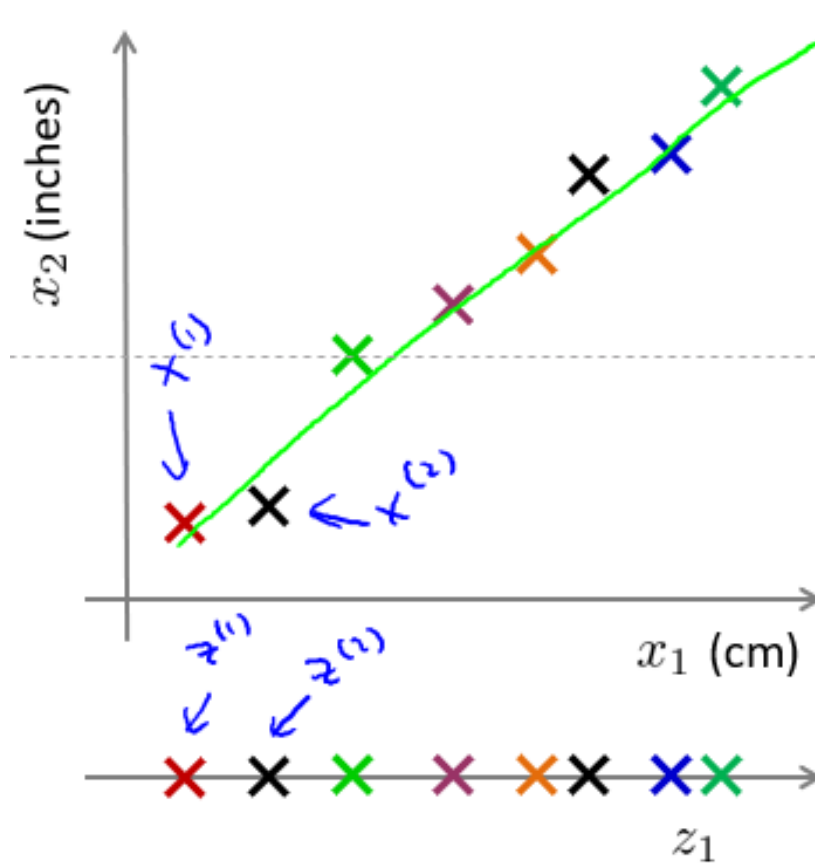


主成分分析

PCA(Principal Component Analysis)

数据压缩2D-1D



Reduce data from
2D to 1D

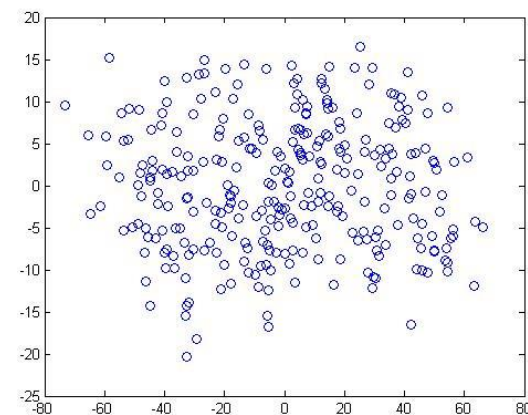
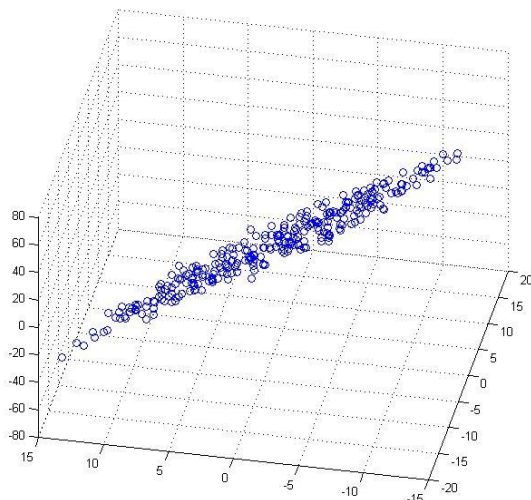
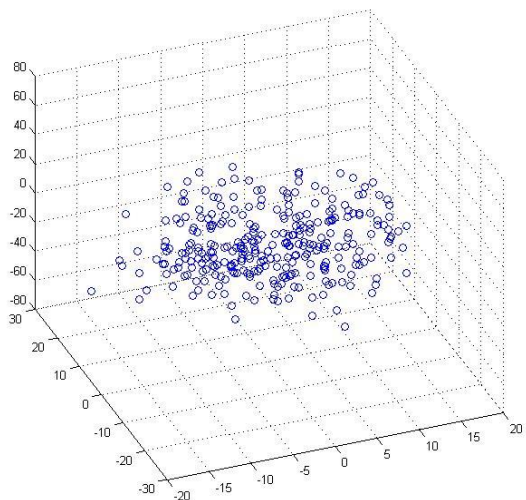
$$x^{(1)} \in \mathbb{R}^2 \rightarrow z^{(1)} \in \mathbb{R}$$

