華東師範大學
EAST CHINA NORMAL UNIVERSITY

数据科学与工程算法基础

Algorithm Foundations of Data Science and Engineering

第九章 SVD 与 PCA

课程提纲

Content

课程提纲

Content

矩阵对角化

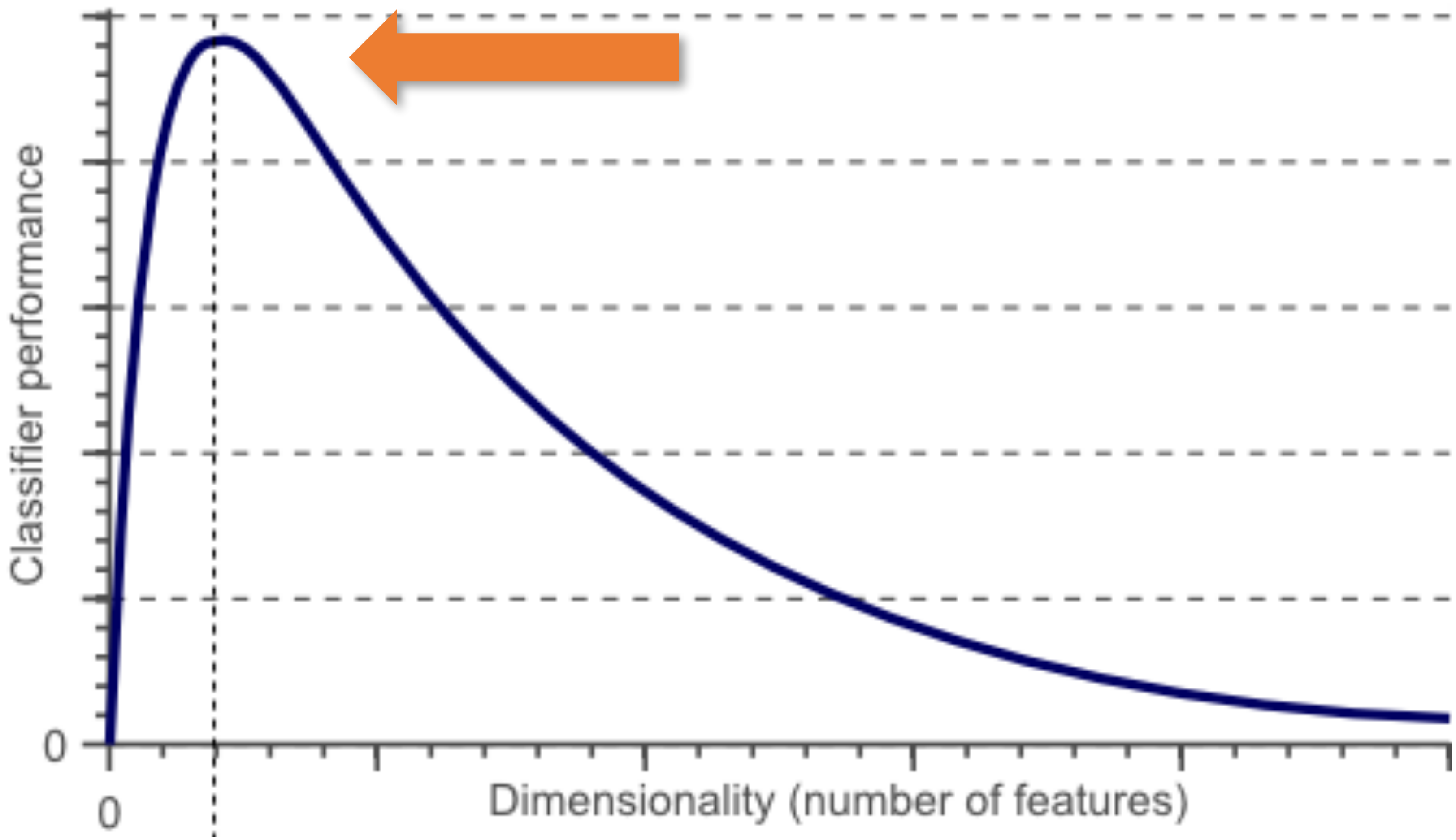
- 特征空间维度为n的矩阵$A \in \mathbb{R}^{n\times n}$可以正交对角化，即$A = P\Sigma P^\top$
  - P为正交矩阵
  - Σ为对角矩阵
- 很多应用中矩阵并非方阵
  - 二分图邻接矩阵
  - 特征矩阵
  - 图像
  - ……

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1j} & & a_{1N} \\ \cdots & & & & \cdots \\ a_{i1} & & a_{ij} & & a_{iN} \\ \cdots & & & & \cdots \\ a_{M1} & \cdots & a_{Mj} & & a_{MN} \end{pmatrix}$$

- 不是方阵可以对角化吗?

