PCA算法原理

主成分分析（Components Analysis，PCA）是机器学习中最经典的降维方法，也是面试中的家常便饭，因而有必要认真梳理一遍PCA的原理，甚至需要手动推导一遍。

数据降维

在理解PCA的概念之前，我们先来认识一下什么是数据降维。**降维就是用低维度的向量来表示原始高维度的特征。**

例如：三维空间中分布在同一个平面上的一些点，用x,y,z三个轴来表示，就需要三个维度；而实际上，因为这些点是分布在一个平面上的，所以可以通过坐标系的旋转变换使得只需要x,y两个轴来表示这些点的数据关系，而且不会有任何损失，从而达到数据降维的目的。

降维的作用： 1.增大样本密度，可以缓解维数灾难

2.减小计算开销

3.去噪

PCA概念

PCA是**数据降维**的一种方式，**旨在找到数据中的主成分，并利用这些主成分来表征原始数据**。简单地说，就是将n维的特征映射到k维上（k<n），这k维的正交特征，就是主成分。

用周志华的《机器学习》书上的话来理解就是：对于正交特征空间中的样本点，如何用一个超平面（直线的高维推广）来对所有的样本进行恰当的表达？如果存在这样的超平面（我的理解就是由k维特征重构出的主成分），那么它应该具有这样的性质：

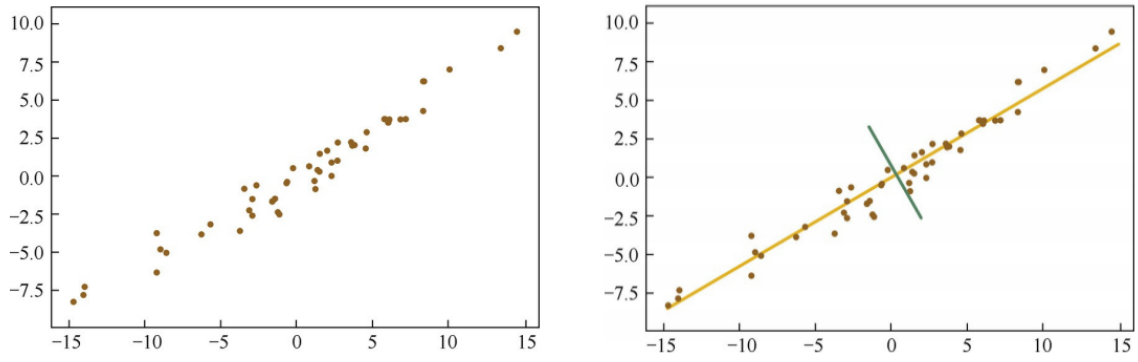
①最大可分性：样本点在这个超平面上的投影尽可能的分开（最大化方差）

②最近重构性：样本点到这个超平面的距离都足够近（最小化平方误差）

如何理解最大可分性（最大方差）和最近重构性（最小平方误差）这两种性质呢？以及怎样才能找到这个k维的主成分呢？下面分别展开分析：

PCA之最大可分性（最大方差）

在信号处理中认为，信号具有较大的方差，噪声具有较小的方差，两者的比值称之为信噪比。信噪比越大意味着数据的质量越好，因此，我们很容易想到**PCA的优化目标，就是最大化投影方差。换种说法就是，让数据在某个超平面（主轴）上投影的方差最大**。



（a）二维空间中经过中心化的一组数据　　（b）该组数据的主成分

理解了最大方差的含义和PCA的优化目标，接下来将是具体的公式推导。

最大化方差公式推导

①给定一组样本点$\{v_1, v_2, \ldots, v_n\}$，首先将其中心化后表示为
$\{x_1, x_2, \ldots, x_n\} = \{v_1 - \mu, v_2 - \mu, \ldots, v_n - \mu\}$，其中，$\mu = \frac{1}{n}\sum_{i=1}^{n} v_i$。

②因为一个向量x_i在ω（单位方向向量）上的投影可以表示为两者的内积$< x_i, \omega >= x_i^T\omega$，而**PCA的目标就是找到一个投影方向$\omega$，使得所有的数据$\{x_1, x_2, \ldots, x_n\}$在$\omega$上的投影方差尽可能地大**，因此：

③投影后的方差可以表示为：

$$
\begin{aligned}
D(x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i^T\omega)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^T\omega)^T(x_i^T\omega) \\
&= \frac{1}{n}\sum_{i=1}^{n}\omega^T x_i x_i^T\omega \\
&= \omega^T\left(\frac{1}{n}\sum_{i=1}^{n}x_i x_i^T\right)\omega
\end{aligned}
\tag{1}
$$

