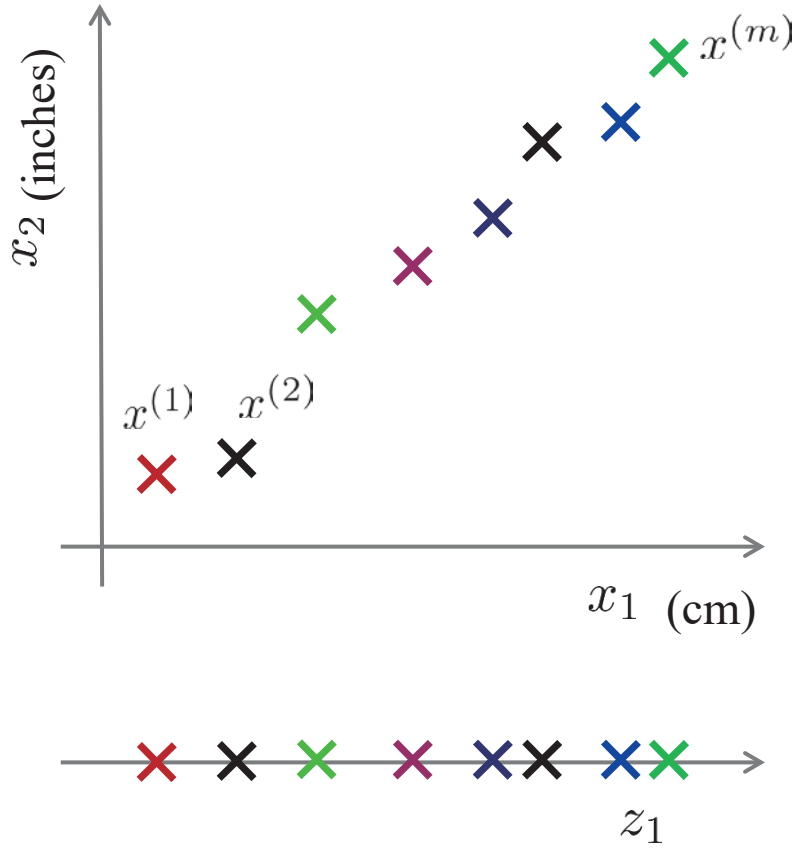
降维问题

降维的作用

数据压缩（提取主要信息）

2D to 1D数据压缩

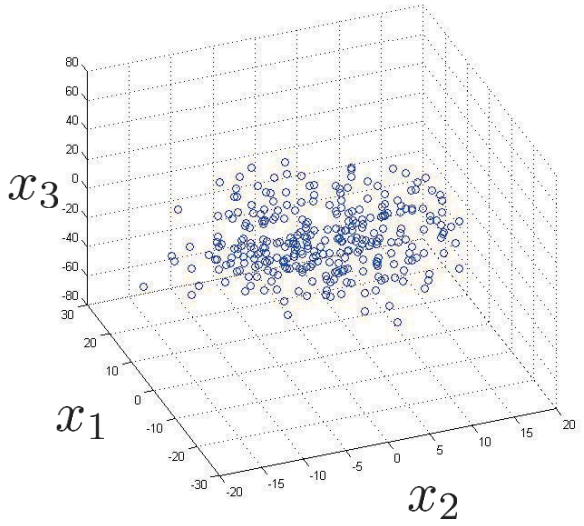


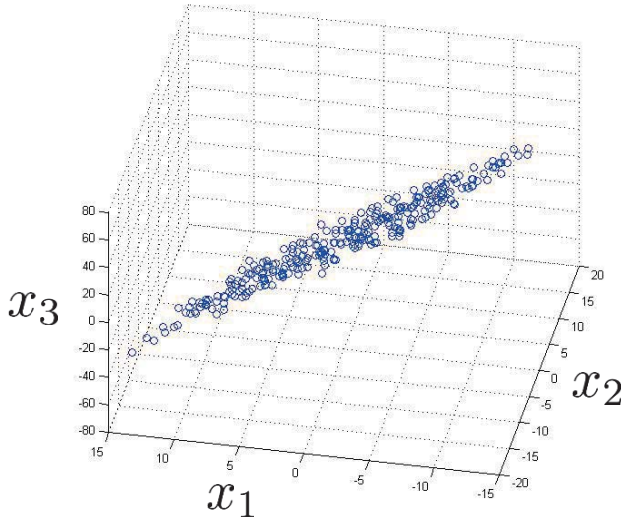
$$x^{(1)} \rightarrow z^{(1)}$$
$$x^{(2)} \rightarrow z^{(2)}$$
$$\vdots$$
$$x^{(m)} \rightarrow z^{(m)}$$

