线性代数

机器学习中数学基础第三课

何老师 慧科AI讲师

- **标量、向量、矩阵和张量**
  - **标量（scalar）：** 一个标量就是一个单独的数，它不同于线性代数中研究的其他大部分对象（通常是多个数的数组）。
  - **向量（vector）：** 一个向量是一列数。这些数是有序排列的。通过次序中的索引，我们可以确定每个单独的数。

- **矩阵（matrix）：** 矩阵是具有相同特征和纬度的对象的集合，表现为一张二维数据表。如果一个实数矩阵高度为mm，宽度为nn，那么定义 $A \in R_{m \times n}$

- **张量（tensor）：** 在某些情况下，我们会讨论坐标超过两维的数组。一般地，一个数组中的元素分布在若干维坐标的规则网格中，我们将其称之为张量。张量A中坐标为(i,j,k)的元素记作$A_{i,j,k}$

  - tensorflow

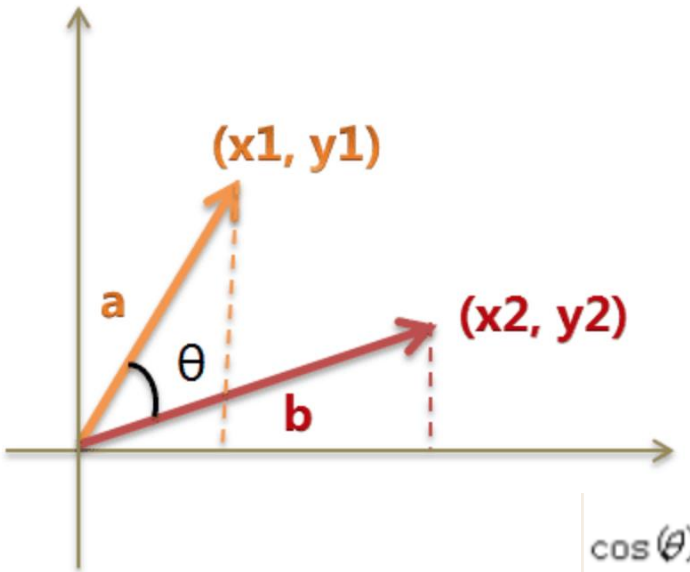
- 向量四则运算

  - 坐标表示　若$\boldsymbol{a}=(a_1, a_2, a_3)$，$\boldsymbol{b}=(b_1, b_2, b_3)$.

向量运算	坐标表示
$\boldsymbol{a}+\boldsymbol{b}$	$(a_1+b_1, a_2+b_2, a_3+b_3)$
$\boldsymbol{a}-\boldsymbol{b}$	$(a_1-b_1, a_2-b_2, a_3-b_3)$
$\lambda\boldsymbol{a}$	$(\lambda a_1, \lambda a_2, \lambda a_3)$
$\boldsymbol{a}\cdot\boldsymbol{b}$	$a_1b_1+a_2b_2+a_3b_3$

- 相似度

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{(x_1, y_1) \cdot (x_2, y_2)}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$



句子A：(1，1，2，1，1，1，0，0，0)
句子B：(1，1，1，0，1，1，1，1，1)

$$\cos(\theta) = \frac{1\times1+1\times1+2\times1+1\times0+1\times1+1\times1+0\times1+0\times1+0\times1}{\sqrt{1^2+1^2+2^2+1^2+1^2+1^2+0^2+0^2+0^2} \times \sqrt{1^2+1^2+1^2+0^2+1^2+1^2+1^2+1^2+1^2}}$$

$$= \frac{6}{\sqrt{7} \times \sqrt{8}}$$

$$= 0.81$$

基本概念

- 矩阵乘法

$A_{m\times s}B_{t\times n}$ 有意义的条件是 $s=t$

$A_{m\times s}B_{s\times n}=C_{m\times n}$

$$A=\begin{pmatrix}1 & 3 & 4\\ 5 & 7 & 2\end{pmatrix} \quad B=\begin{pmatrix}1 & 2\\ 2 & 5\end{pmatrix}$$

$$C=\begin{pmatrix}1 & 5 & 8\\ 5 & 7 & 2\end{pmatrix} \quad D=\begin{pmatrix}1 & 2 & 9\\ 2 & 5 & 2\\ 7 & 6 & 4\end{pmatrix}$$

