Dimensionality Reduction 降维

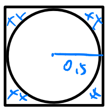
一、背景

降维是抑制过拟合的手段之一。

维度灾难：每增加一个特征，若加二值，它也会以指数的速度增长。

几何角度：

例一、



当维度增加时

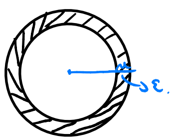
$V_{超立方体} = 1$

$V_{超球体} = K \cdot r^{D}$

当D增大时，$V_{超球体} \to 0$。

故在高维空间中，立方体几乎是空的，所以在高维空间中，样本分布在角落中，造成样本的稀疏性，并且分布不均，这种分布很难做分类。

内部的圆表示空间内部。



$V_{外} = K \cdot 1^{D} = K$

$V_{环} = V_{外} - V_{内} = K - K \cdot (1-\varepsilon)^{D}$

当 $D \to \infty$ 时，$V_{环} \to K$。

同样，可以发现，在D很大时，其内部几乎不存在任何空间，这导致其样本都分部在了外球壳上。

降维方法 →
- 直接降维(即特征选择)：如L1 (Lasso)
- 线性降维：PCA，MDS(多维缩放)
- 非线性降维：流形

二、样本均值 & 方差矩阵

$$X=(x_1,x_2,\cdots x_N)^T=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}_{N\times p}$$

$x_i\in\mathbb{R}^p,\ i=1,2,\cdots N.$

样本均值：$\bar{X}_{p\times 1}=\frac{1}{N}\sum_{i=1}^{N}x_i$

样本方差：$S_{p\times p}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{X})(x_i-\bar{X})^T$

$$\bar{X}=\frac{1}{N}\sum_{i=1}^{N}x_i=\frac{1}{N}(x_1\ x_2\ \cdots x_N)\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}_{N\times 1}=X^T\cdot 1_N \qquad 1_N=\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}_{N\times 1}$$

$$S=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{X})(x_i-\bar{X})^T$$

$$=\frac{1}{N}\underbrace{(x_1-\bar{X},\ x_2-\bar{X},\ \cdots\ x_N-\bar{X})}_{(x_1\ x_2,\cdots x_N)-\bar{X}(1\ 1\ \cdots\ 1)}\underbrace{\begin{pmatrix}(x_1-\bar{X})^T\\(x_2-\bar{X})^T\\\vdots\\(x_N-\bar{X})^T\end{pmatrix}}$$

$X^T-\bar{X}\cdot 1_N^T$

$X^T-\frac{1}{N}X^T\cdot 1_N\cdot 1_N^T$

$$\frac{1}{N}\cdot X^T\underbrace{(I_N-\frac{1}{N}1_N\cdot 1_N^T)}_{\text{令 } H_N \text{ 为 centering Matrix (中心矩阵)}}\cdot(I_N-\frac{1}{N}1_N\cdot 1_N^T)^T X=\frac{1}{N}X^T H\cdot H^T X$$

H 的性质：$H=H^T$，$H^n=H$

$\therefore S=\frac{1}{N}X^T H X$

PCA：

一个中心：原始特征空间的重构。
将原来线性相关的维度转换成线性无关的维度。

服务于

两个基本点：
- 最大投影方差
- 最小重构距离

最大投影方差：

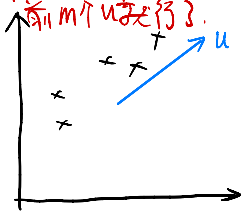
目的：找到 u，使得中心化后的样本在 u 上的投影的方差之和最大。

即最大化 $J = \sum_{i=1}^{N}\left((x_i-\bar{x})^T u\right)^2$

$$= \sum_{i=1}^{N}\left(u^T(x_i-\bar{x})\cdot(x_i-\bar{x})^T u\right)$$

$$= u^T\left(\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T\right)u$$

u就称作主成分，要想降到m维，那就找前m个u就行了。



令 $\|u\|=1$ 即 $u^Tu=1$

样本 x_i 在 u 上的投影为 $x_i\cdot u$，即 $x_i^T u$

均值 $\bar{x}$ 的投影为 $\bar{x}\cdot u$ 即 $\bar{x}^T u$，因此 x_i 在 u 上投影的方差为 $\left((x_i-\bar{x})^T u\right)^2$.

