PCA提要

基本思路

假设有伸缩程度不一样的两维（由图中 $\mathbf{u_1}$ 和 $\mathbf{u_2}$ 决定）：

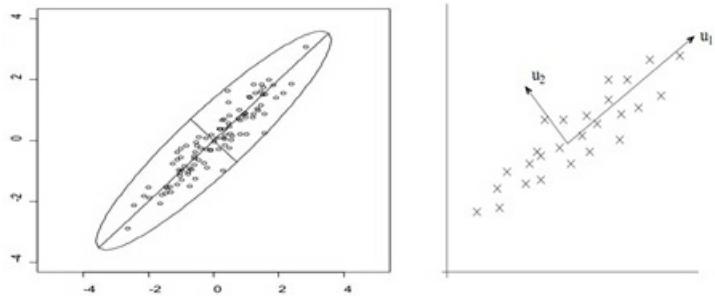

图 3

那么在降维的过程中选择将保留数据方差大（$\mathbf{u_1}$）的那个维度。

方法

假设一组中心化（减掉均值之后）后的数据：$\{\mathbf{x_1}, \mathbf{x_2}, \dots \mathbf{x_N}\}$，其中$\mathbf{x_i}$是D维向量；要找到方差最大的维度，由$\mathbf{u_1}$表示.

即求：$argmax_{u_1} \sum (x_i^T u_1)^2$

令 $f(u_1) = \sum (x_i^T u_1)^2$ 并作如下变形：

$$f(u_1) = \sum (x_i^T u_1)^2 = \sum (x_i^T u_1)^T (x_i^T u_1) = \sum u_1^T x_i x_i^T u_1 = u_1^T \sum x_i x_i^T u_1 = u_1^T X X^T u_1$$

其中，$X = [x_1, x_2, \dots, x_N]$

此外，u1需满足约束条件：$u_1^T u_1 = 1$（标准基）

拉格朗日乘子法：

$$f_2(u_1) = u_1^T X X^T u_1 + \lambda(1 - u_1^T u_1)$$

$$\frac{\partial f_2}{\partial u_1} = 2XX^T u_1 - 2\lambda u_1 = 0$$

所以$XX^T u_1 = \lambda u_1$，即：满足最大值条件的u_1一定是XX^T的特征向量。并且：$f(u1) = u_1^T X X^T u_1 = u_1^T \lambda u_1 = \lambda$

因此：最大的特征值对应的特征向量即是方差最大的维度

所以，求得XX^T的特征值和特征向量，按特征值大小排序并取前K个对应的特征向量按行排列成矩阵P，那么PX即为降维后的数据