降维

- 随着维数的增加，分类器的性能会先增后降



图片压缩

- 无人机性能参数
  - 速度100–300 m/s
  - 视频帧率为25–30 帧/s
  - 帧与帧间有重叠
- 航拍图像相关性非常强
- 如何利用相关性降低图像存储与传输量
- 图像压缩旨在解决这一问题
  - 无损压缩
  - 有损压缩

课程提纲

Content

1 算法引入

2 SVD

3 PCA

相似矩阵

- 对于矩阵$A, B \in \mathbb{R}^{n\times n}$，如果存在一个可逆的$n \times n$矩阵 P，使得$A = PBP^{-1}$，则称它们为相似矩阵
- **定理：** 若两个$n \times n$矩阵相似，则它们有相同的特征多项式，而且相同特征值的重数也相同
  - 方阵A可对角化：相似于一个对角矩阵，即存在一个可逆矩阵P，使得$A = PDP^{-1}$，其中D为对角矩阵
  - 如果A是可对角化的，那么$A^k = PD^kP^{-1}$，$k > 0$
  - 矩阵相似是一种等价关系：自反性、对称性和传递性
- 如果$v_1, \ldots, v_n$是矩阵A的线性无关的特征向量，λ_i是v_i对应的特征值，那么$A = PDP^{-1}$，其中$P = [v_1 \ldots v_n]$且$D = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$

对角化条件

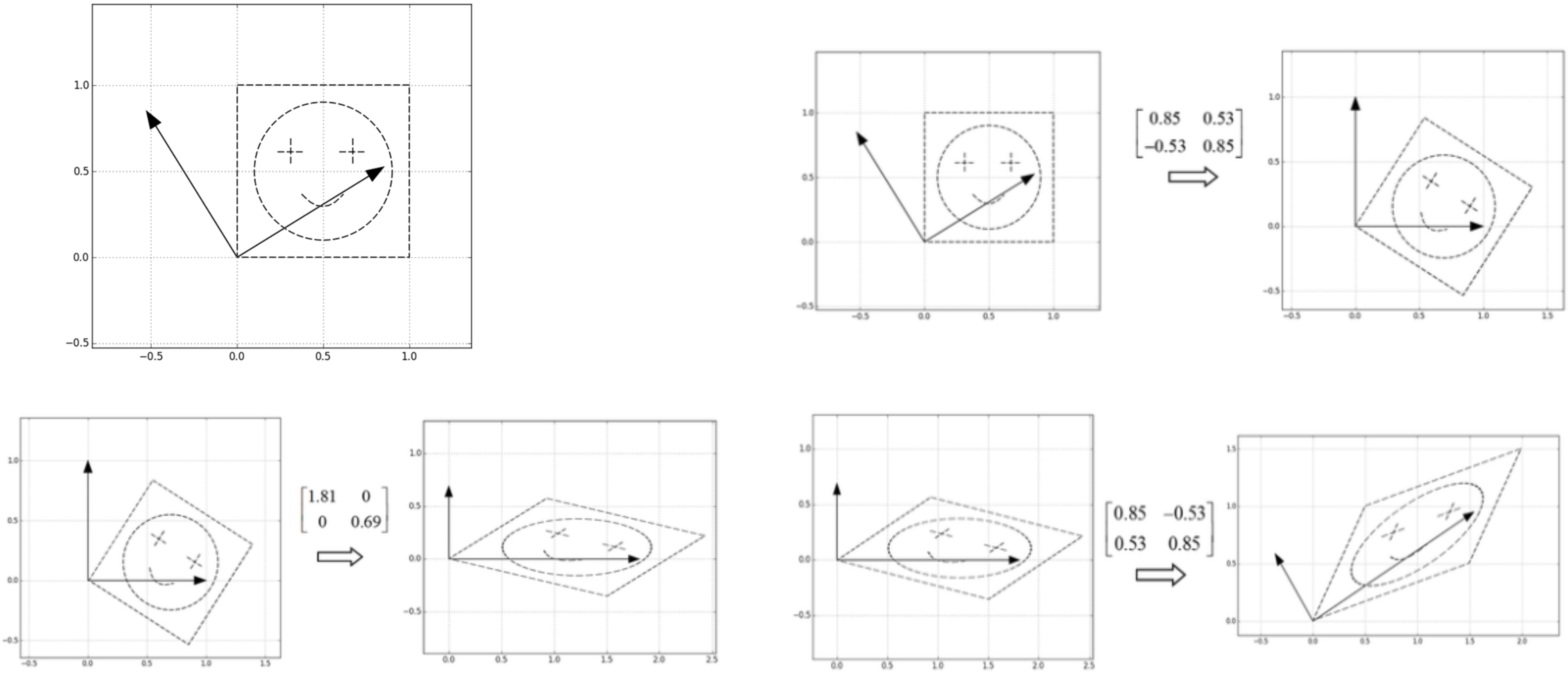
- 矩阵$A \in \mathbb{R}^{n\times n}$可对角化当且仅当$A$的每个特征值的几何重数都等于其代数重数
  - 矩阵$A \in \mathbb{R}^{n\times n}$有$n$个不重复的特征值时可对角化
  - 例如$A = \begin{pmatrix} 2 & 4 & 6 \\ 0 & 2 & 2 \\ 0 & 0 & 4 \end{pmatrix}$的特征值为$\lambda_1 = \lambda_2 = 2$，$\lambda_3 = 4$
    - $\lambda_1 = \lambda_2 = 2$对应的特征向量为$(1,0,0)^\top$，解$(A - 2I)v = 0$可得
    - 代数重数≠几何重数
    - 因此，矩阵A不可对角化