$$x^{(2)} \in \mathbb{R}^2 \rightarrow z^{(2)} \in \mathbb{R}$$

⋮

$$x^{(m)} \in \mathbb{R}^2 \rightarrow z^{(m)} \in \mathbb{R}$$

数据压缩3D-2D

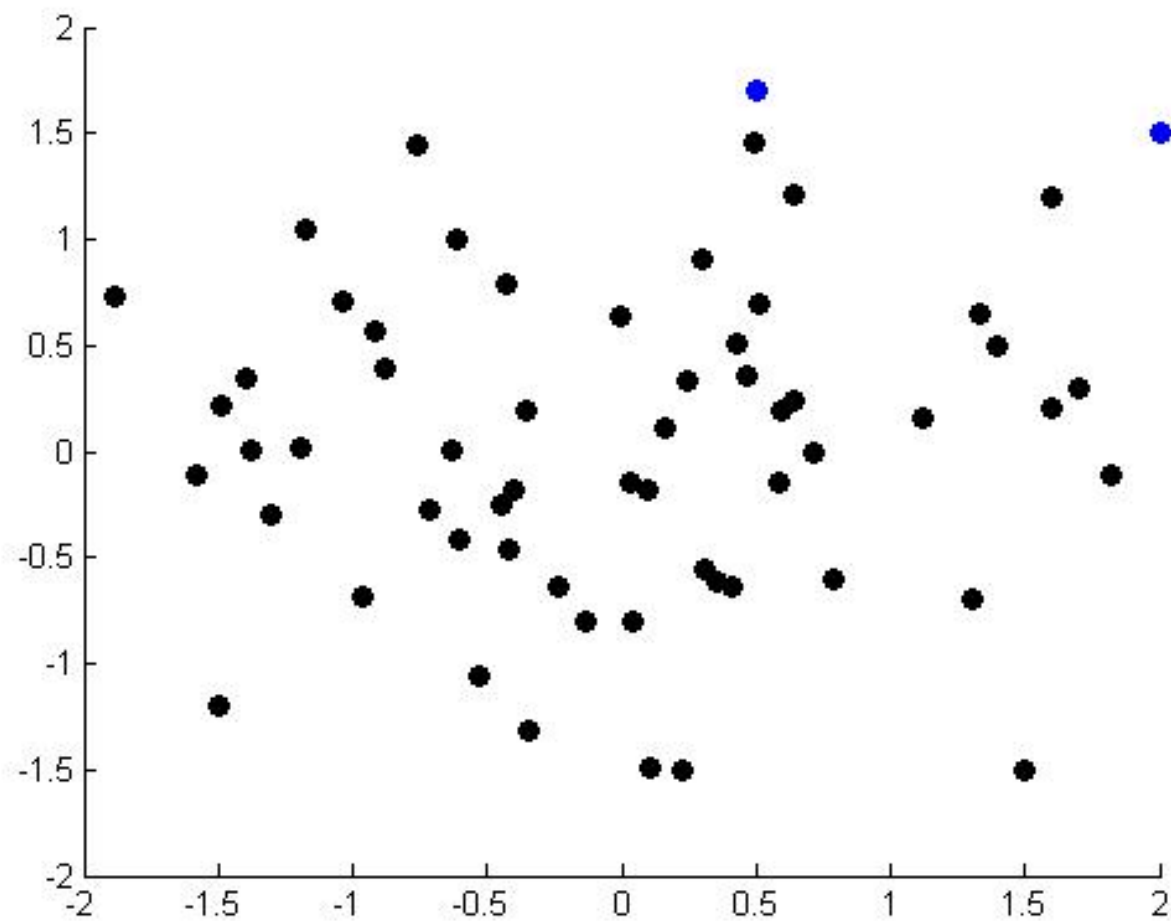




Country	GDP (trillions of US\$)	Per capita GDP (thousands of intl. \$)	Human Develop- ment Index	Life expectancy	Poverty Index (Gini as percentage)	Mean household income (thousands of US\$)	...
Canada	1.577	39.17	0.908	80.7	32.6	67.293	...
China	5.878	7.54	0.687	73	46.9	10.22	...
India	1.632	3.41	0.547	64.7	36.8	0.735	...
Russia	1.48	19.84	0.755	65.5	39.9	0.72	...
Singapore	0.223	56.69	0.866	80	42.5	67.1	...
USA	14.527	46.86	0.91	78.3	40.8	84.3	...
...