基本概念

- 行向量 $\left(a_1\ a_2\ \ldots\ a_n\right)$
- 列向量 $\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}$
- 方阵 $\begin{pmatrix} a_{11}\ a_{12} \cdots\cdots a_{1n} \\ a_{21}\ a_{22} \cdots\cdots a_{2n} \\ \cdots\cdots \\ a_{n1}\ a_{n2} \cdots\cdots a_{nn} \end{pmatrix}$

基本概念

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots\ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots\ldots + a_{2n}x_n = b_2 \\ \quad\ldots\ldots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots\ldots + a_{mn}x_n = b_m \end{cases} \quad \text{——（1）}$$

若记：$A = \begin{pmatrix} a_{11}\ a_{12} \ldots\ldots a_{1n} \\ a_{21}\ a_{22} \ldots\ldots a_{2n} \\ \ldots\ldots \\ a_{m1}\ a_{m2} \ldots\ldots a_{mn} \end{pmatrix} \quad X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad B = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$

则方程组（1）可记为：$AX = B$

基本概念

$A+B=B+A;$

$A+B+C=A+(B+C);$

$A+(-A)=0;$

$A+0=A;$

$\lambda A=A\lambda;$

$(\lambda\mu)A=\lambda(\mu A);$

$(\lambda+\mu)A=\lambda A+\mu A;$

$\lambda(A+B)=\lambda A+\lambda B;$

1) $(A^{T})^{T}=A;$

2) $(A+B)^{T}=A^{T}+B^{T};$

3) $(\lambda A)^{T}=\lambda A^{T};$

4) $(AB)^{T}=B^{T}A^{T};$

1) $ABC=A(BC)$

2) $\lambda(AB)=(\lambda A)B=A(\lambda B)$　（其中λ是数）

3) $A(B+C)=AB+AC$

$(B+C)A=BA+CA$

1) $|A^{T}|=|A|$

2) $|\lambda A|=\lambda^{n}|A|$

3) $|AB|=|A||B|$　（当A,B均为方阵时）

基本概念

- 1. 基本概念

- 考虑一个方程组
  - 3x+4y=10
  - 5x-7y=3

$$\begin{bmatrix} 3 & 4 \\ 5 & -7 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 10 \\ 3 \end{bmatrix}$$

  - 简单表示为：
  - AX=b

基本概念

- 矩阵不都是数字的堆砌
  - 3x+4y=10
  - 5x-7y=3
  - 行向量-图解法

$$\begin{bmatrix}3\\5\end{bmatrix}x+\begin{bmatrix}4\\-7\end{bmatrix}y=\begin{bmatrix}10\\3\end{bmatrix}$$

  - 列向量 –线性组合

基本概念

- $\forall x, y$

$$\begin{bmatrix}3\\5\end{bmatrix}x+\begin{bmatrix}4\\-7\end{bmatrix}y=?\ ]$$

- 组成一个平面

- 向量空间的含义

基本概念

- 线性相关：
- 线性无关：

在向量空间V的一组向量**A:** $\alpha_1, \alpha_2, \cdots \alpha_m$，如果存在**不全为零**的数 $k_1, k_2, \cdots, k_m$，使

$$k_1\alpha_1 + k_2\alpha_2 + \cdots k_m\alpha_m = O$$

则称向量组**A**是线性相关的[1]，否则数 $k_1, k_2, \cdots$,km全为0时，称它是线性无关。

基本概念

- 考虑这个方程组：能解出来么？

- $x+y+2z=9$
- $2x+y+3z=13$
- $x+2y+3z=20$

基本概念

- 矩阵的秩：A的最大无关组中向量的个数为A的秩

- 向量空间的维度

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$$

基本概念

- 线性空间：

1.1. 零空间

n 维向量，是 $Ax = 0$ 的解，所以$N(A) \in \mathbb{R}^n, dimN(A) = n - r$，自由元所在的列即可组成零空间的一组基。