实对称矩阵的对角化

- 如果$P^{-1} = P^\top$，方阵P是正交矩阵
- 矩阵A可正交对角化，如果存在方阵P使得$A = PDP^\top$，其中D为对角矩阵
- 实对称矩阵的性质
  - A有n个特征值（包括重数）
  - 对于任一特征值，其对应的特征向量的几何重数等于代数重数
  - 如果A为实对称矩阵，那么A的任意两个特征向量都是正交的
  - 特征空间是相互正交的
  - A是可正交对角化的

对角化的几何解释

$$\begin{pmatrix}1.5 & 0.5\\ 0.5 & 1.0\end{pmatrix} = \begin{pmatrix}0.85 & -0.53\\ 0.53 & 0.85\end{pmatrix}\begin{pmatrix}1.81 & 0\\ 0 & 0.69\end{pmatrix}\begin{pmatrix}0.85 & -0.53\\ 0.53 & 0.85\end{pmatrix}^{T}$$

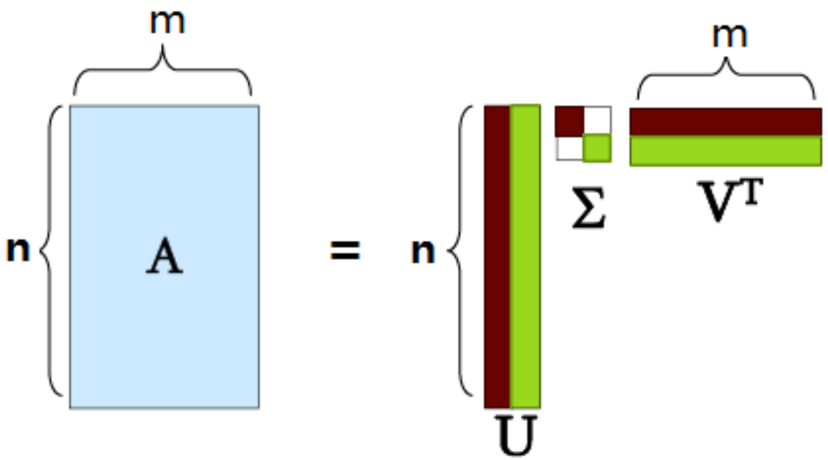


奇异值

- 对任意矩阵A，$AA^\top$和$A^\top A$都是对称半正定矩阵（课后尝试自己证明）
  - 所有特征值非负；可以正交对角化
- 对任意矩阵A，有$\mathrm{rank}(A)=\mathrm{rank}(AA^\top)=\mathrm{rank}(A^\top A)$
  - $AA^\top$和$A^\top A$都有$\mathrm{rank}(A)$个非零特征值
- 实对称矩阵$AA^\top$和$A^\top A$的特征值分别为λ_i 和 μ_i，我们有$\lambda_i=\mu_i>0,\ i=1,2,\ldots,r$ （课后尝试自己证明）