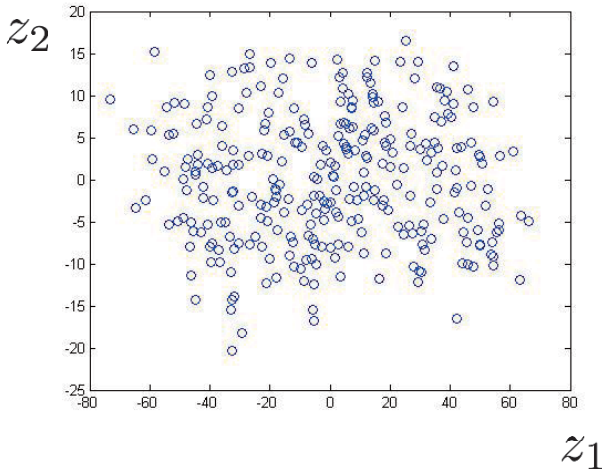
降维的作用

数据压缩（提取主要信息）

3D to 2D数据压缩







降维的作用

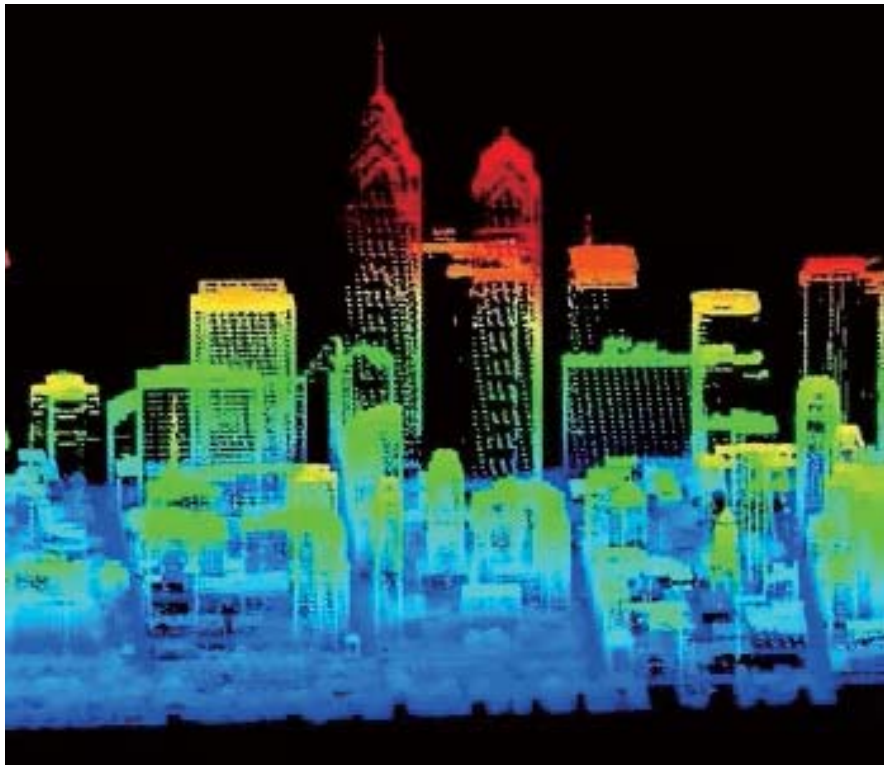
数据压缩（提取主要信息）

降维——将数据的维度降低，便于发现问题，提高数据的处理效率。

降维的作用

数据可视化

点云数据		
22.780	22.920	23.120
58.880	58.660	58.680
37.860	37.510	38.150
39.820	39.910	39.950
46.300	46.390	46.400
39.850	39.990	39.830
42.700	42.640	42.690
44.070	44.210	44.420
49.730	49.470	49.670
50.520	51.140	50.930
90.240	90.350	90.600
90.970	91.000	91.100
69.410	69.450	69.410
70.960	71.110	71.090
83.120	83.040	82.870
54.230	54.310	54.440
23.730	23.710	23.690
43.320	43.160	42.910
33.260	32.890	33.140
17.940	17.990	18.340
20.570	20.640	20.540
11.700	11.590	11.780