1.2. 列空间

列向量是m 维的，所以$C(A) \in \mathbb{R}^m, dimC(A) = r$，主元所在的列即可组成列空间的一组基。

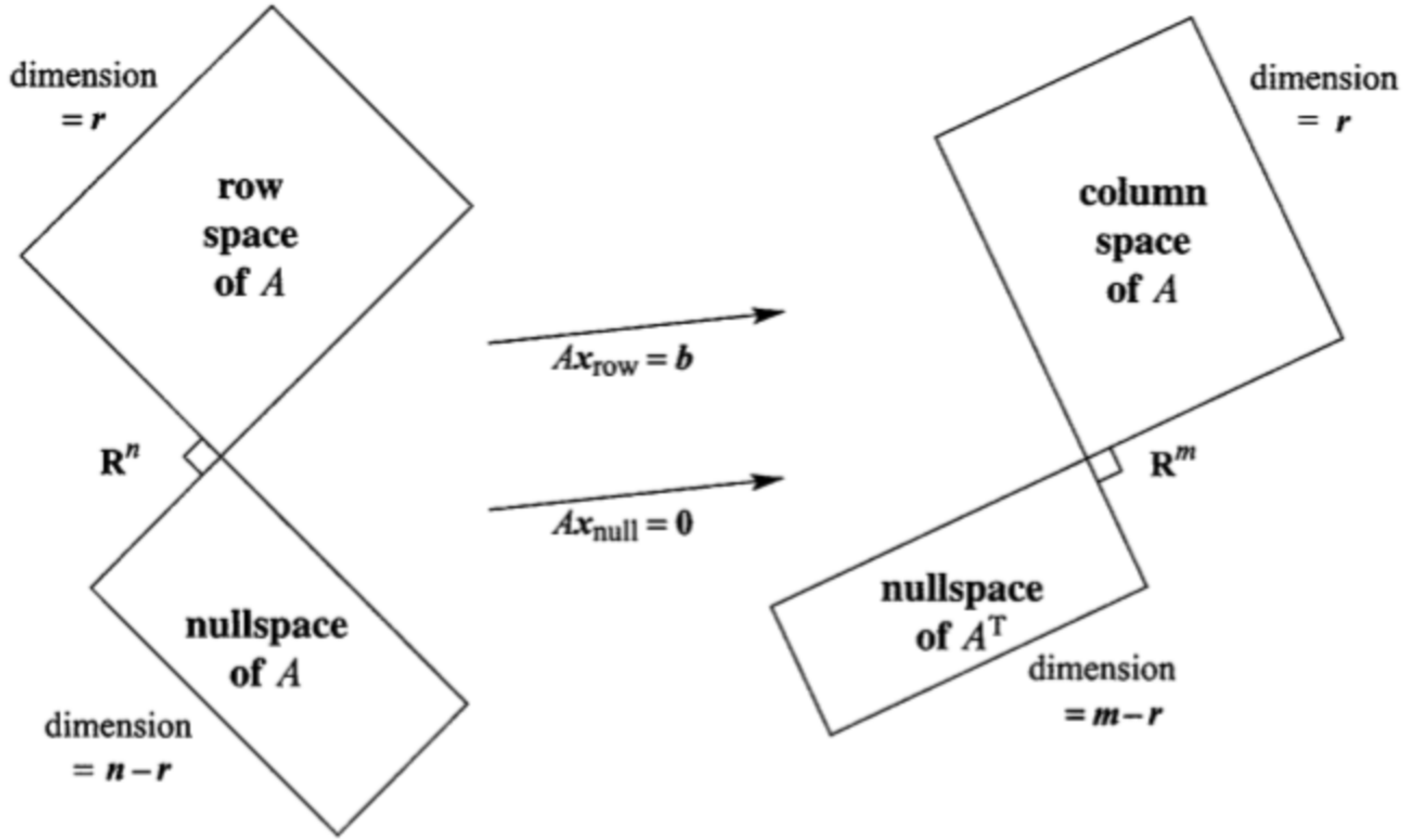
1.3. 行空间

A 的行的所有线性组合，即A 转置的列的线性组合（因为我们不习惯处理行向量），
$C(A^T) \in \mathbb{R}^n, dimC(A^T) = r$

1.4. 左零空间

A 转置的零空间 $N(A^T)$,A的左零空间
$N(A^T) \in \mathbb{R}^m, dimN(A^T) = m - r$

子空间的维度：


dimension
$= r$
row
space
of A
$Ax_{row} = b$
column
space
of A
dimension
$= r$
$\mathbf{R}^n$
$Ax_{null} = 0$
$\mathbf{R}^m$
nullspace
of A
dimension
$= n - r$
nullspace
of A^T
dimension
$= m - r$

基本概念

- 矩阵的秩：

- 向量空间的维度

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$$

基本概念

- 机器学习中经常是这样的数据：
- $x+y=4$
- $2x+3y=7$
- $4x+y=9$

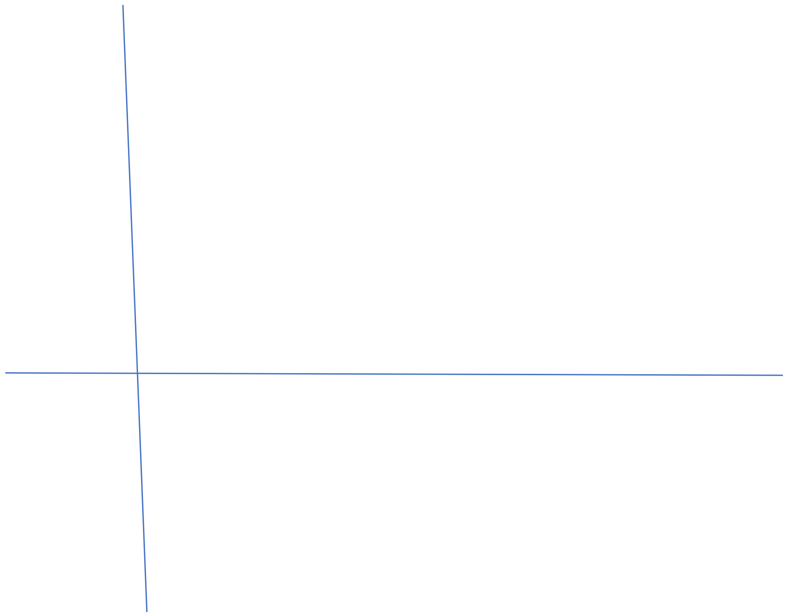
- 方程比未知数多
- $Ax=b$

基本概念

- 事实上就是拟合：

- 找最优的一个解，几何中就是找距离最近的点。