- 对$A\in\mathbb{R}^{m\times n}$，$AA^\top$和$A^\top A$的特征值分别为$\lambda_i$和$\mu_i$，我们称$\alpha_i=\sqrt{\lambda_i}=\sqrt{\mu_i},\ i=1,\ldots,r$为$A$的奇异值

奇异值分解

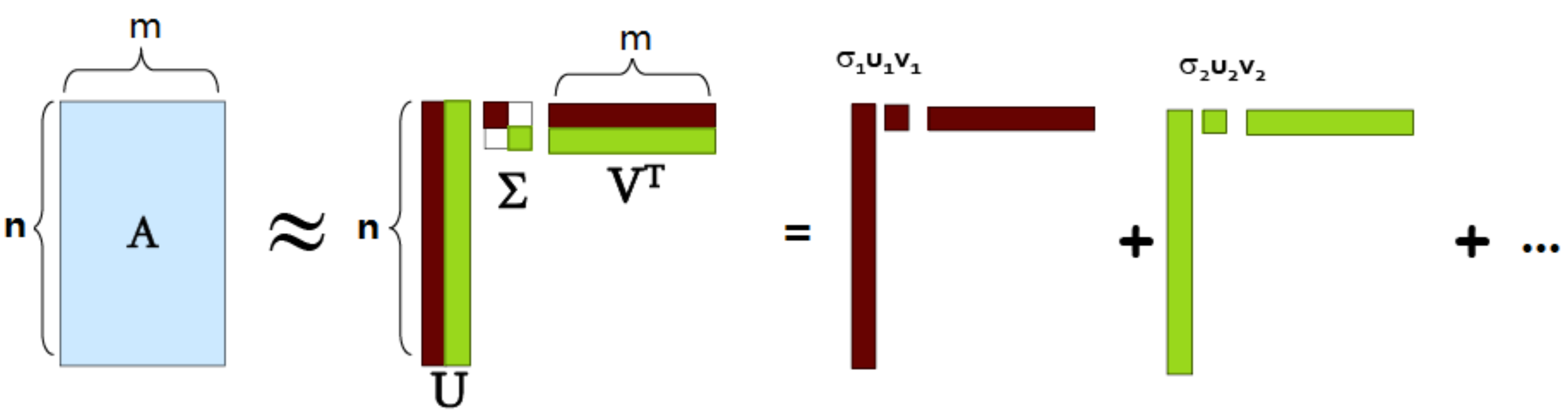
- 奇异值分解（Singular Value Decomposition，SVD）
  - 任意$m \times n$的矩阵A都可以被分解为$A_{[m\times n]} \sim U_{[m\times m]} D_{[m\times n]} V^{\top}_{[n\times n]}$



  - 矩阵A为$m \times n$的矩阵(不一定为方阵)
    - ✓ 例如m个用户，n个物品
    - ✓ U和V：左（m个用户和k种兴趣）和右奇异（k种兴趣和n个商品）矩阵
    - ✓ D：一个$m \times n$的对角矩阵仅有k个非零分量

SVD 分解方法

- 假如$A = UDV^\top$，如何确定矩阵U, D, V
  - 实对称矩阵$AA^\top$的正交对角化$AA^\top = UD^2U^\top$
  - 实对称矩阵$A^\top A$的正交对角化$A^\top A = VD^2V^\top$
  - 其中$D^2 = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_k^2)$，$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k > 0$ 称为矩阵 A 的奇异值