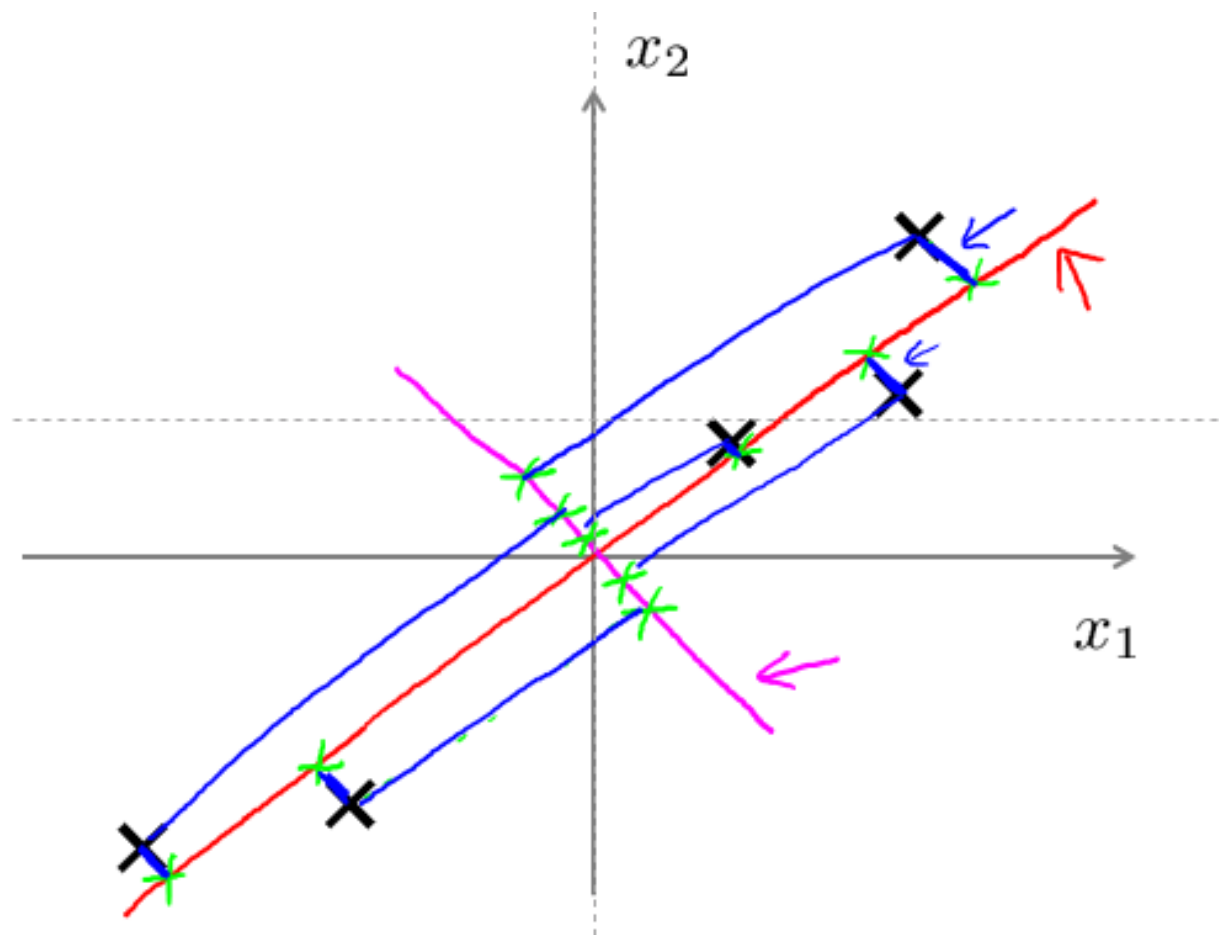


Country		
Canada	1.6	1.2
China	1.7	0.3
India	1.6	0.2
Russia	1.4	0.5
