机器学习常见知识汇总之降维算法

田晓彬
xiaobin9652@163.com

主成份分析（PCA）

在机器学习领域，对原始数据进行特征提取，可能会得到比较高维度的特征向量。在这些向量所在的高维空间，包含很多冗余和噪声。我们希望通过降维的方法来寻找数据内部的特性，从而提升特征表达能力，降低训练复杂度。主成份分析作为降维中最经典的方法，是一种线性、非监督、全局的降维方法。

1.最近重构性——最小平方误差

给定一系列样本点，怎么样才能用一个超平面来对所有样本进行恰当的表示呢？线性回归目标是求解出一个线性函数，使线性函数对应的直线可以更好的拟合样本。那么模仿线性回归，并扩展到高维空间，使用超平面对所有样本进行恰当的表示其实是找到一个 d 维超平面，使得所有数据点到这个超平面的距离和最小。令 $d=1$ 该问题就退化成了求解一条直线，使所有点到该直线的距离平方和最小。

假设数据集中每个点 x_k 到 d 维超平面的距离为

$$distance(x_k, D) = ||x_k - \widetilde{x_k}||_2 \tag{1}$$

其中 $\widetilde{x_k}$ 表示 x_k 在超平面 D 上的投影向量。根据线性代数，假设超平面是由一组正交基 $W = \{w_1, w_2, ..., w_d\}$ 构成，那么 $\widetilde{x_k}$ 可以由这组正交基线性表示

$$\widetilde{x_k} = \sum_{i=1}^{d} ({w_i}^T x_k) w_i \tag{2}$$

其中 $w_i^T x_k$ 表示 x_k 在 w_i 方向上投影的长度。实际上， $\widetilde{x_k}$ 就是 x_k 在 W 这组正交基表示的超平面上的坐标。

根据最近重构性，可以得到PCA待优化的目标为找到一组 W， 使得所有数据点到 W 所代表的超平面的距离和最小。即

$$\begin{cases} \underset{w_1,w_2,...w_d}{\arg\min} \sum_{k=1}^{n} ||x_k - \widetilde{x_k}||_2^2, \\ \text{s.t. } {w_i}^T w_j = \begin{cases} 1, i = j; \\ 0, i \neq j. \end{cases} \end{cases} \tag{3}$$

将式(3)中的距离展开

$$\begin{aligned} ||x_k - \widetilde{x_k}||_2^2 &= (x_k - \widetilde{x_k})^T (x_k - \widetilde{x_k}) \\ &= {x_k}^T x_k - 2{x_k}^T \widetilde{x_k} + \widetilde{x_k}^T \widetilde{x_k} \end{aligned} \tag{4}$$

其中第一项与 W 无关，所以当作常数看待。将式(2)带入式(4)中的后两项可以得到

$$\begin{aligned} {x_k}^T \widetilde{x_k} &= {x_k}^T \sum_{i=1}^{d} ({w_i}^T x_k) w_i \\ &= \sum_{i=1}^{d} {w_i}^T x_k {x_k}^T w_i \end{aligned}$$

$$\begin{aligned} \widetilde{x_k}^T \widetilde{x_k} &= \left(\sum_{i=1}^{d} ({w_i}^T x_k) w_i \right)^T \left(\sum_{i=1}^{d} ({w_i}^T x_k) w_i \right) \\ &= \sum_{i=1}^{d} ((w_i^T x_k) w_i)^T ((w_i^T x_k) w_i) \\ &= \sum_{i=1}^{d} ({w_i}^T x_k)({w_i}^T x_k) = \sum_{i=1}^{d} ({w_i}^T x_k)({x_k}^T w_i) \\ &= \sum_{i=1}^{d} {w_i}^T x_k {x_k}^T w_i \end{aligned}$$

注意到 $\sum_{i=1}^{d} {w_i}^T x_k {x_k}^T w_i$ 实际上就是矩阵 $W^T x_k {x_k}^T W$ 的迹，于是可以将式(4)继续化简

$$||x_k - \widetilde{x_k}||_2^2 = -\sum_{i=1}^{d} {w_i}^T x_k {x_k}^T w_i + {x_k}^T x_k$$

所有数据的距离和就等于

$$\sum_{k=1}^{n}||x_k-\widetilde{x_k}||_2^2=-\sum_{i=1}^{d}{w_i}^T\sum_{k=1}^{n}(x_k{x_k}^T)w_i+\sum_{k=1}^{n}{x_k}^Tx_k$$
$$=-\sum_{i=1}^{d}{w_i}^TXX^Tw_i+C$$

进一步我们可以把优化问题(3)转换成如下形式

$$\begin{cases}\underset{w_1,w_2,...w_d}{\arg\min}-\sum_{i=1}^{d}{w_i}^TXX^Tw_i,\\ \text{s.t. }{w_i}^Tw_j=\begin{cases}1,i=j;\\0,i\neq j.\end{cases}\end{cases}\tag{5}$$