降维的作用

数据可视化（合并特征）

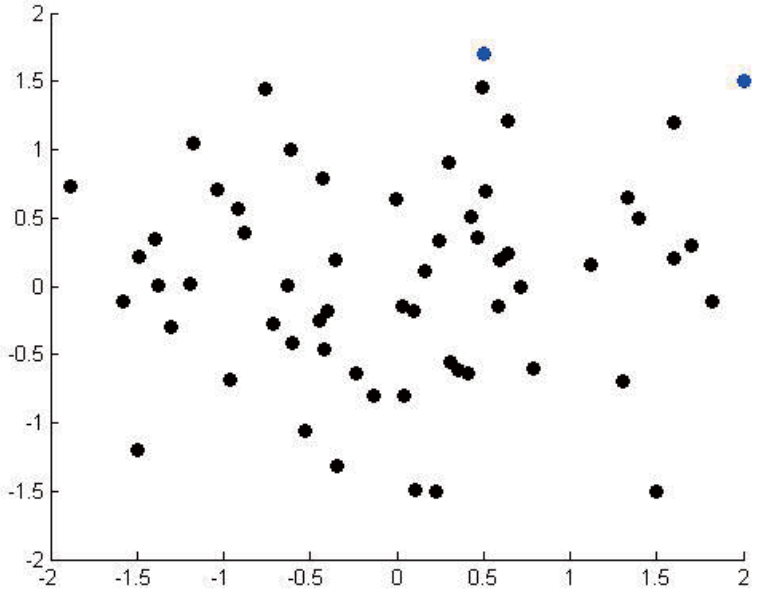
Country	GDP (trillions of US$)	Per capita GDP (thousands of intl. $)	Human Develop-ment Index	Life expectancy	Poverty Index (Gini as percentage)	Mean household income (thousands of US$)	…
Canada	1.577	39.17	0.908	80.7	32.6	67.293	…
China	5.878	7.54	0.687	73	46.9	10.22	…
India	1.632	3.41	0.547	64.7	36.8	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	…
…	…	…	…	…	…	…	