Singapore	0.5	1.7
USA	2	1.5
...





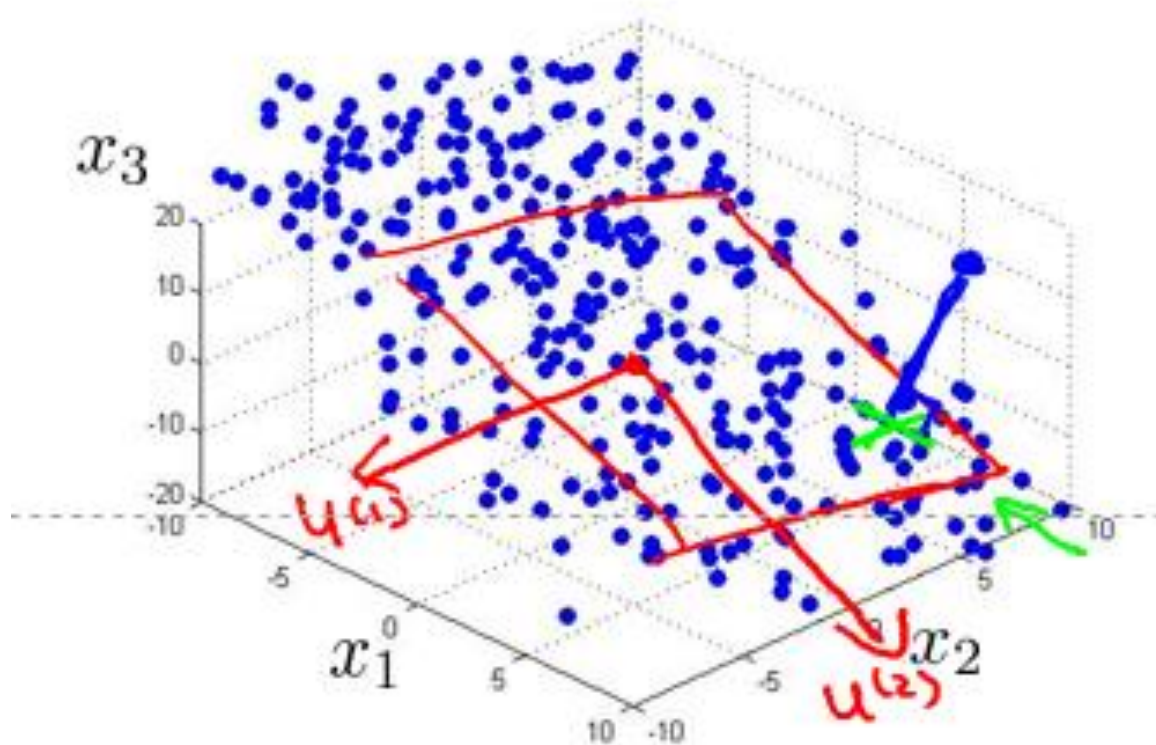
找到数据最重要的方向(方差最大的方向)



