Part I

PCA

1 线性降维基本原理

Proposition 1.1. 给定一个低维的线性空间U，一般$\mathbb{R}^2$用于可视化，线性降维本质上都是寻找一个从特征空间V到U的线性映射$\varphi : V \to U$. 在这个线性映射的作用下，高维空间中的向量可以用坐标表示，从而在平面上展开，并且保留数据的某些特征.

Theorem 1.1 (PCA主定理)**.** 对于$n\times 1$的k个列向量并成的$n\times k$矩阵Y，每个坐标轴上的方差和坐标轴之间的协方差用协方差矩阵D表示. 若列向量矩阵Y是中心化的，即每一个坐标轴上的分量总和为0，将$n\times k$矩阵Y写成行向量形式：

$$Y = \begin{pmatrix} \vec{\alpha_1} & \vec{\alpha_2} & \cdots & \vec{\alpha_k} \end{pmatrix}_{n\times k} = \begin{pmatrix} \vec{\gamma_1} \\ \vec{\gamma_2} \\ \vdots \\ \vec{\gamma_n} \end{pmatrix}_{n\times k}$$

其中每一个$\vec{\gamma_i}$的所有分量之和为0（中心化条件），两个行向量之间的协方差满足

$$cov(\gamma_i,\gamma_j) = \begin{cases} var(\gamma_i) = \dfrac{1}{k}\displaystyle\sum_{t=1}^{k}(a_{it}-0)^2 = \dfrac{1}{k}\gamma_i{\gamma_i}^t, & i=j \\ cov(\gamma_i,\gamma_j) = \dfrac{1}{k}\displaystyle\sum_{t=1}^{k}(a_{it}-0)(a_{jt}-0) = \dfrac{1}{k}\gamma_i{\gamma_j}^t, & i\neq j \end{cases}$$

协方差矩阵D可用矩阵乘法表示：

$$D = \frac{1}{k}YY^t = \frac{1}{k}\begin{pmatrix} \vec{\gamma_1} \\ \vec{\gamma_2} \\ \vdots \\ \vec{\gamma_n} \end{pmatrix} \begin{pmatrix} \vec{\gamma_1}^t & \vec{\gamma_2}^t & \cdots & \vec{\gamma_n}^t \end{pmatrix}$$

$$= \begin{pmatrix} cov(\gamma_1,\gamma_1) & cov(\gamma_1,\gamma_2) & \cdots & cov(\gamma_1,\gamma_n) \\ cov(\gamma_2,\gamma_1) & cov(\gamma_2,\gamma_2) & \cdots & cov(\gamma_2,\gamma_n) \\ \vdots & \vdots & & \vdots \\ cov(\gamma_n,\gamma_1) & cov(\gamma_n,\gamma_2) & \cdots & cov(\gamma_n,\gamma_n) \end{pmatrix}$$

综上，当降维后的协方差矩阵为对角矩阵，且对角线元素从大到小排列的时候，就实现了我们的目标：在每个坐标轴上，样本尽可能分散，在并且不同的坐标轴之间是线性无关的（正交的）,即：

$$cov(\gamma_i, \gamma_j) = \begin{cases} var(\gamma_i)\ \max, & i = j \\ cov(\gamma_i, \gamma_j) = 0, & i \neq j \end{cases}$$

根据我们的线性降维手段，$Y = AX$，带入上述方程:

$$D = \frac{1}{k}YY^t = \frac{1}{k}(AX)(AX)^t = \frac{1}{k}AXX^tA^t = ACA^t$$

其中$C = \frac{1}{k}XX^t$是X的协方差矩阵，D是对角阵. 由于实对角阵总可以相似合同对角化，所以这里的A取正交矩阵，满足$A^t = A^{-1}$ 若C有相似对角化:

$$\begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} = A_{n\times n}CA^t$$

这里ACD都为同形方阵，取前m个最大的特征值$\{\lambda_1 \ldots \lambda_m\}$构成的对角阵$D$，从而得到取前$m$个坐标的截短的特征向量矩阵$A_{m\times n}$，使得：

$$D = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_m \end{pmatrix} = A_{m\times n}C_{n\times n}A^t$$

从而，对于有k个样本点的原始数据矩阵X，我们找到了一个A，使得结果$Y = AX$满足$\frac{1}{k}YY^t$是对角线元素递减的对角阵。 ∎