不能数据可视化

降维的作用

数据可视化（合并特征）

Country	z_1	z_2
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
…	…	…



降维后新特征的实际意义是什么？

降维问题

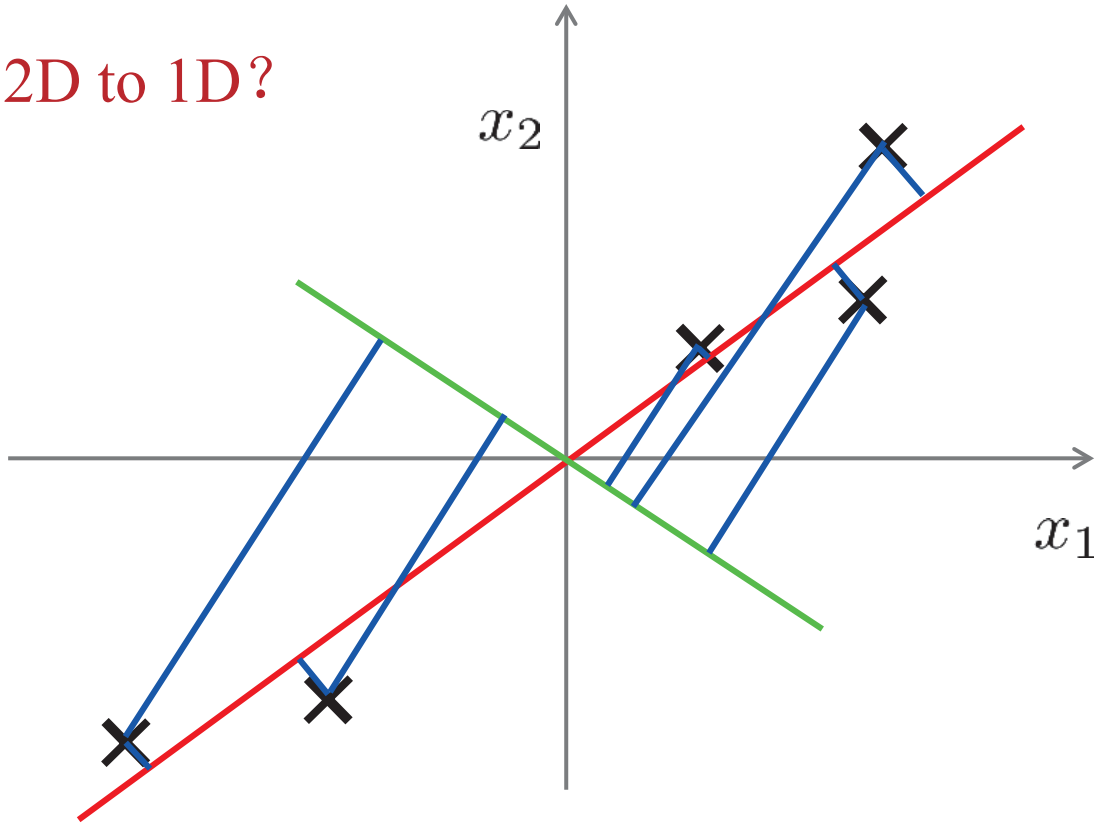
主成分分析方法

主成分分析方法

主成分分析：PCA（Principal Components Analysis）

PCA：找到一个方向向量（Vector direction），所有数据投影到该变量上，投影平均均方误差尽可能小。

2D to 1D?



