

第16章 主成分分析

VCG

2025-11-07

特征维度灾难

➤ 高维数据的挑战

- 高维空间中的数据点变得极其稀疏，距离度量失去意义
- 算法的计算和存储开销随维度指数级增长
- 人类无法直观理解超过三维的空间

➤ 现实世界的高维数据往往冗余的

- 许多特征高度相关，它们共同描述着某种更本质、更低维的潜在结构
- 数据冗余。如一张1000x1000像素的图片（100万维），其内容可能只是一个简单的物体

➤ 如何在信息损失最小的前提下，对数据进行降维？

核心思想：寻找一个新的、更具信息量的坐标系

➤ PCA的哲学

- 信号蕴含于变化之中，方差越大的（线性意义）方向，包含的信息越丰富

➤ 目标

- 希望找到一组全新的正交坐标轴（主成分，主方向）来重新描述数据
- 数据在主方向上的投影方差最大化

➤ 对数据进行正交变换，数据由线性相关基表示，变成正交基表示，且

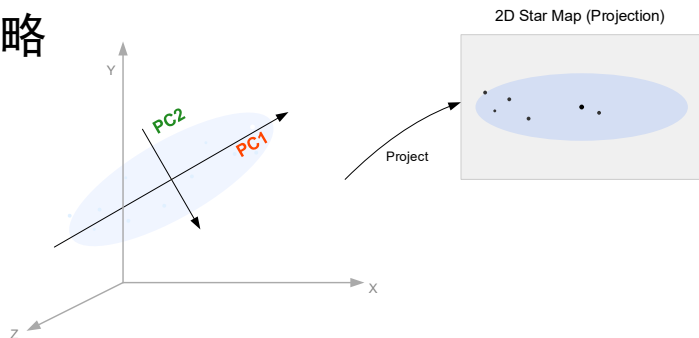
- 第一主成分 (PC1)，数据在其上投影的方差最大
- 第二主成分 (PC2)，与PC1正交，且数据在其上投影的方差次大
- 以此类推...

➤ 通过保留方差最大的前 k 个主成分

- 可以用一个 k 维向量（远小于原始维度）来近似表示一个数据点

为星云绘制最精确的二维星图

- 观测三维上的扁平状星云。如何在二维星图上展示它，并最大程度保留其结构？
 - 寻找主方向 (PC1)
 - 旋转视角，直到找到一个方向，使得该方向看过去，星云的“影子”最长、最舒展：该方向捕捉了星云最大的变异
 - 寻找次要方向 (PC2)
 - 固定主方向，围绕它旋转，寻找一个与之垂直的方向，使得星云的“影子”次长
 - 绘制星图
 - 将所有星星投影到由这两个方向构成的平面上。这样，虽然丢失了厚度信息，但最大程度保留了星云的整体形状
- “寻找方差最大的投影方向”为PCA的核心策略



两种等价的数学视角

➤ PCA的优化目标的等价方式

➤ 最大投影方差 (Maximal Variance)

- 寻找一个投影方向，使得数据点在该方向上的投影尽可能分散
- 保留了数据中最主要的“变化”或“信号”

➤ 最小重建误差 (Minimal Reconstruction Error)