降维分析



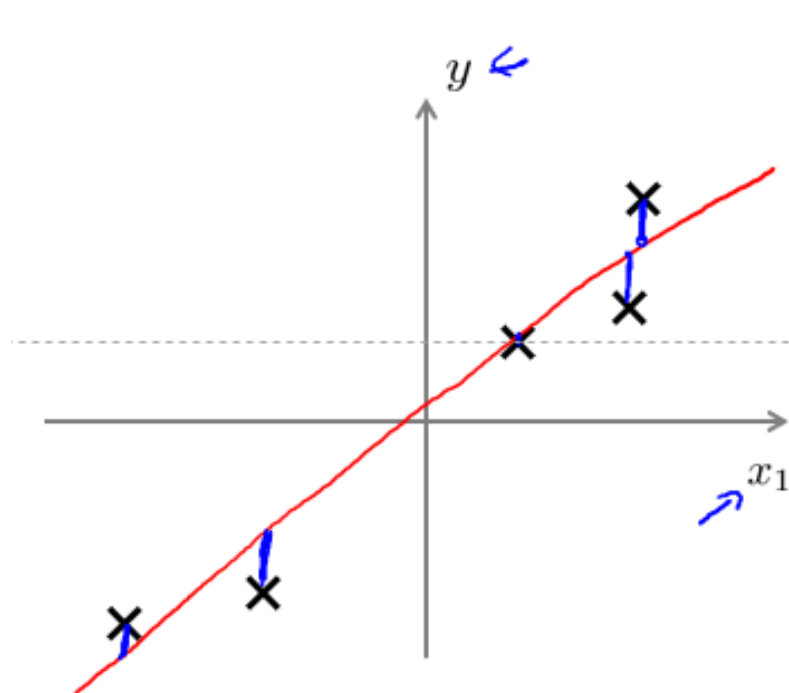
第一个主成分就是从数据差异性最大(方差最大)的方向提取出来的，第二个主成分则来自于数据差异性次大的方向，并且要与第一个主成分方向正交。



