PCA+SVD

2025/8/16

1 WHAT IS PCA AND WHY

降维可以提取数据内部的本质结构，减少冗余信息和噪声信息造成误差。维度降低便于计算和可视化，有效信息的提取综合以及无用信息摒弃。

主成分分析 (Pincipal Component Analysis)，将众多具有一定相关性的指标重新合成一组少量互相无关的综合指标。

降维后样本的方差尽可能大，数据的均方误差尽可能小。

2 数学推导

设样本矩阵 $X \in R^{n\times p}$，每行是一个样本。首先进行去中心化处理:

$$\tilde{X} = X - \mathbf{1}\mu^\top, \quad \mu = \frac{1}{n}\sum_{i=1}^{n} X_{i\cdot}.$$

样本协方差矩阵为:

$$C = \frac{1}{n-1}\tilde{X}^\top\tilde{X} \in R^{p\times p}.$$

一主成分 希望找到一个单位向量 w，使得投影 $\tilde{X}w$ 的方差最大:

$$\max_{\|w\|=1} \operatorname{Var}(\tilde{X}w) = \frac{1}{n-1}\|\tilde{X}w\|_2^2 = w^\top Cw.$$

引入拉格朗日乘子:

$$\mathcal{L}(w,\lambda) = w^\top Cw - \lambda(w^\top w - 1).$$

令梯度为零:

$$2Cw - 2\lambda w = 0 \quad \Rightarrow \quad Cw = \lambda w.$$

因此 w 是协方差矩阵 C 的特征向量，λ 为对应特征值。取最大特征值对应的特征向量为第一主成分 w_1。

多主成分　继续在与已选主成分 $w_1, \ldots, w_{k-1}$ 正交的约束下，最大化投影方差，可得：

$$CW = W\Lambda, \quad W^\top W = I_k,$$

其中 $W = [w_1, \ldots, w_k]$ 按特征值从大到小排序，$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$。低维表示为：

$$Z = \tilde{X}W.$$

解释方差比　第 j 个主成分的方差为 λ_j，解释方差比为：

$$\mathrm{EVR}_j = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}.$$

3　代码

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.datasets import load_iris

data = load_iris()
X = data.data
y = data.target

pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

print("Explained variance ratio for each principal component:", pca.explained_variance_ratio_)
print("Cumulative explained variance ratio:", np.sum(pca.explained_variance_ratio_))

```