降维方法

$$A = \left[\begin{array}{ccc|ccc} u_1 & \cdots & u_k & u_{k+1} & \cdots & u_m \end{array}\right] \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_k & \\ \hline & 0 & & 0 \end{array}\right] \left[\begin{array}{c} v_1^T \\ \vdots \\ v_k^T \\ \hline v_{k+1}^T \\ \vdots \\ v_n^T \end{array}\right]$$

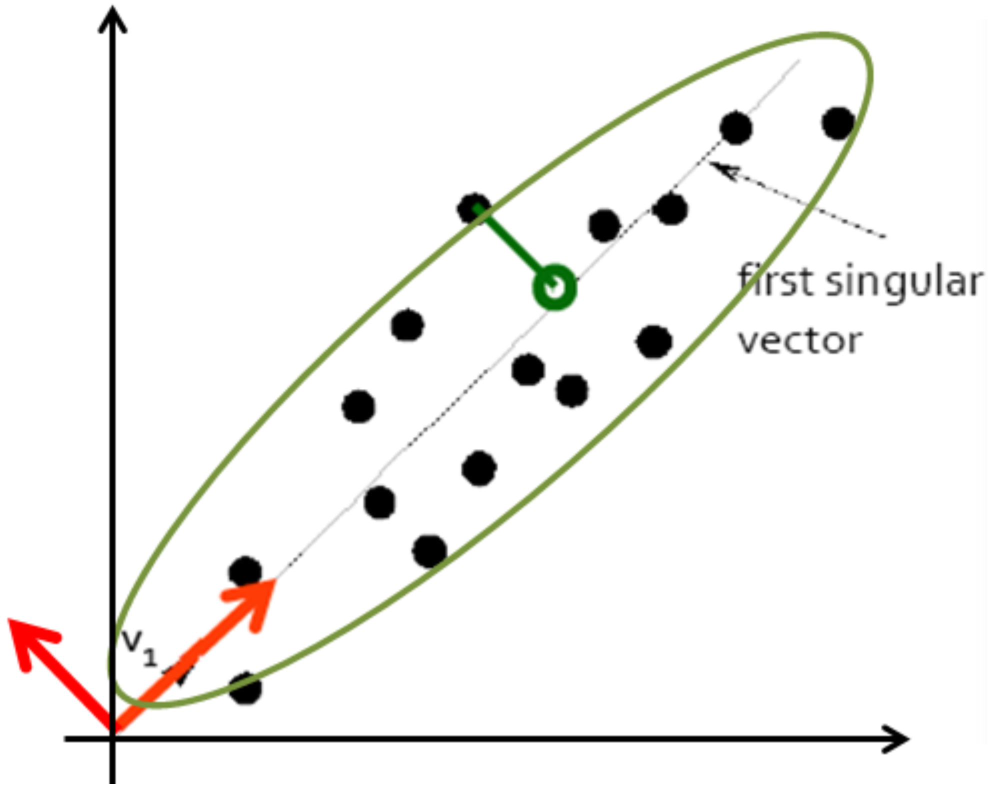
$$A = \left[\begin{array}{ccc} u_1 & \cdots & u_k \end{array}\right] \left[\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{array}\right] \left[\begin{array}{c} v_1^T \\ \vdots \\ v_k^T \end{array}\right]$$

一些小的奇异值可以忽略不计

$k = \arg\min_{l} \left\{ \frac{\sum_{i=1}^{l} \sigma_i}{\sum_{i=k}^{l} \sigma_i} > 90\,\% \right\}$个最大的奇异值重构的矩阵为$A^{(k)} \approx \sum_{i=1}^{k} \sigma_i u_i v_i^\top$

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \approx \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

SVD 的几何解释



- SVD给出来最佳投影，“最佳”是指最小化总的投影误差SVD给出最小重构误差的降维方法

重构误差

- 矩阵A的2范数和Frobenius范数定义如下:
  - $\|A\|_2 = \max_{x \neq 0} \dfrac{\|Ax\|_2}{\|x\|_2} = \sigma_{\max}(A)$
  - $\|A\|_F = \sqrt{\sum_{i,j} a_{i,j}^2} = \sqrt{\mathrm{Trace}(A^\top A)}$
- 如果$r < k$，$A^{(r)} \approx \sum_{i=1}^{r} \sigma_i u_i v_i^\top$
- 重构误差将为
  - $\|A^{(r)} - A\|_2 = \sigma_{r+1}$
  - $\|A^{(r)} - A\|_F = \sqrt{\sum_{r+1}^{k} \sigma_i^2}$