- 寻找一个低维超平面（由主成分定义），使得原始数据点到它们在该平面上投影点的距离之和最小
- 保证了降维后的数据能最忠实地“重建”出原始数据

➤ 这两个看似不同的目标，会导出完全相同的解

红酒列表的PCA特性

➤ 寻找在所有红酒中很不相同的属性

➤ 尽可能体现红酒差异的属性

➤ 属性差异=>独立=>正交

➤ 显著：投影分散

➤ 方差何时最大？

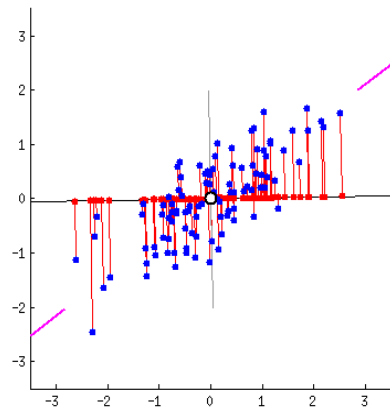
➤ 属性容易重建原本的红酒特性

➤ 重建原本特性的属性

➤ 基

➤ 寻找能够尽可能好地重建

➤ 红线的长度和为重建误差，何时最小化？



数学工具之协方差矩阵

➤ 如何衡量数据在不同方向上的“伸展”程度？

➤ 数据的协方差矩阵 (Covariance Matrix) S

➤ 协方差矩阵 S 告诉我们什么？

➤ 对称，可对角化

➤ 对角线元素 S_{ii} ：第 i 个特征自身的方差

➤ 非对角线元素 S_{ij} ：第 i 和第 j 个特征的协方差，衡量它们的线性相关性

➤ 特征之间线性相关：某个特征的值一定程度上可由其它特征的线性组合来预测

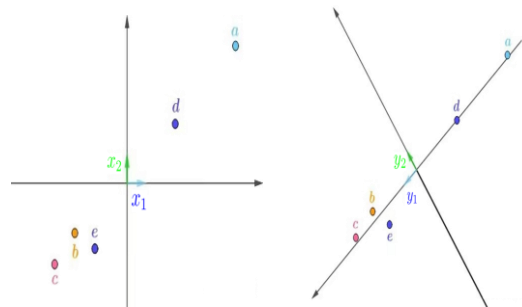
➤ 协方差矩阵完全描述了数据分布的（二次）几何形状

➤ 如果数据点形成一个“超椭球体”