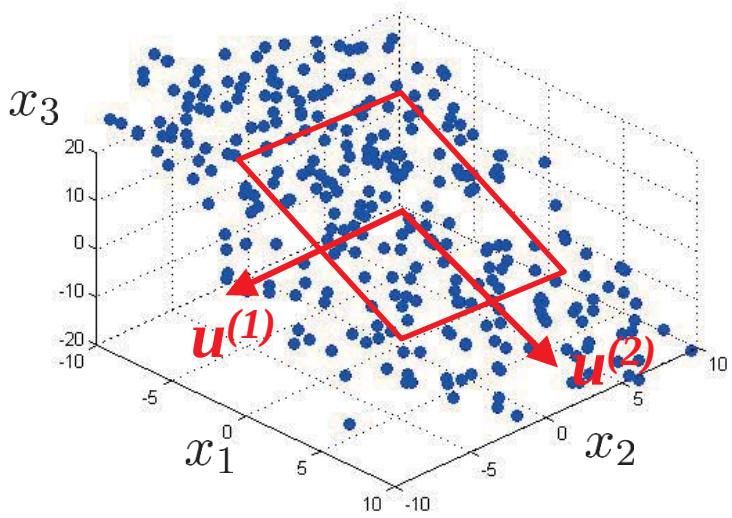
主成分分析方法

主成分分析问题的描述：

2D to 1D：找到方向（向量）在其上投影，最小化投影距离。

$u^{(1)} \in \mathbb{R}$

从n维减少到k维：找到向量（ $u^{(1)}, u^{(2)}, \ldots, u^{(k)}$），在其上投影数据具有最小化的投影误差。



主成分分析方法

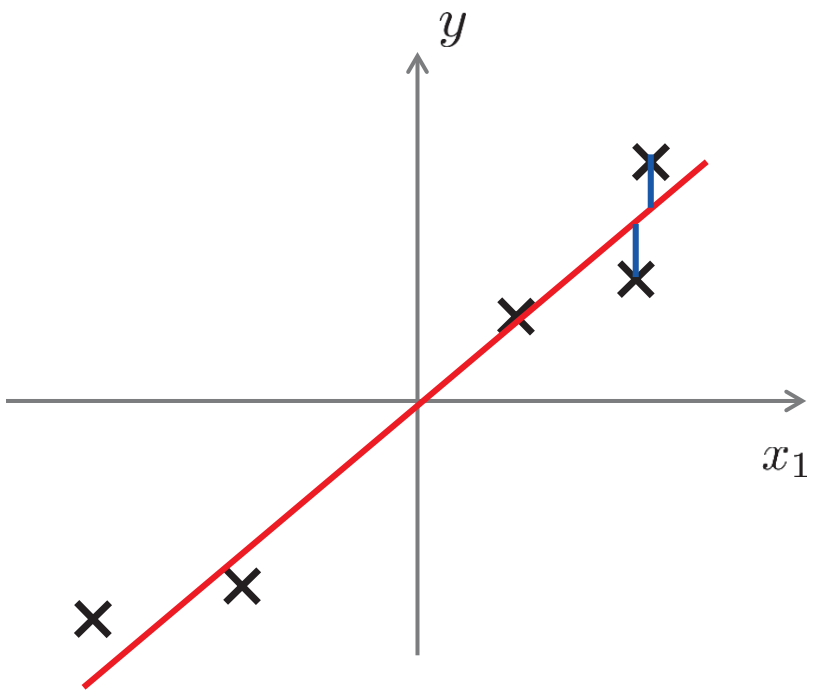
PCA的思想：

- 将n维特征映射到k维上（k<n），k维特征称为主元（主成分）。

k维特征是重新构造出来，不是简单地从n维特征中去掉了其余n-k维特征。

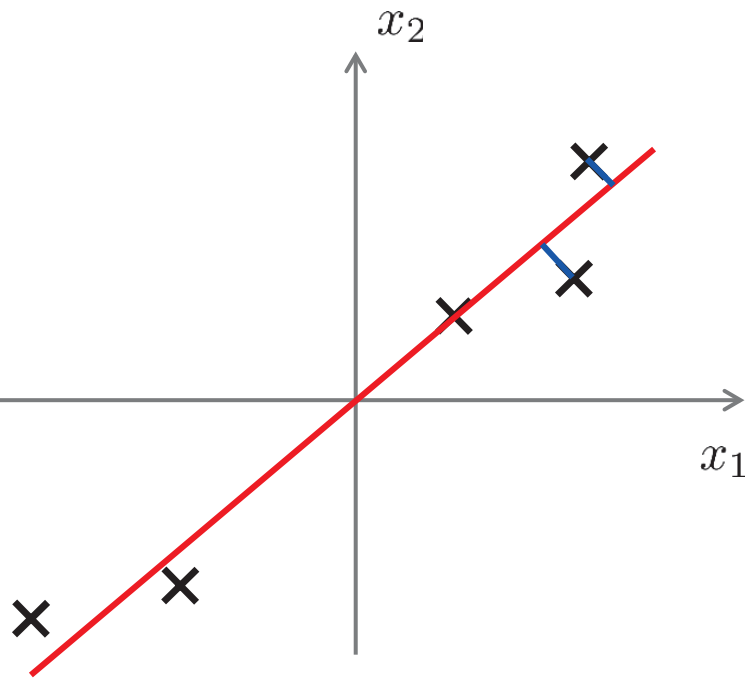
PCA与线性回归

线性回归



最小化预测误差

主成分分析方法



最小化投射误差

主成分分析方法

算法步骤

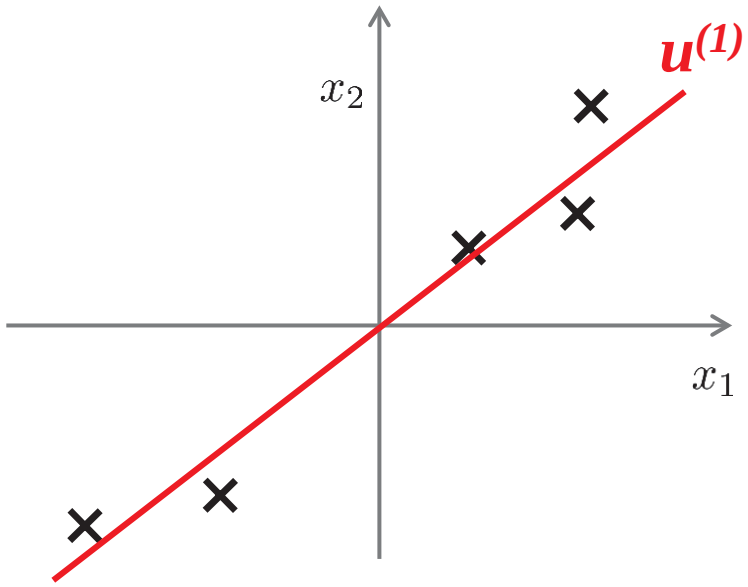
假设有数据集：$x^{(1)}, x^{(2)}, \ldots, x^{(m)}$

