PCA 降维

华南师范大学 数学科学学院 朱致鹏

前面我们说过 MDS 通过线性变换来进行有效降维，若要求低维子空间对样本具有最大可分性，则得到一种极为常用得线性降维方法——PCA

PCA 算法本质就是找一些投影方向，使得数据在这些投影方向上的方差最大，而且这些投影方向是相互正交的。这其实就是找新的正交基的过程，计算原始数据在这些正交基上投影的方差，方差越大，就说明在对应正交基上包含了更多的信息量。原始数据协方差矩阵的特征值越大，对应的方差越大，在对应的特征向量上投影的信息量就越大。反之，如果特征值较小，则说明数据在这些特征向量上投影的信息量很小，可以将小特征值对应方向的数据删除，从而达到了降维的目的。

对于正交属性空间中得样本点，如何让用一个超平面对所有样本进行恰当得表达？

这样的超平面由下面两个性质：

①最大可分性：样本点在这个超平面上的投影能尽可能分开

②最近重构性：样本点到这个超平面得距离都足够近

①首先针对最大可分性，它要使投影后的样本点的方差最大化!!!

假设原始数据$\{x_1, x_2, x_3, ..., x_n\}$，先对样本点中心化（归一化）可使得特征均值等于 0

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

中心化后$\{z_1, z_2, z_3, ..., z_n\} = \{x_1 - \mu, x_2 - \mu, ..., x_n - \mu\}$，向量$z_i$在 w（超平面）上的投影为$w^T z_i$。方差$D(z) = \frac{1}{n}\sum_{i-1}^{n}(w^T z_i)^T (w^T z_i) = w^T(\frac{1}{n}\sum_{i=1}^{n} z_i z_i^T)w$，其中 z 的方差为$tr(\frac{1}{n}\sum_{i=1}^{n} z_i z_i^T)$

要令方差尽可能大，即求解最优化问题（含等式约束），用拉格朗日乘数法，接下来先介绍求解方法：

（一）无约束

形如：$\min \quad f(x) = x^2 - 2x + 1$

方法：高中所学求导求极值

（二）有约束

1. 等式约束

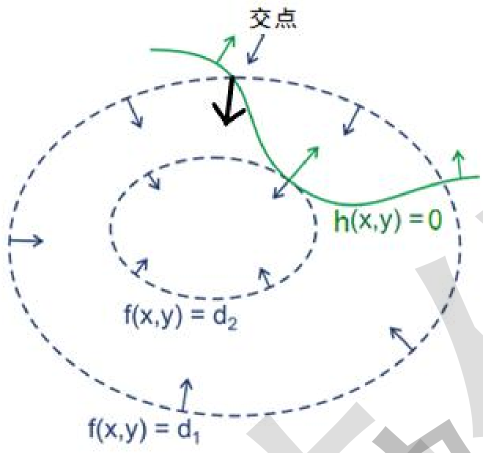
$$\min \quad f(x)$$
$$\{h_i(x) = 0 \quad i = 1, 2, .., n$$

从几何角度看，该问题的目标是在方程$h(x) = 0$确定的 d-1 维曲面上寻找能

使目标函数 $f(x)$ 最小化的点。此时不难得到如下结论：

①对于约束曲面上的任意点 x，该点的梯度 $\nabla h(x)$ 正交于约束曲面

这点很容易理解，因为梯度是下降最快的方向，显然该点曲面正交，下图可以帮助理解：



②在最优点 x^*，目标函数在该点的梯度 $\nabla f(x^*)$ 正交于约束曲面

可以用反证法解释：如果 $\nabla f(x^*)$ 不正交于约束曲面，那么在约束曲面上，我们总可以沿着 $\nabla f(x^*)$ 的相反方向移动，使 $f(x)$ 函数值更进一步下降，因为梯度的反方向就是函数值下降最快的方向，所以如果找到了最优点 x^*，则该点梯度 $\nabla f(x^*)$ 必定正交于约束曲面，使我们无法在约束曲面上沿着梯度的负方向移动，就像图中的交点。

由此可知，在最优点 x^*，梯度 $\nabla h(x)$ 和 $\nabla f(x)$ 的方向必定相同或相反，即存在 $\lambda \neq 0$ 使得

$$\lambda \nabla h(x^*) + \nabla f(x^*) = 0 \qquad (8)$$

λ 称为拉格朗日乘子。定义拉格朗日函数

$$L(x, \lambda) = f(x) + \lambda h(x) \qquad (9)$$

可以发现，对（9）求 x 的偏导数其实就是（8），对（9）求 λ 得偏导数即得到约束条件 $h(x) = 0$。于是，原约束优化问题可以转化为对拉格朗日函数得无约束优化问题。

