降维
Dimension Reduction

① 背景

② 样本均值 & 样本方差

③ PCA

④ SVD角度看PCA和PCoA

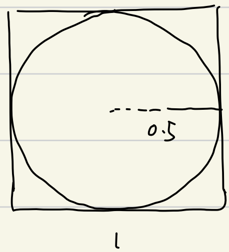
⑤ 概率角度PCA

背景（Background）

① 维度灾难（Curse of Dimensionality）

从几何角度解释.

1)



假设维度为2，有一个圆内嵌于正方形.

正方形边长为1，

则 $V_{正} = 1^2$ $\overset{维度为3}{\Longrightarrow}$ $V_{立方体} = 1^3$

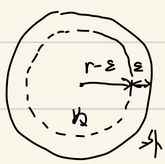
$V_{圆} = \pi 0.5^2$ $V_{球体} = \frac{4}{3}\pi 0.5^3$

维度为D $\Longrightarrow$ $V_{超立方体} = 1^D$

$V_{超球体} = K \cdot 0.5^D$

由上面计算可知：当维度不断增大，则 $V_{超球体}$ 不断逼近于0.
又由于超球体内切于超立方体，则可见超立方体的体积几乎分布于壳上，即样本几乎分布于超立方体表面，内部几乎没有样本

2)



用同心圆举例. 如上述例子所示，令 $r=1$

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

则 $\frac{V_{环}}{V_{外}} = 1 - (1-\varepsilon)^D$ ，当 $D \to \infty$ 时，$(1-\varepsilon)^D \to 0$.

则 $\frac{V_{环}}{V_{外}} = 1$ ，即体积集中在球壳上.

上述两个例子都说明了一件事。数据在高维空间中分布的稀疏性。这就是维度灾难。

②

过拟合

$X \in \mathbb{R}^{N \times p}$

N: 样本量

p: 样本维度

- 增加样本量 $N \uparrow$
- 正则化。$X^TX \to X^TX + \lambda I$，使其可逆，在线性回归有解释
- 降维 $p \downarrow$ ⟶
  - 直接降维 特征选择
  - 线性降维 PCA、MDS
  - 非线性降维 流形
    - Isomap
    - LLE

样本均值 & 样本方差

(Sample Mean & Variance Matrix)

$X = (x_1, x_2, \cdots, x_N)^T$, $x_i \in \mathbb{R}^p$. $X \in \mathbb{R}^{N \times p}$

Sample Mean : $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N} X^T 1_N$

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

转换为矩阵表示 . 令 $1_N = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^{N \times 1}$, $I_N \in \mathbb{R}^{N \times N}$

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$= \frac{1}{N}(x_1 \ x_2 \ \cdots \ x_N)\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

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

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

$$= \frac{1}{N}\begin{pmatrix} x_1-\bar{x} & x_2-\bar{x} & \cdots & x_N-\bar{x}\end{pmatrix}\begin{pmatrix}(x_1-\bar{x})^T \\ (x_2-\bar{x})^T \\ \vdots \\ (x_N-\bar{x})^T\end{pmatrix}$$

对于 $\begin{pmatrix} x_1-\bar{x} & x_2-\bar{x} & \cdots & x_N-\bar{x}\end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \cdots & x_N\end{pmatrix} - \bar{x}\begin{pmatrix}1 & 1 & \cdots & 1\end{pmatrix}$

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

$$= X^T\underbrace{\left(I_N - \frac{1}{N}1_N 1_N^T\right)}_{H_N \to \text{centering matrix}}$$

$$= X^T H$$

则 $S = \frac{1}{N}X^T H\cdot(X^T H)^T = \frac{1}{N}X^T H H^T X$

对于 $H = I_N - \frac{1}{N}1_N 1_N^T$

$$H^T = \left(I_N - \frac{1}{N}1_N 1_N^T\right)^T = I_N - \frac{1}{N}1_N 1_N^T = H$$

$$H^2 = H\cdot H = \left(I_N - \frac{1}{N}1_N 1_N^T\right)\left(I_N - \frac{1}{N}1_N 1_N^T\right)$$

$$= I_N - \frac{2}{N}1_N 1_N^T + \boxed{\frac{1}{N^2}1_N \underbrace{1_N^T 1_N}_{N} 1_N^T} \Rightarrow \frac{1}{N}1_N 1_N^T$$

$$= I_N - \frac{1}{N}1_N 1_N^T$$

$$= H$$

则 $H^n = H \Rightarrow \boxed{S = \frac{1}{N}X^T H\cdot H^T X = \frac{1}{N}X^T H X}$

PCA

一个中心：原始特征空间的重构.

