Chapter 4: Dimension Reduction

2018.12.17 Monday By KuKuXia@github.com

ML/DL/RL深度交流讨论群: 622545311，加群时请备注: 姓名-方向-公司or高校

主成分分析 PCA： Principal Components Analysis

作用

降低高维特征到低维，从而提升特征表达能力，降低训练复杂度

最大方差理论

信号具有较大方差、噪声具有较小方差，信号与噪声之比为信噪比，信噪比越大意味着数据的质量越好，反之，则质量越差

目标：最大化投影方差，也就是让数据在主轴上投影的方差最大

- 数据X投影之后的方差就是协方差矩阵的特征值
- 最大方差就是协方差矩阵的最大特征值，最佳投影方向就是最大特征值对应的特征向量
- 次佳投影方向位于最佳投影方向的正交空间中，是第二大特征值对应的特征向量

线性、非监督、全局的降维算法

- 有一定局限性，可以通过核映射对PCA进行扩展得到核主成分分析KPCA

旨在找到数据中的主成分，并利用这些主成分表征原始数据，从而达到降维的目的

对原始数据中心化的目的是使得投影之后的数据均值为0

求解方法

- 对样本数据进行中心化处理
- 求样本协方差矩阵
- 对协方差矩阵进行特征值分解，将特征值从大到小排列
- 取特征值前d大的对应的特征向量w_1,w_2,...,w_d，通过以下映射将n维样本映射到d维

$$x_i' = \left[\omega_1^T x_i \quad \omega_2^T x_i \quad \dots \quad \omega_2^T x_i\right]^T$$

新的(x_i)'的第d维就是x_i在第d个主成分w_d方向上的投影，通过选取最大的d个特征值对应的特征向量，我们将方差较小的特征(噪声)抛弃，使得每个n维列向量x_i被映射为d维列向量(x_i)'

$$\eta = \sqrt{\frac{\sum_{i=1}^{d} \lambda_i^2}{\sum_{i=1}^{n} \lambda_i^2}}$$

定义降维后的信息占比为

最小平方误差理论

算法没有考虑数据的标签(类别)，只是把原数据映射到一些方差比较大的方向上而已

线性判别分析 LDA： Linear Discriminant Analysis

由Ronald Fisher在1936年发明的，也称Fisher LDA

监督学习方法，降维过程中不会损失数据的类别信息

目标：最大化类间距离和最小化类内距离

适用于对有类别信息的数据进行降维处理

对数据分布有很强假设

- 每个类都是高斯分布
- 各个类的协方差相等

线性模型对于噪声的鲁棒性比较好，但由于模型简单，表达能力有一定局限性，因此可以引入核函数扩展LDA方法以处理分布较为复杂的数据

求解方法

- 计算数据集中每个类别样本的均值向量μ_j，及总体均值向量μ
- 计算类内散度矩阵S_w，全局散度矩阵S_t，并得到类间散度矩阵S_b = S_t - S_w
- 对矩阵(S_w).inv S_b进行特征值分解，将特征值从小到大排列
- 去特征值前d大的对应的特征向量W_1,W_2,..., W_d，通过下列映射将n维样本映射到d维：

$$x_i' = \left[\omega_1^T x_i \quad \omega_2^T x_i \quad \dots \quad \omega_2^T x_i\right]^T$$

方法

- 主成分分析
- 线性判别分析

都是线性降维，无监督用PCA，有监督用LDA

- 等距映射
- 局部线性嵌入
- 拉普拉斯特征映射

非线性降维操作

- 局部保留投影等