2. 不等式约束

$$\min \quad f(x)$$
$$\{ h_i(x) \leq 0 \quad i = 1, 2, .., n$$

这种情况得最优点 x^* 或在 h(x)=0 边界上，或在 h(x)<0 的区域中，对于 h(x)<0 的情形，约束 h(x)<=0 将不起作用，可直接通过条件 $\nabla f(x)$=0 来获得最优点；这等价于将 λ 置零后对拉格朗日函数求 x 的偏导数置零得到最优点。那 h(x)=0 的时候类似于等式约束，需要注意的是，梯度 $\nabla h(x)$ 和 $\nabla f(x)$ 的方向必定相反，即存在常数 λ>0 使得 $\lambda\nabla h(x^*)+\nabla f(x^*)=0$。结合两种情况，必满足 $\lambda h(x)=0$，此时的约束变成如下：

$$\begin{cases} h(x)\le 0 \\ \lambda \ge 0 \\ \lambda h(x)=0 \end{cases} \quad (10)$$

称为 KKT 条件。

3. 多约束

推广到多约束：

$$\min \quad f(x)$$

$$\begin{cases} h_i(x)\le 0 \quad i=1,2,..,n \\ g_j(x)=0 \quad j=1,2,..,m \end{cases}$$

此时拉格朗日函数为：

$$L(x,\lambda,\mu)=f(x)+\mu_i\sum_{i=1}^{n}h_i(x)+\lambda_j\sum_{j=1}^{m}g_j(x) \quad (11)$$

KKT 条件：

$$\begin{cases} h_i(x)\le 0 \\ \mu_i \ge 0 \\ \mu_i h_i(x)=0 \end{cases} \quad (12)$$

回到求解方差问题上，我们要求解最优化问题：

$$\begin{cases} \max \quad tr(W^TZZ^TW) \\ s.t. \quad W^TW=I \end{cases}$$

建立拉格朗日函数：

$$L(w,\lambda)=tr(W^TZZ^TW)+\lambda)(W^TW-I)$$

对 w 偏导置零：

$$\nabla_w L(w,\lambda)=2(ZZ^T)W+2\lambda W=0$$

得到：

$$Z^TZW=\lambda W$$

代入原问题：

$$tr(W^T ZZ^T W)=tr(W^T \lambda W)=\lambda$$

即要求$Z^T Z$的特征值及对应的特征向量构成的 W。

②最近重构性（最小平方误差）

样本点到超平面的距离：

$$dist(x,D)=\left\|x_k-\hat{x}_k\right\|_2$$

其中，$\hat{x}_k$是x_k在超平面 D 上的投影向量：$\hat{x}_k=\sum_{j=1}^{d}(w_j{}^T x_k)w_j=\sum_{j=1}^{d}z_j w_j$。

（$w_j{}^T x_k$是x_k在w_j方向上投影的长度）

要令$dist(x,D)$最小，即求解如下问题：

$$\begin{cases}\min \quad \sum_{k=1}^{n}\left\|x_k-\hat{x}_k\right\|_2^2 \\ s.t. \quad w_j^T w_i=\begin{cases}1, i=j\\0, i\neq j\end{cases}\end{cases}$$

$$\left\|x_k-\hat{x}_k\right\|_2^2=(x_k-\hat{x}_k)^T(x_k-\hat{x}_k)=x_k^T x_k-x_k^T\hat{x}_k-\hat{x}_k{}^T x_k+\hat{x}_k{}^T\hat{x}_k$$

$$=x_k^T x_k-2x_k^T\hat{x}_k+\hat{x}_k{}^T\hat{x}_k=x_k^T x_k-\sum_{j=1}^{d}w_j^T x_k x_k^T w_j$$

那么

$$\min \quad \sum_{k=1}^{n}\left\|x_k-\hat{x}_k\right\|_2^2=\sum_{k=1}^{n}(-tr(W^T x_k x_k^T W)+x_k^T x_k$$

那么，原问题转化为：

$$\begin{cases}\min \quad -tr(W^T XX^T W)\\ s.t. \quad W^T W=I\end{cases}$$

与①一致。