➤ 协方差矩阵的特征向量指向这个椭球体的主轴方向

➤ 特征值代表了这些主轴方向上的方差大小（轴的长度平方）

➤ PCA实际上就是对数据的协方差矩阵进行特征值分解



直观解释 – 方差和最大

➤ PCA旨在选取正交变换中方差最大的变量作为第一主成分

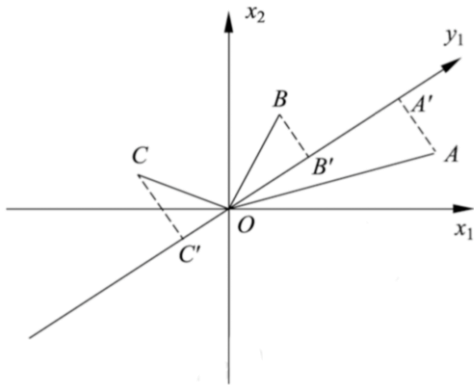
➤ 方差

➤ 坐标值的平方和: $OA'^2 + OB'^2 + OC'^2$, 最大化

➤ 误差和

➤ 样本点到 y_1 轴的距离的平方和: $AA'^2 + BB'^2 + CC'^2$, 最小化

➤ 方差和最大等价于投影误差和最小



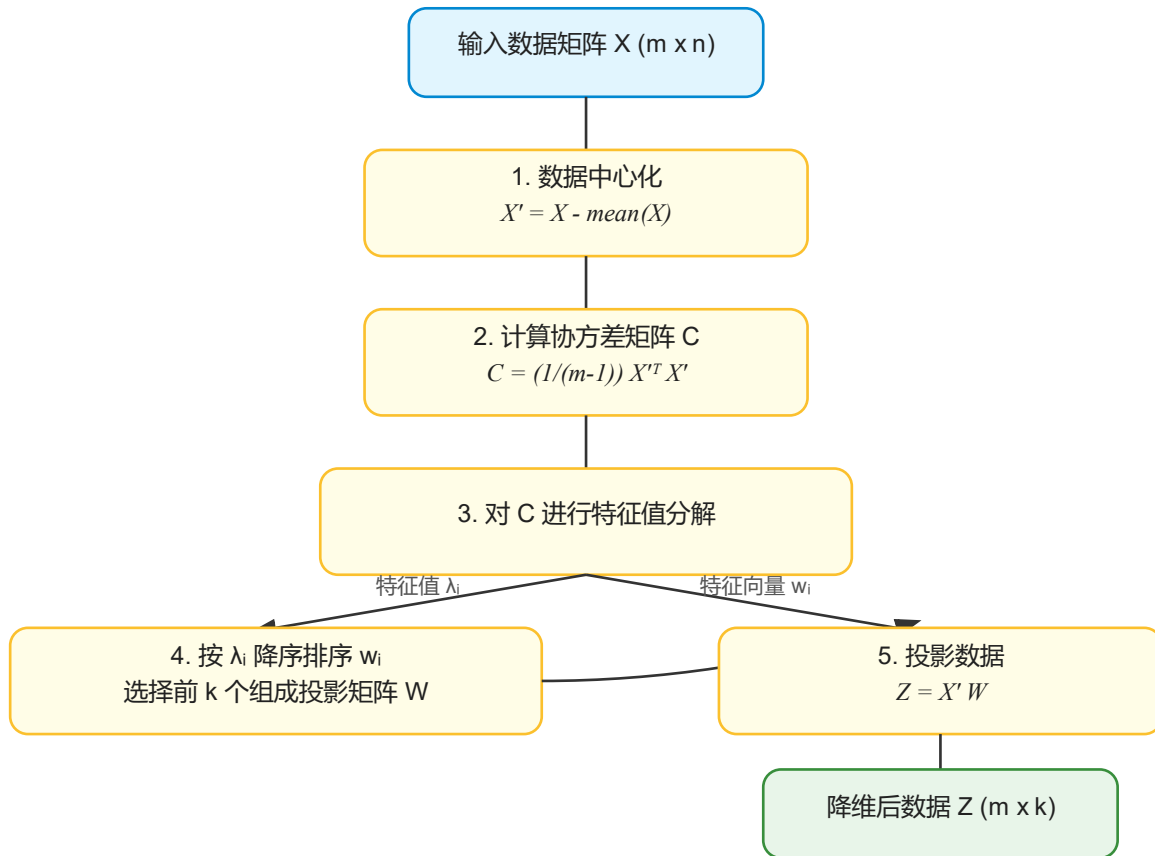
PCA推导：从投影方差最大化到特征值分解

- 优化目标：找到一个单位方向向量 $\mathbf{u}$ ($\|\mathbf{u}\| = 1$)，使得数据 $\mathbf{X}$ 在其上的投影方差最大
 - 数据中心化，使得均值为0
 - 投影方差 $Var(\mathbf{Xu}) = \frac{1}{N}(\mathbf{Xu})^T(\mathbf{Xu}) = \mathbf{u}^T(\frac{1}{N}\mathbf{X}^T\mathbf{X})\mathbf{u} = \mathbf{u}^T\mathbf{Su}$ ，其中， $\mathbf{S}$ 为数据协方差矩阵
- 于是，得优化问题

$$\max_{\mathbf{u}} \quad \mathbf{u}^T\mathbf{Su} \quad \text{subject to} \quad \mathbf{u}^T\mathbf{u} = 1$$

- 求解 (拉格朗日乘子法)
 - 拉格朗日函数： $L(\mathbf{u}, \lambda) = \mathbf{u}^T\mathbf{Su} - \lambda(\mathbf{u}^T\mathbf{u} - 1)$
 - 对 $\mathbf{u}$ 求导，并令其为零： $\frac{\partial L}{\partial \mathbf{u}} = 2\mathbf{Su} - 2\lambda\mathbf{u} = 0$
 - 得： $\mathbf{Su} = \lambda\mathbf{u}$
- 投影向量 $\mathbf{u}$ ，为数据协方差矩阵 $\mathbf{S}$ 的特征向量
 - 最大的投影方差方向 $\mathbf{u}$ ，该方向上的投影方差 $\mathbf{u}^T\mathbf{Su} = \mathbf{u}^T\lambda\mathbf{u} = \lambda$ ，为最大特征值
 - $\mathbf{u}$ 为对应的特征向量

PCA流程



PCA算法

➤ 算法目标：给定数据矩阵 $\mathbf{X}$ (N 个样本, m 个维度), 降维到 k 维

➤ 数据中心化（均值为0），标准化（方差为1）

➤ 计算协方差矩阵

$$\text{➤ } \mathbf{S} = \frac{1}{N-1} \mathbf{X}^T \mathbf{X}$$

➤ 特征值分解

➤ 对协方差矩阵 $\mathbf{S}$ 进行特征值分解，得到特征值 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$ 和对应的特征向量