SVD 的应用

- 方阵A是非奇异的，则对$\forall i$都有$\sigma_i > 0$
  - 如果A是非奇异矩阵，它的逆矩阵为$A^{-1} = V^\top D^{-1} U$
  - 如果A是奇异矩阵，可以用SVD来近似它的逆矩阵
    - $A^{-1} = (UDV^\top)^{-1} \approx V^\top D_0^{-1} U$，其中$D_0^{-1} = \begin{cases} \frac{1}{\delta_i} & \text{if } \sigma_i > t \\ 0 & \text{otherwise} \end{cases}$，其中$t$是一个阈值
- 其他应用
  - 数据压缩
  - 主成分分析
  - 个性化推荐（协同过滤）

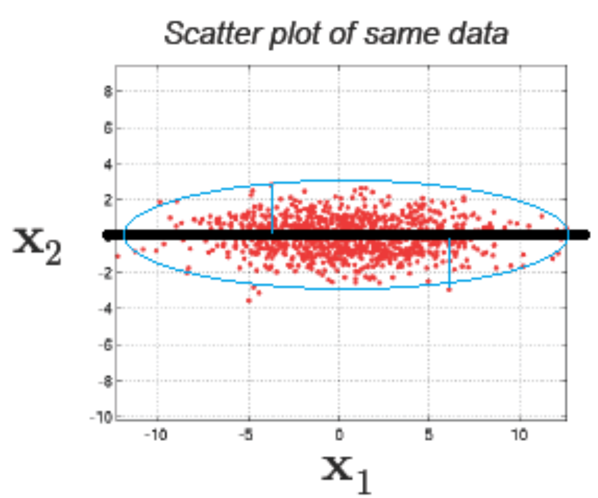
课程提纲

Content

降维

$$x = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_N \end{bmatrix} \text{-}\text{-}> \quad \text{降维} \quad \text{-}\text{-}> y = \begin{bmatrix} b_1 \\ b_2 \\ \cdots \\ b_K \end{bmatrix} \quad (K << N)$$

- 高维数据可能会出现问题，如维度灾难、过拟合
  - 将数据映射到低维空间
  - 但保留原数据中尽可能多的信息
- 降维的目标
  - 找到好的特征表示
  - 减少数据冗余



主成分分析

- 主成分分析（PCA）
  - 删除数据间的冗余关系
  - 保留尽可能的信息
- PCA：SVD 方法的一个应用
  - 将数据投影到最佳低维空间，其中“最佳”意味着投影误差最小
  - 即最小化重构误差
  - 根据高维数据的协方差矩阵确定最佳的投影方向
    - 协方差矩阵“最大”特征值对应的特征向量方向
    - 也称为“主成分”

协方差

- 样本均值$\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$
- 样本方差$S^2 = \dfrac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$
- 对于两个维度X和Y，协方差定义为

$$\mathrm{Cov}(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

  - X_i和Y_i偏离$\overline{X}$和$\overline{Y}$同号：正相关
  - X_i和Y_i偏离$\overline{X}$和$\overline{Y}$异号：负相关
  - X_i和Y_i偏离$\overline{X}$和$\overline{Y}$正负没有明显特征：不相关

