PCA 误差分析

数据的变换、降维：

m 个特征方向

$$\begin{aligned} \boldsymbol{y}&=[y_1,y_2,...,y_m]^T \\ &=[\boldsymbol{w}_1^T\boldsymbol{x},\boldsymbol{w}_2^T\boldsymbol{x},...,\boldsymbol{w}_m^T\boldsymbol{x}]^T \\ &=\boldsymbol{W}^T\boldsymbol{x} \end{aligned}, \qquad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}=\begin{bmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \vdots \\ \boldsymbol{w}_m^T \end{bmatrix}\boldsymbol{x}$$

选取 l 个方向：

$$\begin{aligned} \boldsymbol{y}&=[y_1,y_2,...,y_l]^T \\ &=[\boldsymbol{w}_1^T\boldsymbol{x},\boldsymbol{w}_2^T\boldsymbol{x},...,\boldsymbol{w}_l^T\boldsymbol{x}]^T \\ &=\boldsymbol{W}^T\boldsymbol{x} \end{aligned}, \qquad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_l \end{bmatrix}=\begin{bmatrix} \boldsymbol{w}_1^T \\ \boldsymbol{w}_2^T \\ \vdots \\ \boldsymbol{w}_l^T \end{bmatrix}\boldsymbol{x}, \qquad l<m$$

数据的反变换、重构：

无损截断重构（舍弃特征值为 0 的方向）

$$\begin{aligned} \boldsymbol{x}&=\boldsymbol{W}\boldsymbol{y} \\ &=\sum_{j=1}^{m}y_j\boldsymbol{w}_j \end{aligned}, \qquad \boldsymbol{W}^T\boldsymbol{W}=\boldsymbol{I}$$

有损截断重构（舍弃部分特征值大于 0 的方向）

$$\hat{\boldsymbol{x}}=\sum_{j=1}^{l}y_j\boldsymbol{w}_j=[\boldsymbol{w}_1,\boldsymbol{w}_2,...,\boldsymbol{w}_l]\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_l \end{bmatrix}, \quad l<m$$

误差分析

证明 1：误差向量和重构空间是正交的

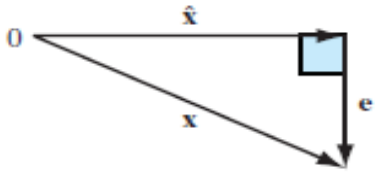
误差向量 = 无损重构 － 有损重构

$$\boldsymbol{e}=\boldsymbol{x}-\hat{\boldsymbol{x}}=\sum_{i=l+1}^{m}a_i\boldsymbol{w}_i\,,$$ 代入下式

$$\begin{aligned}\boldsymbol{e}^T\hat{\boldsymbol{x}}&=\sum_{i=l+1}^{m}a_i\boldsymbol{w}_i\sum_{j=1}^{l}a_j\boldsymbol{w}_j\\&=\sum_{i=l+1}^{m}\sum_{j=1}^{l}a_ia_j\boldsymbol{w}_i^T\boldsymbol{w}_j\\&=0\end{aligned}$$

证明 2：重构误差等于截断特征值之和

误差向量的模长?



分别算一下无损重构、有损重构、误差的 2 范数，

$$\|\boldsymbol{e}\|^2=\|\boldsymbol{x}\|^2-\|\hat{\boldsymbol{x}}\|^2=\sum_{j=l+1}^{m}a_j^2=\sum_{j=l+1}^{m}\boldsymbol{w}_j^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{w}_j\quad,$$

$$\|\hat{\boldsymbol{x}}\|^2=\sum_{j=1}^{l}a_j^2\,,\qquad \text{其中，}\ a_j=\boldsymbol{x}^T\boldsymbol{w}_j$$

$$\|\boldsymbol{e}\|^2=\|\boldsymbol{x}\|^2-\|\hat{\boldsymbol{x}}\|^2=\sum_{j=l+1}^{m}a_j^2=\sum_{j=l+1}^{m}\boldsymbol{w}_j^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{w}_j$$

对于样本集（N 个样本）的平均误差：

$$\varepsilon = \frac{1}{N}\sum_{n=1}^{N}\left\|\boldsymbol{e}_n\right\|^2$$

$$= \frac{1}{N}\sum_{j=l+1}^{m}\boldsymbol{w}_j^T\left(\sum_{n=1}^{N}\boldsymbol{x}\boldsymbol{x}^T\right)\boldsymbol{w}_j$$

$$= \sum_{j=l+1}^{m}\boldsymbol{w}_j^T\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}\boldsymbol{x}^T\right)\boldsymbol{w}_j$$

$$= \sum_{j=l+1}^{m}\boldsymbol{w}_j^T\boldsymbol{S}\boldsymbol{w}_j$$

$$= \sum_{j=l+1}^{m}\sigma_j^2 = \sum_{j=l+1}^{m}\lambda_j$$

因此样本集的方差、重构后的方差，以及重构误差表示为：

$$\sum_{j=1}^{m}\sigma_j^2 = \sum_{j=1}^{m}\lambda_j$$

$$\sum_{j=1}^{l}\sigma_j^2 = \sum_{j=1}^{l}\lambda_j$$

$$\sum_{j=l+1}^{m}\sigma_j^2 = \sum_{j=l+1}^{m}\lambda_j$$