主成分分析

cloud

2017.1.11

1 概述

实际问题中往往需要研究多个特征，而这些特征之间存在一定的相关性。过多的特征增加了问题的复杂性，所以需要将特征组合为较少的代表性特征，这些组合的特征能代表原始特征的绝大部分信息，且组合后的特征之间互不相关，而这一种方法就是主成分分析 (PCA)。PCA 是最常用的一种降维方法，首先直接给出 PCA 的基本流程：

输入：样本集 $D = (x^{(1)}, x^{(2)}, ..., x^{(N)})$，低维空间维数 d'

输出：投影矩阵 $W = (\omega_1, \omega_2, ..., \omega_{d'})$

（1）对所有样本进行中心化 $x^{(i)} \leftarrow x^{(i)} - \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$

（2）计算样本的协方差矩阵 XX^T

（3）对协方差矩阵 XX^T 做特征值分解

（4）取最大的 d' 个特征值所对应的特征向量 $\omega_1, \omega_2, ..., \omega_{d'}$

在上面的算法中，有 N 个样本，每个样本有 d 个特征，经过 PCA 之后特征维数降为 d'。那么为什么通过 PCA 可以得到较优的 d' 维特征呢，下面给出两种理解。

1.1 方差角度理解

对于有 d 个特征的 N 个样本，假定已经中心化了，将每个样本写成行向量得到矩阵 X^T 如下：

$$X^T = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \cdots & x_d^{(N)} \end{pmatrix} = \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(N)} \end{pmatrix}$$

将 N 个样本投影到某条直线 L 上，计算 N 个投影点的方差，我们认为方差最大的直线方向是样本的主方向 u。

（1）首先计算 $X^T \cdot u$ 的值为：

$$X^T \cdot u = \begin{pmatrix} x^{(1)} \cdot u \\ x^{(2)} \cdot u \\ \vdots \\ x^{(N)} \cdot u \end{pmatrix}$$

（2）因为中心化之后样本均值为 0，所以求 $X^T \cdot u$ 的方差为：

$$Var(X^T \cdot u) = (X^T u - E)^T (X^T u - E) = (X^T u)^T (X^T u) = u^T X X^T u$$

（3）那么目标函数就是要极大化 $J(u)$：

$$J(u) = u^T X X^T u$$

由于 u 代表的是主方向与模无关，所以令 $||u||_2 = 1$，从而 $u^T u = 1$。增加约束之后的目标函数为：

$$\begin{aligned} J(u) &= u^T X X^T u \\ s.t. \quad & u^T u = 1 \end{aligned}$$

建立拉格朗日方程 $L(u) = u^T X X^T u - \lambda(u^T u - 1)$ 对 u 求导取 0 可得 $(XX^T)u = \lambda u$，所以知道 λ 是 XX^T 的特征值，u 是对应的特征向量。将这个式子代入上式可得：

$$\begin{aligned} J(u) &= u^T X X^T u \\ &= u^T \lambda u \\ &= \lambda \end{aligned}$$

由上式可知，要想使 $J(u)$ 极大，那么 λ 也要极大，所以最佳的投影直线是特征值 λ 最大时对应的特征向量，其次是 λ 第二大对应的特征向量。依次类推，我们只需要对协方差矩阵进行特征值分解，得到的前 d' 大特征值对应的特征向量就是最佳的 d' 维新特征，而且这 d' 维新特征是正交的。

1.2 最近重构性/最大可分性理解

在说明 PCA 之前，先考虑这样一个问题。对于正交属性空间中的样本点，如何用一个超平面来合理表示样本的分布？如果存在这样的超平面，应该具备如下的性质：

（1）最近重构性：样本点到这个超平面的距离都足够近

（2）最大可分性：样本点在这个超平面的投影尽可能分开

基于这两种性质，能够得到 PCA 的两种等价推导。

首先对样本进行中心化 $x^{(i)} \leftarrow x^{(i)} - \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$，就有 $\sum_{i=1}^{N} x^{(i)} = 0$。假定投影变换后得到的新坐标系为 $\{\omega_1, \omega_2, ..., \omega_d\}$，其中 ω_i 是标准的正交基向量有 $||\omega_i||_2 = 1, \omega_i^T\omega_j = 0(i \neq j)$。若丢弃新坐标系中的部分坐标，将维度降为 d'，则某个样本点 $x^{(i)}$ 在低维坐标系中的投影是 $z^{(i)} = (z_1^{(i)}, z_2^{(i)}, ..., z_{d'}^{(i)})$，其中 $z_j^{(i)} = \omega_j^T x^{(i)}$ 是 $x^{(i)}$ 在低维坐标系下第 j 维的坐标。若基于 $z^{(i)}$ 来重构 $x^{(i)}$，则会得到 $\hat{x}^{(i)} = \sum_{j=1}^{d'} z_j^{(i)}\omega_j$。

考虑整个训练集，原样本点 $x^{(i)}$ 和基于投影重构的样本点 $\hat{x}^{(i)}$ 之间的距离为：

$$
\begin{aligned}
\sum_{i=1}^{N} ||\hat{x}^{(i)} - x^{(i)}||_2^2 &= \sum_{i=1}^{N} || \sum_{j=1}^{d'} z_j^{(i)}\omega_j - x^{(i)}||_2^2 \\
&= \sum_{i=1}^{N} (z^{(i)})^T z^{(i)} - 2\sum_{i=1}^{N} (z^{(i)})^T W^T x^{(i)} + constant \\
&\propto -tr(W^T(\sum_{i=1}^{N} x^{(i)}(x^{(i)})^T)W)
\end{aligned}
$$

（1）根据最近重构性上式应被最小化，又因为 ω_j 是标准正交基，$\sum_{i=1}^{m} x^{(i)}(x^{(i)})^T$ 是协方差矩阵有：

$$
\begin{aligned}
&\min_W -tr(W^T X X^T W) \\
&s.t. \quad W^T W = I
\end{aligned} \tag{1}
$$

这就是主成分分析的优化目标，其中 tr 表示矩阵的迹，迹是所有对角元的和且迹是所有特征值的和，迹正比于数据在各个坐标轴方向上的方差的和。

（2）从最大可分性出发，样本点 $x^{(i)}$ 在新空间中超平面的投影是 $W^T x^{(i)}$。若所有样本点的投影尽可能分开，则应该使投影后样本点方差最大化，那么优化目标可以写成：

$$
\begin{aligned}
&\max_W tr(W^T X X^T W) \\
&s.t. \quad W^T W = I
\end{aligned} \tag{2}
$$

式 (5)(6) 是等价的，同样这个式子也等价于上一节的目标函数。通过拉格朗日乘子法求解可得：

$$X X^T W = \lambda W$$

于是问题最终转化为求 $tr(XX^T)$ 的极大，也就是 XX^T 特征值之和的极大，那么只需要对协方差矩阵 XX^T 进行特征值分解，对求得的特征值按照从大到小排序，取前 d' 大的特征值对应的特征向量构成 $W = (\omega_1, \omega_2, ..., \omega_{d'})$，这就是 PCA 的解。

2 参考文献

1. 小象学院邹博的机器学习课件
2. 周志华的机器学习