2.最大可分性——最大方差

PCA旨在找到数据中的主成份，并利用这些主成份表征数据。使用数据的主成份来表征数据可以使信息损失达到最少。如果将数据的主成份看作是一个超平面，那么数据在该超平面上的投影将尽可能的分开。数据的投影尽可能的分开意味着数据在该超平面上尽可能的分散，也就意味着数据在这个超平面上每个正交基上投影的方差都尽可能的大。

假定所有数据进行了中心化，即均值为0。 数据 x_k 在超平面上的投影为 W^Tx_k 其中， $W=\{w_1,w_2,...,w_d\}$ 为超平面的正交基向量构成的矩阵。因此所有数据在正交基 w_i 上投影的方差为：

$$\begin{aligned}D(x,w_i)=\frac{1}{n}\sum_{k=1}^{n}({w_i}^Tx_k)^2&=\frac{1}{n}\sum_{k=1}^{n}({w_i}^Tx_k)({w_i}^Tx_k)^T\\&=\frac{1}{n}\sum_{k=1}^{n}{w_i}^Tx_kx_k^Tw_i\\&={w_i}^T\left(\frac{1}{n}\sum_{k=1}^{n}x_k{x_k}^T\right)w\\&=\frac{1}{n}{w_i}^TXX^Tw\end{aligned}\tag{6}$$

显然 XX^T 是样本的协方差矩阵，因此我们可以根据最大可分性得到如下优化条件：

$$\begin{cases}\underset{W}{\arg\max}\,{w_i}^TXX^Tw,\\ \text{s.t. }WW^T=I.\end{cases}\tag{7}$$

3.优化方程的求解

通过使用最近重构性和最大可分性分别获得优化问题(5)和(7)。显然式子(5)可以分解分求 d 项 $-{w_i}^T X X^T w_i$ ，即和式子(7)的优化问题相同。接下来通过描述求解式子(7)来介绍PCA优化问题的求解。

首先对公式(7)使用拉格朗日乘子法可得

$$XX^T w_i = \lambda w_i \tag{8}$$

公式(8)很容易看出来 w_i 的解为协方差矩阵 XX^T 的特征向量，并且最大的方差就位协方差矩阵最大的特征值。由此我们可以计算出数据的协方差矩阵特征值前 d 大的特征值对应的特征向量作为最优的超平面。

由此我们可以总结PCA的求解过程如下：

1. 对样本数据进行预处理。
2. 求样本数据的协方差矩阵。
3. 对协方差矩阵进行特征值分解，并将特征值按照从大到小顺序排列。
4. 取特征值前 d 大对应的特征向量 $w_1, w_2, ...w_d$ ，并将数据映射到 d 维超平面上。

作为最常用的降维方法，怎么样确定降维后的特征维度 d 是很关键的。可以通过使用不同的 d 值对开销较小的分类器进行交叉验证，选取效果最好的值。另外还可以通过降维后的信息占比来确定 d。定义降维后的信息占比为 $\eta = \sqrt{\frac{\sum_{i=1}^{d} {\lambda_i}^2}{\sum_{i=1}^{n} {\lambda_i}^2}}$。可以设置一个阈值，并找到使信息占比大于等于阈值的最小的 d 值。

线性判别分析(LDA)

线性判别分析是一种有监督学习算法，同时也经常被用来进行数据降维。

1.二分类问题

LDA是有监督的算法，核心思想是最大化类间距离和最小化类内距离。我们从简单的二分类问题出发，并在下一节中扩展到多分类问题。

假设有 C_1、C_2 两个类别的样本，两种类别的均值分别为 $\mu_1 = \frac{1}{N}\sum_{x \in C_1} x, \mu_2 = \frac{1}{N}\sum_{x \in C_2} x$ 。若希望投影后样本两类之间的距离尽可能大，距离表示为

$$D(C_1, C_2) = ||\widetilde{\mu_1} - \widetilde{\mu_2}||_2^2 \tag{1}$$

其中 $\widetilde{\mu_1}, \widetilde{\mu_2}$ 表示两个类中心在 w 上的投影向量。 $\widetilde{\mu_1} = w^T \mu_1, \widetilde{\mu_2} = w^T \mu_2$, 因此，最大化类间距离可以表示为下面式子

$$\begin{cases}\max\limits_{w} ||w^T(\mu_1-\mu_2)||_2^2, \\ \text{s.t.} w^Tw=1.\end{cases} \tag{2}$$

