1、降维

这里所学的第二个无监督学习算法是降维，降维有很多显然的优点，第一是减小占用空间，计算更快；第二是方便表示数据；第三，去除潜在的冗余数据。

2、PCA 主成分分析

主成分分析又称主分量分析，主成分回归分析，旨在利用降维的思想，把多指标转化为少数几个综合指标，同时保持数据集中对方差贡献最大的特征。在统计学中，它是一种简化数据集的技术。

3、计算步骤

A、将样本按列组成矩阵 X：n*m，m 为样本个数，n 为特征数量。

B、先进行均值归一化。

C、求出协方差矩阵 $C=1/m \cdot X \cdot X^T$

D、求 C 的特征值和特征向量

E、将特征向量按特征值对应的大小进行从上往下排序，取前 k 行组成的矩阵 P

F、Y = PX 即降维到 j 维后的数据集。

举个例子，原始数据为 2 维，有 5 个数据。组成的矩阵为：

$$\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

进行均值归一化，然后求协方差 C 矩阵：

$$C = \frac{1}{5}\begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 & -2 \\ -1 & 0 \\ 0 & 0 \\ 2 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{6}{5} & \frac{4}{5} \\ \frac{4}{5} & \frac{6}{5} \end{pmatrix}$$

求 C 的特征值和特征向量：

$$\lambda_1 = 2, \lambda_2 = 2/5$$

$$c_1\begin{pmatrix} 1 \\ 1 \end{pmatrix}, c_2\begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

化标准特征向量为：

$$\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}, \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$$

根据特征值进行排序，得：

$$\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$$

降到 1 维，取第一行，此即为原 2 维数据降维到 1 维后的数据集。

$$\begin{aligned} Y &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix} \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} -3/\sqrt{2} & -1/\sqrt{2} & 0 & 3/\sqrt{2} & -1/\sqrt{2} \end{pmatrix} \end{aligned}$$