可以乘上一个 $\frac{1}{N}$，对结果无影响，即 $J = \frac{1}{N}\sum_{i=1}^{N}\left((x_i-\bar{x})^T u\right)^2$

$$= u^T\underbrace{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T\right)}_{\text{协方差矩阵 } S}u$$

$$= u^T\cdot S\cdot u$$

即求
$$\begin{cases}\hat{u} = \arg\max\limits_{u} u^T S u \\ s.t.\ u^T u = 1\end{cases}$$

拉格朗日乘子法：$L(u,\lambda) = u^T S u + \lambda(1-u^T u)$

令 $\frac{\partial L}{\partial u} = 2Su - 2\lambda u = 0$，

即 $S\underset{\text{特征向量}}{u} = \underset{\text{特征值}}{\lambda} u$

故 u 是协方差矩阵的特征向量。

最后解得符合条件的向量u是协方差矩阵S的特征向量，共p个，若想降到q维（q<p），可以选择对应特征值最大的前q个特征向量取出来作为投影方向，然后获得数据在这些方向上的投影即为重构坐标，即：

$$\begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix}_{N\times p} \begin{pmatrix} u_1 & u_2 & \cdots & u_q \end{pmatrix}_{p\times q} = \begin{pmatrix} x_1^T u_1 & x_1^T u_2 & \cdots & x_N^T u_q \\ x_2^T u_1 & x_2^T u_2 & \cdots & x_2^T u_q \\ \vdots & & & \vdots \\ x_N^T u_1 & x_N^T u_2 & \cdots & x_N^T u_q \end{pmatrix}_{N\times q}.$$

特征向量表示投影变换的方向，特征值表示投影变换的强度.

最小重构距离.

由于共p个特征向量符合条件，则原来的中心化了的数据可表示为：

$$x_i - \bar{x} = \sum_{k=1}^{p} \left((x_i-\bar{x})^T u_k \right) u_k$$

降维后的数据：

$$\hat{x}_i = \sum_{k=1}^{q} \left((x_i-\bar{x})^T u_k \right) u_k$$

因此重构距离为：

$$(x_i - \bar{x}) - \hat{x}_i$$

故应最小化J，其中：

$$\begin{aligned} J &= \frac{1}{N}\sum_{i=1}^{N} \| (x_i-\bar{x}) - \hat{x}_i \|^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} \left\| \sum_{k=q+1}^{p} \left((x_i-\bar{x})^T u_k \right) u_k \right\|^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} \sum_{k=q+1}^{p} \left((x_i-\bar{x})^T u_k \right)^2 \\ &= \sum_{i=1}^{N} \frac{1}{N} \sum_{k=q+1}^{p} \left((x_i-\bar{x})^T u_k \right)^2 \\ &= \sum_{k=q+1}^{p} \underbrace{\frac{1}{N}\sum_{i=1}^{N} \left((x_i-\bar{x})^T u_k \right)^2}_{u_k^T S \cdot u_k} \\ &= \boxed{\sum_{k=q+1}^{p} u_k^T S \cdot u_k}, \quad s.t.\ u_k^T u_k = 1 \end{aligned}$$

即求解下列最优化问题：

$$\begin{cases} \hat{u} = \arg\min\limits_{u} \sum_{k=q+1}^{p} u_k^T S \cdot u_k \\ s.t.\ u_k^T u_k = 1 \end{cases}$$

tips：

最大投影方差求得的p个特征向量，我们要的是特征值最大的前q个作为新的坐标方向向量；所以剩下的p-q个特征向量是对应的最小的几个特征值，所以最小重构距离所求得的u_k都是对应的最小的几个特征值.

概率角度的PCA (P-PCA)