相关 ⟶ 无关

一所学样中每个学生都有学校和姓名两个属性
在学校维度，大家都一样，方差为0，不含信息.

两个基本点：① 最大投影方差　　投影后的数据方差尽可能大，代表所含信息量越多.

② 最小重构误差　　降维后的数据重构回原始数据的代价最小.

① 最大投影方差 (maximum variance perspective)

假设投影的轴为 u_1，且 $u_1^T u = 1$.（不考虑大小，只考虑方向）

首先，对数据中心化：$x_i - \bar{x}$

中心化后的数据在 u_1 上的投影为 $u_1^T (x_i - \bar{x})$

投影后的数据依然是中心化的. → $u_1^T (x_i - \bar{x}) = \underbrace{u_1^T x_i - u_1^T \bar{x}}_{\text{可以看作先投影，再中心化}}$

则方差为
$$\frac{1}{N}\sum_{i=1}^{N}\left[u_1^T (x_i - \bar{x})\right]^2 = \frac{1}{N}\sum_{i=1}^{N} u_1^T (x_i - \bar{x})(x_i - \bar{x})^T u_1$$
$$= u_1^T \left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T\right] u_1$$
$$= u_1^T S u_1$$

则
$$\hat{u}_1 = \underset{u_1}{\arg\max}\ u_1^T S u_1$$
$$\text{s.t. } u_1^T u = 1$$

用拉格朗日乘子法，得：

$$\mathcal{L}(u_1,\lambda_1) = u_1^T S u_1 + \lambda_1(1-u_1^T u)$$

$$\frac{\partial \mathcal{L}(u_1,\lambda_1)}{\partial u_1} = 2Su_1 - 2\lambda_1 u_1 = 0.$$

则 $Su_1 = \lambda_1 u_1$

即 u_1 为 S 的关于特征值 λ 的特征向量

又 ∵ $\hat{u}_1 = \underset{u_1}{\operatorname{argmax}}\ u_1^T S u_1$

对 $Su_1 = \lambda_1 u_1$ 两边左乘 u_1^T，得 $u_1^T S u_1 = \lambda_1$.

∴ λ_1 为 S 的最大特征值

u_1 为 S 的关于最大特征值的特征向量.

上面这是降到一维，如果要降到m维，就需要取m根轴.

$(u_1, u_2, \dots, u_m)$ 其中，对应的特征值为 $\lambda_1 > \lambda_2 > \cdots > \lambda_m$

也就是依次取前m个大的特征值

② 最小重构误差 (minimum error perspective)

PCA其实是用S的所有特征值对应的，相互正交的特征向量重构特征空间

$x_i \in \mathbb{R}^p$，则用 $(u_1, u_2, \dots, u_p)$ 重构特征空间. 假设 x_i 已经经过中心化

x_i 在 u_k 方向上的投影为：$x_i^T u_k$，加上方向 => $(x_i^T u_k)u_k$.

则经过降维后. $\hat{x}_i = \sum_{k=1}^{q}(x_i^T u_k)u_k$ 其中. q表示选取前q个特征向量. 指降到q

原始数据：$x_i = \sum_{k=1}^{p}(x_i^T u_k)u_k$ 维，但是这里 $\hat{x}_i$ 还是p维向量.

则重构误差为: $x_i - \hat{x}_i$

则目标是最小化总体的误差: $J = \frac{1}{N}\sum_{i=1}^{N}\|x_i - \hat{x}_i\|^2 = \frac{1}{N}\sum_{i=1}^{N}\|\sum_{k=q+1}^{p}(x_i^T u_k)u_k\|^2$

由于 $\|au\|_2^2 = a^2\|u\|_2^2$. 且 $u_k^T u_k = 1$

则 $J = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=q+1}^{p}(x_i^T u_k)^2$

$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=q+1}^{p} u_k^T x_i x_i^T u_k$$

$$= \sum_{k=q+1}^{p} u_k^T\left[\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T\right]u_k.$$

由于 x_i 已经过中心化.