④然后可以发现，上面大括号内的$\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$就是**原始样本的协方差矩阵**，令其等于$\Sigma$。

> 补充理解：协方差公式形式：
>
> $$Cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y) \tag{2}$$
>
> 在均值 $\mu = 0$ 时，（当n足够大时，n-1可以约等于n），于是有：
>
> $$Cov(x, y) = \frac{1}{n}\sum_{i=1}^{n}x_i y_i \tag{3}$$
>
> 对于步骤③中 x_i，是原始样本 v_i 中心化后的，因此可以说 $\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$ 就是**原始样本的协方差矩阵**。

⑤因此，上面的最大化方差D(x)的优化问题可以转化为

$$\begin{cases} max\{\omega^T \Sigma \omega\}, \\ s.t.\ \ \omega^T \omega = 1. \end{cases} \tag{4}$$

其中，ω是单位向量，因此有 $\omega^T \omega = 1.$

⑥对于上面的优化目标，可以构造拉格朗日函数来解决：

$$L(\omega) = \omega^T \Sigma \omega + \lambda(1 - \omega^T \omega) \tag{5}$$

对 ω 求导并令其等于0，可得 $\Sigma\omega = \lambda\omega$

> 补充一点矩阵微分的知识，有助于理解上式的求导过程：（非常有用的公式！！！可以记住！）
>
> $$\begin{aligned} &① \ \frac{\partial x^T a}{\partial x} = \frac{\partial a^T x}{\partial x} = a \\ &② \ \frac{\partial x^T A x}{\partial x} = (A + A^T)x \end{aligned} \tag{6}$$
>
> 因此，对拉格朗日函数的求导便很容易理解了：
>
> $$\frac{\partial L(\omega)}{\partial \omega} = (\Sigma + \Sigma^T)\omega - \lambda(I + I^T)\omega \tag{7}$$
>
> 其中I为单位矩阵
>
> 还记得Σ是什么吗？由上面定义可知，$\Sigma = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$，很显然，Σ的转置$\Sigma^T$=Σ，因此上面可化简为：
>
> $$\frac{\partial L(\omega)}{\partial \omega} = 2\Sigma\omega - 2\lambda\omega \tag{8}$$
>
> 令其等于0，便可得 $\Sigma\omega = \lambda\omega$

⑦由此，最终可以得到最大方差：

$$D(x) = \omega^T \Sigma \omega = \lambda \omega^T \omega = \lambda \tag{9}$$

至此，公式推导已经完成，现在我们不难看出，**x 投影后的方差就是协方差矩阵的特征值**，理解了这一点一切就很清晰了。**因此，我们要找到最大的方差，也就是相当于要求协方差矩阵的最大特征值，而最佳投影方向就是最大特征值所对应的特征向量。**

PCA求解过程总结

(1) 对原始样本进行中心化处理，即零均值化

(2) 求出样本的协方差矩阵 $\Sigma = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$

(3) 求解协方差矩阵的特征值和特征向量

(4) 将特征值由大到小排列，取出前 k 个特征值对应的特征向量

(5) 将 n 维样本映射到 k 维，实现降维处理。

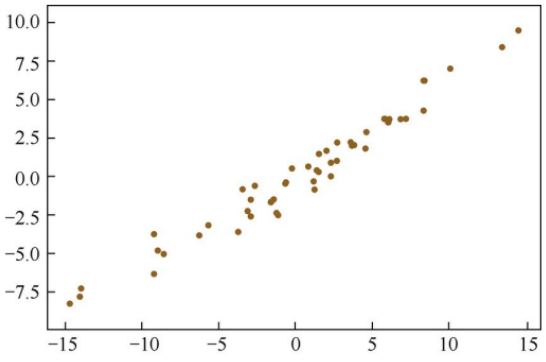
$$x_i^{'} = \begin{bmatrix} \omega_1^T x_i \\ \omega_2^T x_i \\ \vdots \\ \omega_k^T x_i \end{bmatrix} \tag{10}$$

新的$x_i^{'}$的第k维就是x_i在第k个主成分ω_k方向上的投影.