$$\begin{bmatrix}1\\2\\4\end{bmatrix}x+\begin{bmatrix}1\\3\\1\end{bmatrix}y=\begin{matrix}4\\7\\9\end{matrix}$$



- 超定方程集合意义

$$\begin{bmatrix}1\\2\\4\end{bmatrix}x + \begin{bmatrix}1\\3\\1\end{bmatrix}y = \begin{matrix}4\\7\\9\end{matrix}$$

$$G=\begin{bmatrix}1 & 1\\2 & 3\\4 & 1\end{bmatrix} \quad X = \begin{bmatrix}x\\y\end{bmatrix} \quad b = \begin{bmatrix}4\\7\\9\end{bmatrix}$$

- GX=b

$$G=\begin{bmatrix}1 & 1\\ 2 & 3\\ 4 & 1\end{bmatrix} \qquad X=\begin{bmatrix}x\\ y\end{bmatrix} \qquad b=\begin{bmatrix}4\\ 7\\ 9\end{bmatrix}$$

- x* y*:
- x*+y*=4
- 2x*+3y*=7
- 4x*+y*=9

数学再进一步

- 1.最小二乘法
  - 什么是最小二乘?
  - 高斯使用的最小二乘法的方法发表于1809年他的著作《天体运动论》中。法国科学家勒让德于1806年独立发明“最小二乘法”，但因不为世人所知而默默无闻。勒让德曾与高斯为谁最早创立最小二乘法原理发生争执。1829年，高斯提供了最小二乘法的优化效果强于其他方法的证明，因此被称为高斯-马尔可夫定理。（来自于wikipedia）