则 $J = \sum_{k=q+1}^{p} u_k^T S u_k$.

则 $\{u_k\} = \underset{\{u_k\}}{\arg\min}\sum_{k=q+1}^{p} u_k^T S u_k$

$$\text{s.t. } u_k^T u_k = 1$$

和最大投影方差求解方式一样 特别地,由于 u_k 之间线性无关,则每个 u_k 可以单独进行求解.即 $u_k = \underset{u_k}{\arg\min}\, u_k^T S u_k$. s.t. $u_k^T u_k = 1$.

所以.最终结果是 $\{u_k\}$ 是 S 的前 $p-q$ 个小的特征值所对应的特征向量,即降维是选前 q 个大小特征向量.

和①中的结论相同

SVD角度看PCA和PCoA

① 特征分解

对于方阵 A. ⇒ $Ax=\lambda x$. ⇒ $AX=X\Sigma$ ⇒ $A=X\Sigma X^{-1}$

则 $A=X\Sigma X^{-1}$ 称为A的特征分解.

特别地，当A是对称矩阵时，特征向量是正交的

即 $XX^T=X^TX=I$, ⇒ $A=X\Sigma X^T$

② 奇异值分解

设 $A\in C^{m\times n}$, $rank(A)=r$, 则存在m阶、n阶酉阵U, V.

使 $A=U\begin{pmatrix}\Sigma & 0\\ 0 & 0\end{pmatrix}V^H$, 其中 $\Sigma=diag(\sigma_1,\sigma_2,\cdots,\sigma_r)$

σ_i 为矩阵A的非零奇异值, $\underbrace{U^HU=I}_{\text{列正交}}$, $V^HV=VV^H=I$.

③ PCA

X为数据集.

则对X进行中心化，并对中心化后的数据进行特征值分解.

$HX=U\Sigma V^T$ (为了写得方便，这里使用约化后的奇异值分解，并将共轭转置写为转置).

则 $S=X^THX=X^TH^THX=V\Sigma U^TU\Sigma V^T=\boxed{V\Sigma^2V^T}$ 特征分解

则V为S的特征向量组成的矩阵. Σ^2 为特征值组成的对角矩阵.

即对HX进行PCA后的V即为重构空间的一组基底

假设一个基底为 u_k，则重构后，x_i在u_k方向上的坐标为 $x_i^T u_k$.
写成矩阵形式为 XV.
则中心化后的数据在新空间的坐标为 HXV
化简 $HXV = U\Sigma V^T V = U\Sigma$.
令 $T = HXX^TH^T = U\Sigma V^T V \Sigma U^T = U\Sigma^2 U^T$ （U为T的特征向量组成的矩阵）

$$TU\Sigma = U\Sigma^2 U^T U\Sigma = U\Sigma^3 = (U\Sigma)\Sigma^2$$

可以发现，$U\Sigma$ 也为T的特征向量组成的矩阵，可以看作是U的缩放.
那么，意味着对T进行特征值分解能直接得到坐标.

综上，S和T有相同的eigen value
且 S → 特征分解 → 得到主成份（基底的方向）
T → 特征分解 → 得到坐标.
那么S代表主成份分析（principal component analysis，PCA）
T代表主坐标分析（principal coordiate analysis，PCoA）

概率角度PCA (probabilistic PCA)