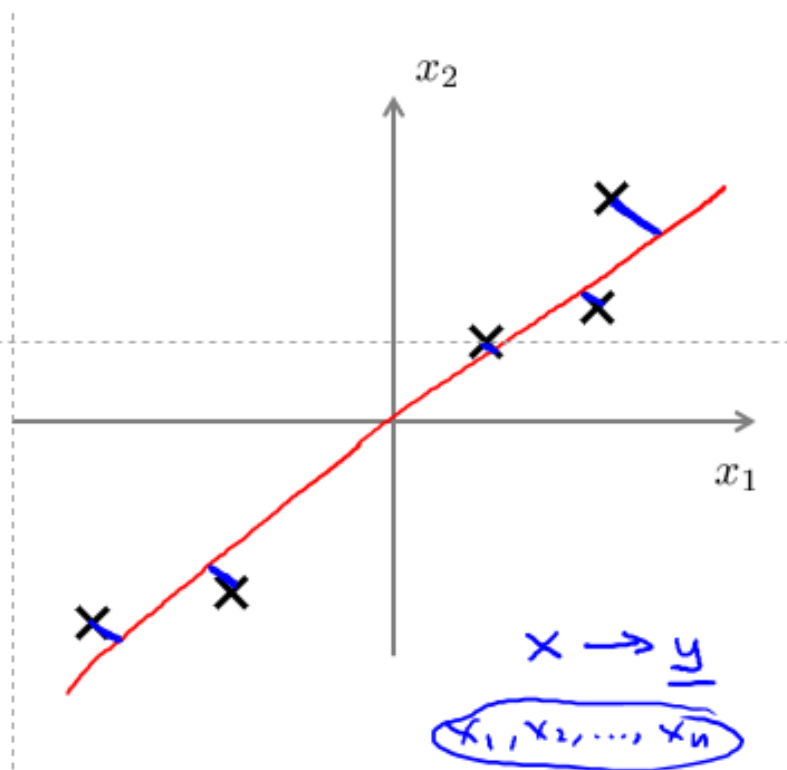
PCA不是线性回归



线性回归



PCA





- 1.数据预处理：中心化 $X - \bar{X}$ 。
- 2.求样本的协方差矩阵 $\frac{1}{m}XX^T$ 。
- 3.对协方差 $\frac{1}{m}XX^T$ 矩阵做特征值分解。
- 4.选出最大的k个特征值对应的k个特征向量。
- 5.将原始数据投影到选取的特征向量上。
- 6.输出投影后的数据集。



方差描述一个数据的离散程度：

$$\text{var}(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{n - 1}$$

协方差描述两个数据的相关性，接近1就是正相关，接近-1就是负相关，接近0就是不相关。

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$



协方差只能处理二维问题，那维数多了自然需要计算多个协方差，我们可以使用矩阵来组织这些数据。
协方差矩阵是一个对称的矩阵，而且对角线是各个维度的方差。

二维的例子：

$$C = \begin{pmatrix} cov(x, x) & cov(x, y) \\ cov(y, x) & cov(y, y) \end{pmatrix} = \begin{pmatrix} \frac{1}{m} \sum_i x_i^2 & \frac{1}{m} \sum_i x_i y_i \\ \frac{1}{m} \sum_i y_i x_i & \frac{1}{m} \sum_i y_i^2 \end{pmatrix}$$

三维的例子：

$$C = \begin{pmatrix} cov(x, x) & cov(x, y) & cov(x, z) \\ cov(y, x) & cov(y, y) & cov(y, z) \\ cov(z, x) & cov(z, y) & cov(z, z) \end{pmatrix}$$



n个特征，m个样本。n行m列

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & \dots & x_1^{(m)} \\ \vdots & \dots & \vdots \\ x_n^{(1)} & \dots & x_n^{(m)} \end{bmatrix}$$

n行m列乘m行n列->n行n列

$$\mathbf{X}\mathbf{X}^T = \mathbf{\Sigma} = \begin{bmatrix} \Sigma_{11} & \dots & \Sigma_{1n} \\ \vdots & \dots & \vdots \\ \Sigma_{n1} & \dots & \Sigma_{nn} \end{bmatrix}$$

$$\mathbf{C} = \begin{pmatrix} cov(x, x) & cov(x, y) \\ cov(y, x) & cov(y, y) \end{pmatrix} = \begin{pmatrix} \frac{1}{m} \sum_i x_i^2 & \frac{1}{m} \sum_i x_i y_i \\ \frac{1}{m} \sum_i y_i x_i & \frac{1}{m} \sum_i y_i^2 \end{pmatrix}$$



通过数据集的协方差矩阵及其特征值分析，我们可以得到协方差矩阵的特征向量和特征值。我们需要保留 k 个维度的特征就选取最大的 k 个特征值。

PCA-简单例子



手写数字识别降维可视化



sklearn-手写数字降维预测





「关注 AI MOOC 官方公众号」

这里持续分享
Python、机器学习和深度学习的
各种资源干货及
有趣又实用的硬知识



覃秉丰

AI MOOC创始人

网易计算机视觉微专业核心讲师

机器学习、深度学习多年开发经验

受邀为中国移动、国家电网、华夏银行、
太平洋保险等世界五百强企业总部做AI内训