协方差矩阵

- 对于n维数据集，我们可以计算每个维度之间的协方差，共$n \times n$个协方差
- 协方差矩阵Σ：实对称矩阵

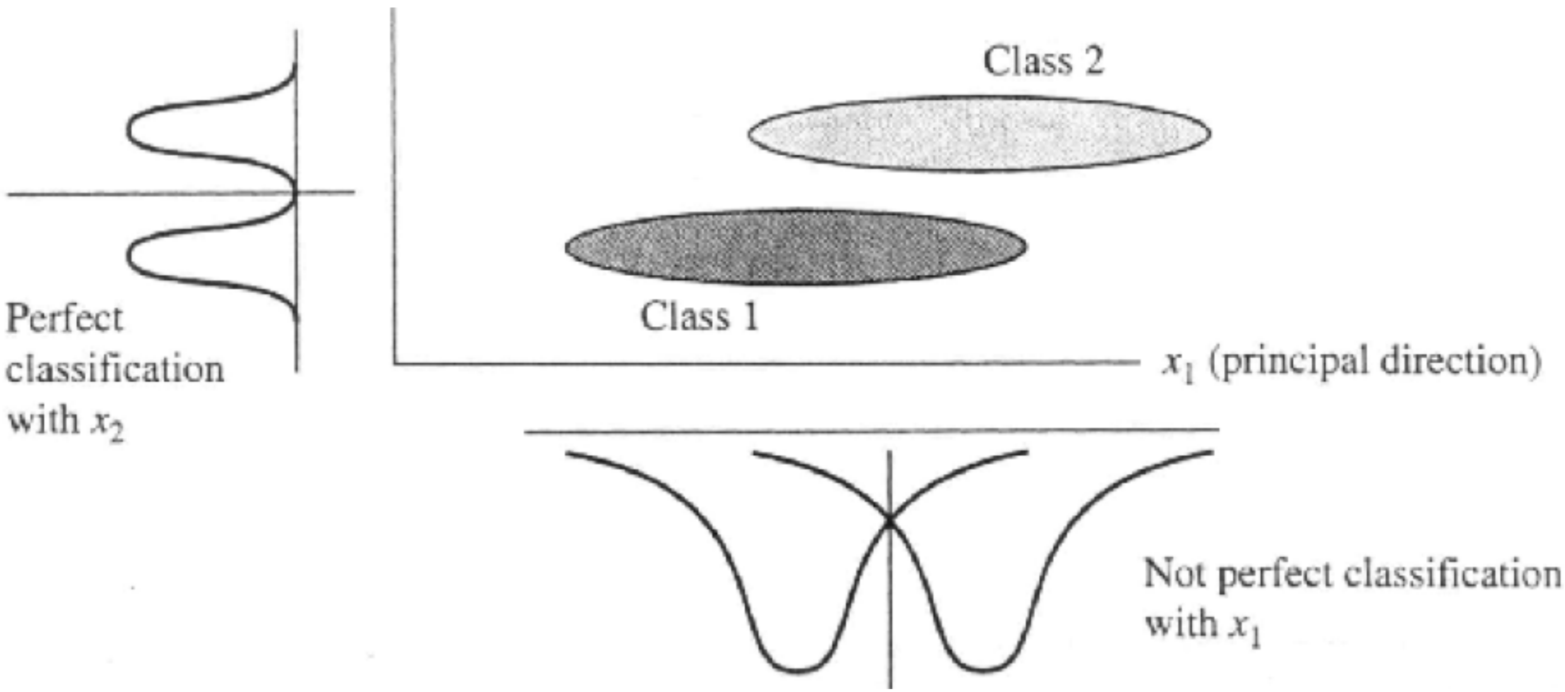
$$\Sigma = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \dots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \dots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \dots & \mathrm{Cov}(X_n, X_n) \end{bmatrix}$$

PCA 方法

- 假设$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$是$d \times 1$维向量
  - 计算中心化向量$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$并去中心化$\mathbf{y}_i = \mathbf{x}_i - \bar{\mathbf{x}}$
  - 构造$d \times n$维矩阵$Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]$，计算$d \times d$维协方差矩阵$C = \dfrac{YY^\top}{n-1}$
  - 对C矩阵进行正交对角化（对Y矩阵进行 SVD 分解），假定λ_i对应的特征向量为$\mathbf{u}_i$
  - 保留前k个大于α的特征值对应的特征向量
  - 降维重建向量$\widehat{\mathbf{x}_i} = [\mathbf{u}_1, \ldots, \mathbf{u}_k]^\top \mathbf{x}_i$

PCA 分析

- PCA 是 SVD 方法的一个应用，因此 PCA 方法最小化了重构误差
- PCA 并不总是最优的降维方法，如分类问题



本章小结

- SVD方法实现非方阵的对角化分解
  - 重构误差最小化
  - PCA是SVD方法的一个应用
- 高维数据处理
  - 高维数据在应用中随处可见
    - 文本，图像，图，……
  - 降维是高维数据的常用处理方法
    - 线性降维方法
    - 非线性降维方法，如t–SNE