$$\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m$$

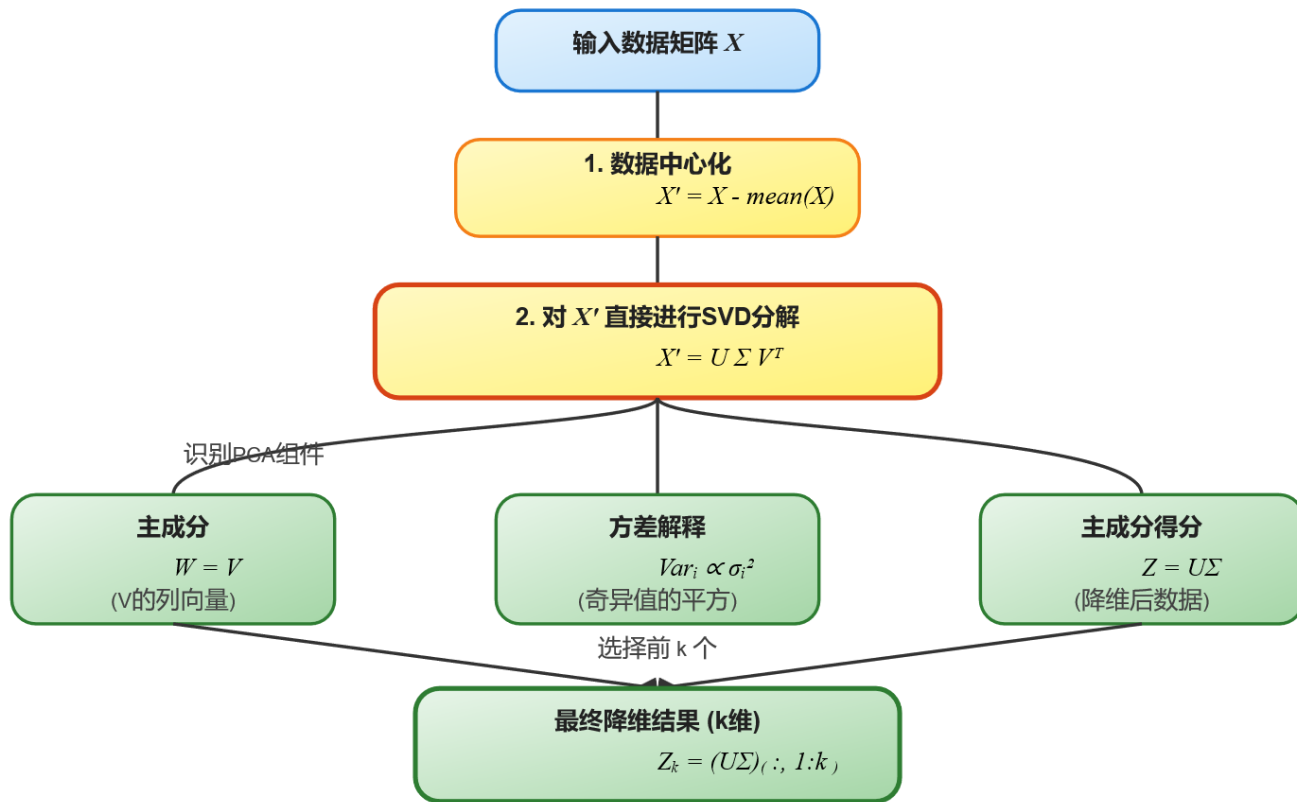
➤ 选择主成分

➤ 选择前 k 个最大的特征值对应的特征向量，构成投影矩阵 $\mathbf{W} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_k]$

➤ 数据投影 (降维)

➤ 新的低维数据 $\mathbf{Z} = \mathbf{XW}$

SVD路线



另一种实现路径：奇异值分解 (SVD)

➤ 什么是SVD?

➤ 任何矩阵 $\mathbf{X}$ 都可以分解为： $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$

➤ $\mathbf{U}, \mathbf{V}$ 是正交矩阵， $\mathbf{\Sigma}$ 是对角矩阵（对角线元素为奇异值）

➤ 通过SVD进行PCA分解

➤ 对中心化数据 $\mathbf{X}$ 进行SVD分解， $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$

➤ 协方差矩阵 $\mathbf{S} = \frac{1}{N-1} \mathbf{X}^T \mathbf{X} = \frac{1}{N-1} (\mathbf{V}\mathbf{\Sigma}^T \mathbf{U}^T)(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T) = \mathbf{V}(\frac{\mathbf{\Sigma}^2}{N-1})\mathbf{V}^T$

➤ 为 $\mathbf{S}$ 的特征值分解形式

➤ 主成分 (特征向量) 就是矩阵 $\mathbf{V}$ 的列向量

➤ 特征值 $\lambda_i = \frac{\sigma_i^2}{N-1}$ ，其中 σ_i 是奇异值

➤ 为什么有时更倾向于用SVD?

➤ 避免了计算 $\mathbf{X}^T \mathbf{X}$ 这一步，可以减少计算过程中的精度损失

➤ SVD存在优化算法，对于某些数据，直接对 $\mathbf{X}$ 做SVD可能比计算并分解 $\mathbf{X}^T \mathbf{X}$ 更高效

总结与局限

➤PCA总结

- 一种强大的无监督线性降维技术
- 通过寻找最大化数据方差的方向（主成分）来工作
- 数学上等价于对数据协方差矩阵进行特征值分解
- 在数据可视化、去噪、特征提取等领域应用广泛

➤PCA的局限性

- PCA只能发现数据中的线性结构。对于非线性结构（如瑞士卷），它会失效
- 假设方差越大的方向越重要。在某些场景下（如某些分类问题），方差小的方向可能包含关键的判别信息
- 主成分是原始特征的线性组合，其物理意义可能难以解释
- 异常值会极大地影响协方差矩阵的计算，从而影响主成分的方向

Thanks