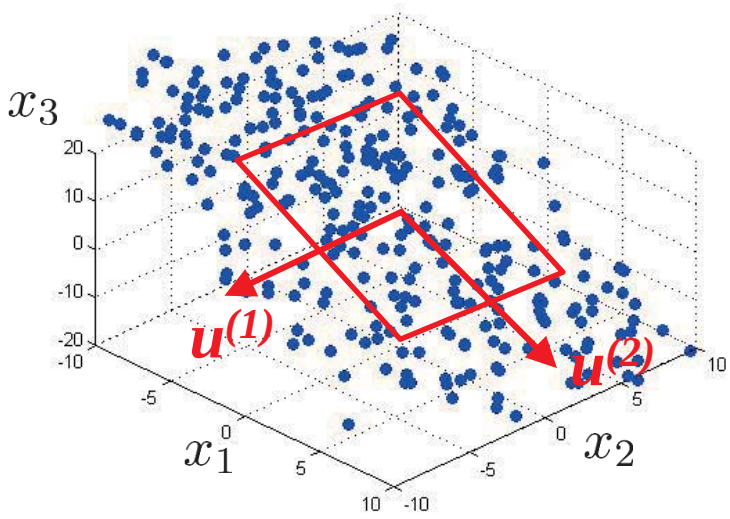
要采用PCA方法，首先要进行数据的预处理：

均值归一化

- 计算出所有特征的均值 $\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$
- 令 $x_j^{(i)} = x_j - \mu_j$。
- 如果特征是在不同的数量级上，还需要将其除以标准差σ，也就是 $x_j^{(i)} = （x_j - \mu_j）/ \sigma$

主成分分析方法





如何得到这些向量？
如何得到数据在新特征下的数值？

主成分分析方法

算法步骤

PCA减少 n 维到 k 维的步骤如下：

首先，计算协方差矩阵

$$\Sigma = \frac{1}{m}\sum_{i=1}^{n}(x^{(i)})(x^{(i)})^T$$

$x^{(i)}$是一个$nx1$的矩阵

然后，计算协方差矩阵的Σ的特征向量，在Matlab、Octave、

Python numpy 里可以利用奇异值分解来求解： [U, S, V]=svd(Σ)

$$U = \begin{bmatrix} | & | & & | \\ u^{(1)} & u^{(2)} & \cdots & u^{(n)} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n\times n}$$

主成分分析方法

算法步骤

$$U = \begin{bmatrix} | & | & & | \\ u^{(1)} & u^{(2)} & \cdots & u^{(n)} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times n}$$

U 是一个具有与数据之间最小投射误差的方向向量构成的矩阵。从 U 中 选取前 K 个向量，获得一个 n×k 维度的矩阵U_{reduce}

新特征向量$z^{(i)}$为：

$$z^{(i)} = U_{reduce}^{T} X x^{(i)};$$

数据“恢复”

通过PCA算法，可把1000维的数据压缩100维特征，或将三维数据压缩到一二维表示。

在压缩过数据后，可以采用如下方法来近似地获得原有的特征：

$$x_{appox} = U_{reduce} \bullet z$$

降维问题

PCA的应用建议

PCA的应用建议

应用场景

- 数据压缩
  - 减小内存、硬盘存储空间需求
  - 提高计算效率

- 可视化

PCA的Python实现

作业

1. 用PCA算法将pca_data.txt中的数据压缩至2维