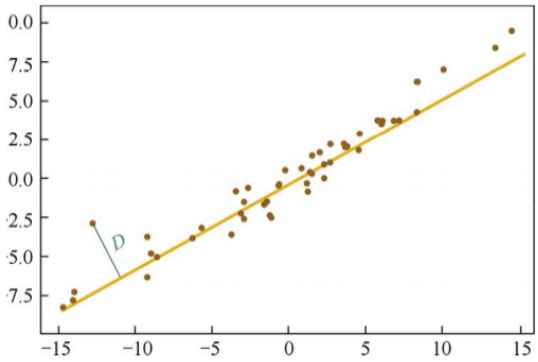
PCA之最近重构性（最小平方误差）

如何理解最近重构性或最小平方误差呢？我们先回顾一下前面讲的最大化方差方法：对于二维空间中的样本点，最大化方差的思想是找到一条直线，使得样本点投影到该直线上的方差最大。因此，也很容易想到，**我们可以找到一条直线来更好的拟合这些样本点。从这个角度来理解，求解PCA的问题就可以转化为一个回归问题了。**

上面说的是二维空间，可以用直线来拟合，那对于高维空间呢？当然也是可以的。超平面是直线在高维空间的推广，因此，最大化方差就是寻找一个超平面使得样本点在超平面上的投影方差最大，而**最小平方误差就是寻找一个超平面使得样本点到这个超平面的距离平方和最小，也就是最近重构性**。



二维空间中经过中心化的一组数据



最小化样本点到直线的距离平方之和

下面给出最小化平方误差的优化目标，具体推导就不展开啦~（一般熟悉最大化方差的推导面试就够用了，最小化平方误差的推导作为了解，如果有需要可以参考《百面机器学习》这本书）

最小化平方误差优化目标

①假设超平面D由k个标准正交基$W = \{\omega_1, \omega_2, \ldots, \omega_k\}$构成，$\widehat{x_d}$是样本点$x_d$（中心化后）在超平面D上的投影向量，则每个样本点到 k 维超平面D的距离为：

$$dist(x_i, D) = ||x_d - \widehat{x_d}||_2 \tag{11}$$

其中，投影向量$\widehat{x_d}$可以通过k维正交基线性表示为$\widehat{x_d} = \sum_{i=1}^{k}(\omega_i^T x_d)\omega_i$，而$\omega_i^T x_d$是$x_d$在$\omega_i$方向上的投影长度

②则PCA的优化目标为：

$$\begin{cases} \underset{\omega_1,\ldots,\omega_k}{arg\,min} \sum_{d=1}^{n} ||x_d - \widehat{x_d}||_2^2, \\ s.t. \quad \underset{\forall i,j}{\omega_i^T \omega_j} = \delta_{i,j} = \begin{cases} 1, & i = j; \\ 0, & i \neq j. \end{cases} \end{cases} \tag{12}$$

通过进一步化简，可得优化目标为：$\begin{cases} \underset{W}{arg\,max}\ tr(W^T X X^T W), \\ s.t. \quad W^T W = I. \end{cases}$

因此，可以发现，最小化方差方法求解问题的形式，和最大化方差方法是一致的，因此也同样可以通过求解协方差的特征值所对应的特征向量，从而得到降维后的主成分。

重要的事情不妨多说一遍，因此这里再把求解过程写一遍：

PCA求解过程总结

(1) 对原始样本进行中心化处理，即零均值化

(2) 求出样本的协方差矩阵 $\Sigma = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$

(3) 求解协方差矩阵的特征值和特征向量

(4) 将特征值由大到小排列，取出前 k 个特征值对应的特征向量

(5) 将 n 维样本映射到 k 维，实现降维处理。

$$x_i^{'} = \begin{bmatrix} \omega_1^T x_i \\ \omega_2^T x_i \\ \vdots \\ \omega_k^T x_i \end{bmatrix} \tag{13}$$

新的$x_i^{'}$的第k维就是x_i在第k个主成分ω_k方向上的投影.

PCA的优缺点

优点： ①它是无监督学习算法，完全无参数限制。

②降维，减小计算开销

缺点： ①特征值分解有一些局限性，比如变换的矩阵必须是方阵

②如果用户对观测对象有一定的先验知识，掌握了数据的一些特征，却**无法通过参数化等方法对处理过程进行干预**，可能会得不到预期的效果，效率也不高

总结

PCA是一种线性的、无监督的、全局的降维算法。

PCA的应用也很广泛，这里列举其中几项：

①数据降维

②去噪

③高位数据集的可视化

④数据压缩

⑤图像分析

参考资料

周志华--《机器学习》

葫芦娃--《百面机器学习》

机器学习--主成分分析(PCA)算法的原理及优缺点