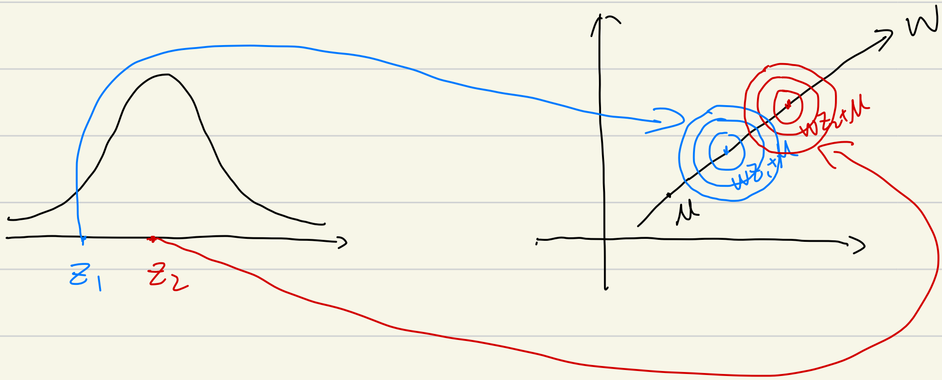
P-PCA

$x \in \mathbb{R}^p$, $z \in \mathbb{R}^q$, $q<p$

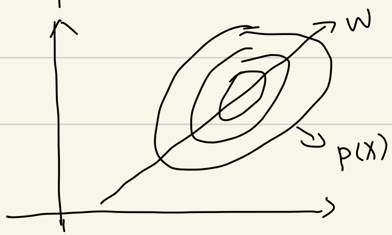
observed data　latent variable.

设 $\begin{cases} z \sim N(0_q, I_q) \\ x \sim Wz+\mu+\varepsilon \\ \varepsilon \sim N(0_p, \sigma^2 I_p), \ \varepsilon \perp z. \end{cases}$ => Linear Gaussian Model.

生成角度看 P-PCA. 取 $q=1$, $p=2$.



当z有足够多的采样数据后，所有z的分布形成的分布如下所示



这就是x的分布.

P-PCA与GMM不同之处在于，GMM中高斯分布的数量是确定的，而P-PCA中高斯分布的数量是采样次数，即GMM中，z是离散的。P-PCA中，z是连续的

对于如何求解P-PCA，$\begin{cases} \text{inference} \rightarrow p(z|x) \\ \text{learning} \rightarrow W, \mu, \sigma^2 \end{cases}$ (可以用最大似然，也可以用EM，这里不讨论learning).

本节主要关心inference，即如何求得 $p(z|x)$.

由已知的线性高斯模型，可得：

$$E[x|z] = Wz + \mu + E[\varepsilon] = Wz + \mu$$

$$Var[x|z] = Var[\varepsilon] = \sigma^2 I.$$

则 $x|z \sim N(Wz+\mu, \sigma^2 I)$.

$$\therefore E[x] = E[Wz+\mu] + E[\varepsilon] = \mu$$

$$Var[x] = Var[Wz] + Var[\varepsilon] = WIW^T + \sigma^2 I = WW^T + \sigma^2 I$$

则 $x \sim N(\mu, WW^T + \sigma^2 I)$

为了求解 $z|x$，可以使用数学基础部分求条件概率的方法.

构建 $\begin{pmatrix} x \\ z \end{pmatrix} \sim N\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \begin{bmatrix} WW^T+\sigma^2 I & \Delta \\ \Delta^T & I \end{bmatrix}\right)$

$$\begin{aligned}\Delta &= Cov(x, z) \\ &= E[(x-\mu)(z-0)^T] \\ &= E[(x-\mu)z^T] \\ &= E[(Wz+\varepsilon)z^T] \\ &= E[Wzz^T] + \underbrace{E[\varepsilon z^T]}_{\varepsilon \perp z} \\ &= WE[zz^T] + E[\varepsilon]E[z^T] \\ &= W \cdot I \\ &= W\end{aligned}$$

则 $\begin{pmatrix} x \\ z \end{pmatrix} \sim N\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \begin{bmatrix} WW^T+\sigma^2 I & W \\ W^T & I \end{bmatrix}\right)$

将 x 看作 x_a，z 看作 x_b，套用

$$\boxed{\begin{cases} x_{b\cdot a} = x_b - \Sigma_{ba}\Sigma_{aa}^{-1}x_a \\ \mu_{b\cdot a} = \mu_b - \Sigma_{ba}\Sigma_{aa}^{-1}\mu_a \\ \Sigma_{bb\cdot a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \end{cases} \Rightarrow x_b \mid x_a \sim N(\mu_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}x_a,\ \Sigma_{bb\cdot a})}$$

得：$z \mid x \sim N\left(W^T(WW^T+\sigma^2 I)^{-1}(x-\mu),\ I - W^T(WW^T+\sigma^2 I)^{-1}W\right)$