}_i$ 是 $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的第 i 列； r 是 $\boldsymbol{\Sigma}$ 中非零对角元素的数量



$\boldsymbol{A}$ $\quad$ $\boldsymbol{U}'$ $\quad$ $\boldsymbol{\Sigma}'$ $\quad$ $\boldsymbol{V}'^T$

$M \times N$ $\quad$ $M \times r$ $\quad$ $r \times r$ $\quad$ $r \times N$

$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} = \begin{bmatrix} \mathbf{0.13} & 0.02 & -0.01 \\ \mathbf{0.41} & 0.07 & -0.03 \\ \mathbf{0.55} & 0.09 & -0.04 \\ \mathbf{0.68} & 0.11 & -0.05 \\ 0.15 & \mathbf{-0.59} & \mathbf{0.65} \\ 0.07 & \mathbf{-0.73} & \mathbf{-0.67} \\ 0.07 & \mathbf{-0.29} & \mathbf{0.32} \end{bmatrix} \begin{bmatrix} \mathbf{12.4} & 0 & 0 \\ 0 & \mathbf{9.5} & 0 \\ 0 & 0 & \mathbf{1.3} \end{bmatrix} \begin{bmatrix} \mathbf{0.56} & \mathbf{0.59} & \mathbf{0.56} & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & \mathbf{-0.69} & \mathbf{-0.69} \\ 0.40 & \mathbf{-0.80} & 0.40 & 0.09 & 0.09 \end{bmatrix}$$

- 在 $\boldsymbol{A}$ 的奇异值分解（SVD）中，向量 $\boldsymbol{u}_i$ 是矩阵 $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ 对应于其第 i 大特征值的特征向量

- 通过设置 $\boldsymbol{A} = \tilde{\boldsymbol{X}}$ ，其中 $\tilde{\boldsymbol{X}} \triangleq [\boldsymbol{x}^{(1)} - \overline{\boldsymbol{x}}, \boldsymbol{x}^{(2)} - \overline{\boldsymbol{x}}, \cdots, \boldsymbol{x}^{(N)} - \overline{\boldsymbol{x}}]$ ，我们可以看到

$$\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T = \sum_{n=1}^{N} \left(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}^{(n)} - \overline{\boldsymbol{x}}\right)^T$$

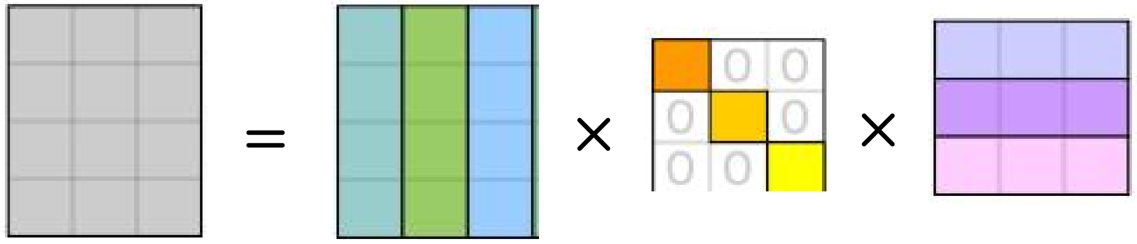
$$= N \cdot \boldsymbol{S}$$

$\Longrightarrow \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T$ 的特征向量与矩阵 $\boldsymbol{S}$ 相同

通过对 $\tilde{\boldsymbol{X}}$ 执行 SVD，可以直接获得数据 $\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$ 的主方向，这些主方向是矩阵 $\boldsymbol{U}$ 的列

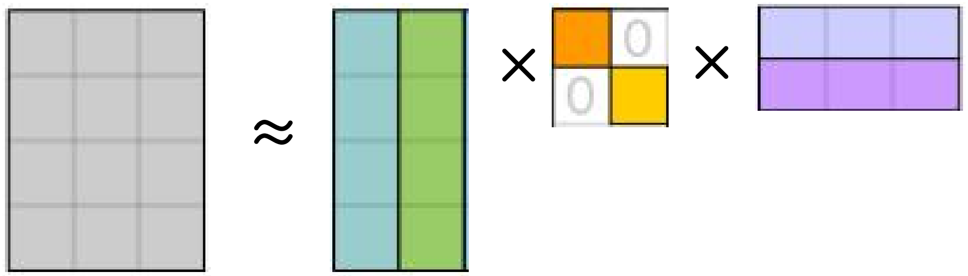
- 问题:

给定 $\tilde{\boldsymbol{X}}$ 的 SVD 分解如下所示， $\tilde{\boldsymbol{x}}^{(n)}$ 的主方向和系数 a_i's是什么？



$\tilde{\boldsymbol{X}}$ $\boldsymbol{U}$ $\boldsymbol{\Sigma}$ $\boldsymbol{V}^T$

如果只保留前两个方向，系数 a_i's 是什么？



$\tilde{\boldsymbol{X}}$ $\boldsymbol{U}$ $\boldsymbol{\Sigma}$ $\boldsymbol{V}^T$

线性降维PCA大纲

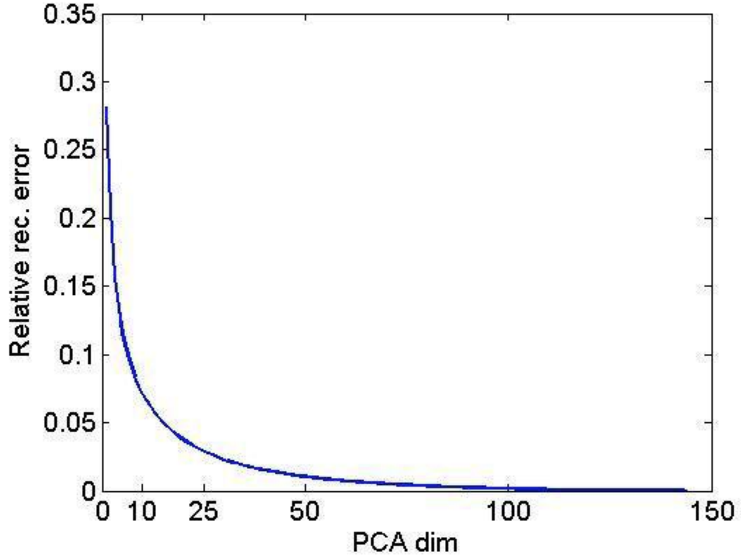
- 动机
- 视角 1： 最小化重构误差
- 视角 2： 最大化方差
- 视角 3: 奇异值分解（SVD）
- PCA的其他应用

图像压缩

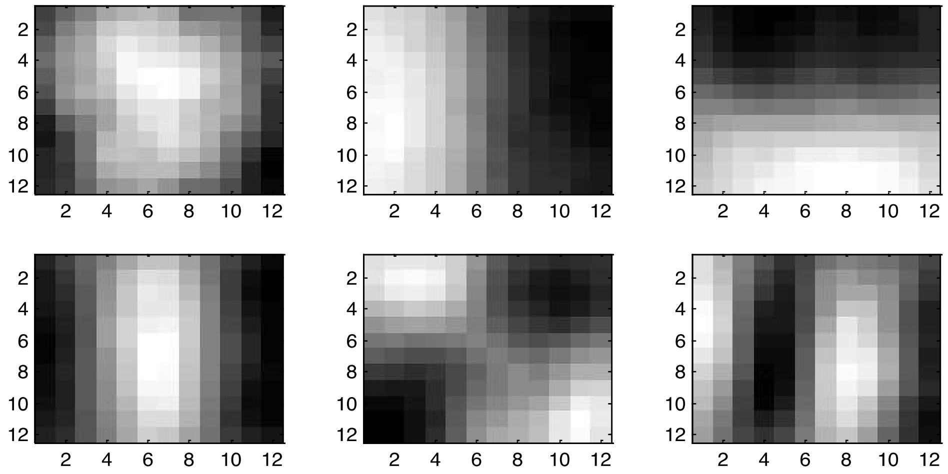
将一个 372×492 的图像分割成多个 12×12 的图像块

- 每个图像块被视为一个数据实例
- 在这些图像块上执行 PCA

重构误差 vs PCA 分量数



前 6 个 PCA 分量图示



使用前 60 个分量进行重构

使用前 16 个分量进行重构

图像去噪

噪声图像

去噪图像

由前 15 个主成分重构

课堂小结

- PCA的最小重建误差视角
- PCA的方差最大化视角
- PCA的矩阵分解（SVD）视角