容易发现，当 w 与 $\mu_1-\mu_2$ 方向一致时，该距离达到最大值。

为了得到数据集的类内距离，我们将全部数据的类内距离定义为各个类分别的方差之和。

$$D(C_1)=\sum_{x\in C_1}(w^Tx-w^T\mu_1)^2=\sum_{x\in C_1}w^T(x-\mu_1)(x-\mu_1)^Tw$$

$$D(C_2)=\sum_{x\in C_2}w^T(x-\mu_2)(x-\mu_2)^Tw \tag{3}$$

接下来，我们定义LDA的优化目标为最大化类间距离和类内距离的比值。

$$\begin{aligned}\max_w J(w)&=\frac{||w^T(\mu_1-\mu_2)||_2^2}{\sum_{i=1}^2\sum_{x\in C_i}w^T(x-\mu_i)(x-\mu_i)^Tw} \\ &=\frac{w^T(\mu_1-\mu_2)(\mu_1-\mu_2)^Tw}{\sum_{i=1}^2\sum_{x\in C_i}w^T(x-\mu_i)(x-\mu_i)^Tw}\end{aligned} \tag{4}$$

定义两类数据的类间散度矩阵 $S_b=(x-\mu_i)(x-\mu_i)^T$，类内散度矩阵为两个类别的散度矩阵的和 $S_w=\sum_{i=1}^2\sum_{x\in C_i}(x-\mu_i)(x-\mu_i)^T$。则式子(4)可以重新写成

$$\max_w J(w)=\frac{w^TS_bw}{w^TS_ww} \tag{5}$$

对式子(5)求导，令倒数等于0。可以解得

$$(w^TS_ww)S_bw=(w^TS_bw)S_ww \tag{6}$$

由于 (w^TS_ww) 和 (w^TS_bw) 是两个标量,我们令 $\lambda=J(w)=\frac{w^TS_bw}{w^TS_ww}$。可以将式子(6)转化成下列形式：

$$\begin{aligned}S_bw&=\lambda S_ww \\ S_w^{-1}S_bw&=\lambda w\end{aligned} \tag{7}$$

可以看出，在对优化方程(5)求解的过程中，转化得到了一个求矩阵特征向量的问题。$J(w)$ 就对应着矩阵 $S_w^{-1}S_b$ 最大的特征值，最优 w 就是最大特征值对应的特征向量。

可以发现，类间散度矩阵 $S_b = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$，因此 S_b 和 $(\mu 1 - \mu_2)$ 的方向始终一致（$(\mu_1 - \mu_2)w$为一标量）。如果我们只考虑 w 的方向，可以直接求得 $w = S_s^{-1}(\mu_1 - \mu_2)$。

2.多分类问题

将LDA扩展到多类情况，假设有 c 个类别，并且需要将特征降到 d 维。我们需要找到一个 d 维的超平面 $W = \{w_1, w_2, ...w_d\}$ 来使投影后的样本满足LDA的目标。

由于多类问题中，类间散度矩阵无法定义，所以我们引入全局散度矩阵

$$\begin{aligned} S_t &= S_b + S_w \\ &= \sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T \end{aligned}$$

其中， μ 是全部数据的均值向量。由此可以得到

$$\begin{aligned} S_b &= S_t - S_w \\ &= \sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T - \sum_{i=1}^{c}\sum_{x \in C_i}(x - \mu_i)(x - \mu_i)^T \\ &= \sum_{i=1}^{c}\left(\sum_{x \in C_i}(x - \mu)(x - \mu)^T - \sum_{x \in C_i}(x - \mu_i)(x - \mu_i)^T\right) \\ &= \sum_{i=1}^{c} m_i(\mu_i - \mu)(\mu_i - \mu)^T \end{aligned}$$

其中 m_i 是第 i 类的样本个数。由上式可以看出，类间散度表示的就是每个类别的中心到全局中心的加权距离。最大化类间散度实际上就是在投影后，使每个类别的中心距离全局中心足够远。

根据LDA的定义，我们可以将优化目标定义如下：

$$J(w) = \frac{tr(W^T S_b W)}{tr(W^T S_w W)}$$

得到 S_b 和 S_w 后，最优解 $W = \{w_1, w_2, ...w_d\}$可以通过求解 $S_w^{-1}S_b$ 的前 d 个最大的特征向量对应的特征值来得到。

3.PCA与LDA的区别和联系。

作为两个最常用的降维方法，PCA和LDA都假设数据的分布服从正态分布。虽然从求解优化方程的过程看，PCA和LDA都是用了特征值分解的方法，但是两种方法的基本思想存在着很大的不同。由于PCA是无监督方法，它假设方差越大，信息量越多。通过最大化数据方差来找到最优超平面来达到降维的效果。而LDA是有监督的学习方法，它选择数据投影后类内方差最小、类间方差最大的方向。它用到了类别信息，使得不同类别的数据投影后尽可能被分开。

从实际应用方面来看，PCA主要用于用于去除噪声，进行特征选择；而LDA可以得到分类特征，使得降维后的特征具有更好的可分性。并且LDA可以用于分类问题。