数学再进一步

- 小的实例

假定x, y有如下数值：

y	1.00	0.90	0.90	0.81	0.60	0.56	0.35
x	3.60	3.70	3.80	3.90	4.00	4.10	4.20

解：将这些数值画图可以看出接近一条直线，故用$y = ax + b$表示，故将上面的数值代入表达式有：

$$
\begin{aligned}
3.6a + b - 1.00 = 0 \\
3.7a + b - 0.90 = 0 \\
3.8a + b - 0.90 = 0 \\
3.9a + b - 0.81 = 0 \\
4.0a + b - 0.60 = 0 \\
4.1a + b - 0.56 = 0 \\
4.2a + b - 0.35 = 0
\end{aligned}
$$

由于直线只有两个未知数a, b，理论上只需要两个方程就能求得，但是实际上是不可能的，因为所有点并没有真正的在同一条直线上，即不可能所有的数值都满足

$$ax + b - y = 0$$

，故只需找到一对儿a、b，使得误差平方和

$$\sum (ax_i + b - y_i)^2 = (ax_0 + b - y_0)^2 + (ax_1 + b - y_1)^2 + \ldots\ldots + (ax_n + b - y_n)^2$$

最小即可。

误差的平方即二乘方，故成为最小二乘法。

数学再进一步

- 小的实例

上面是求最值的问题，我们会想到导数和偏导数，这里在偏导数等于0的地方能取到极值，并且也是最值。
分别对a和b求偏导得到如下表达式：

$$\frac{\partial}{\partial a} = \sum_{i=1}^{n} 2x_i(ax_i + b - y_i)$$
$$\frac{\partial}{\partial b} = \sum_{i=1}^{n} 2(ax_i + b - y_i)$$

通过对二元一次方程组

$$\sum_{i=1}^{n} 2x_i(ax_i + b - y_i) = 0$$
$$\sum_{i=1}^{n} 2(ax_i + b - y_i) = 0$$

进行求解，可以得到如下解：

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$
$$b = \frac{\sum_{i=1}^{n} x_i^2 \sum y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

```python
import numpy as np
import matplotlib.pyplot as plt

def linear_regression(x, y):
    N = len(x)
    sumx = sum(x)
    sumy = sum(y)
    sumx2 = sum(x**2)
    sumxy = sum(x*y)
    A = np.mat([[N, sumx], [sumx, sumx2]])
    b = np.array([sumy, sumxy])
    return np.linalg.solve(A, b)


X = np.arange(0, 10, 0.1)
Z = [8 + 7 * x for x in X]
Y = [np.random.normal(z, 4) for z in Z]
plt.plot(X, Y, 'k.')
plt.show()
```

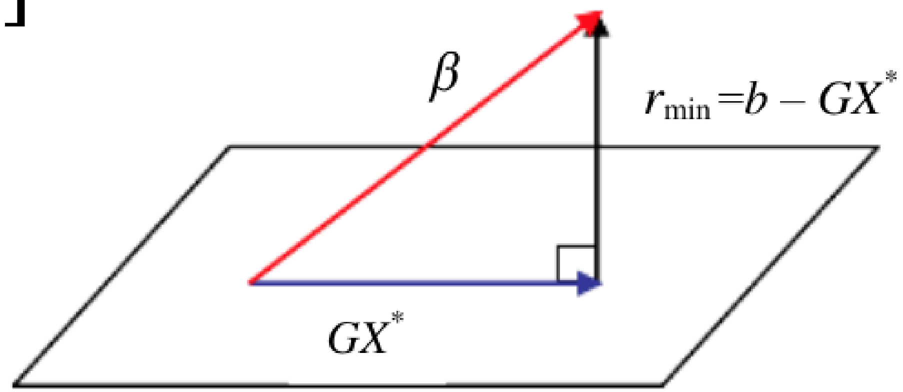
- 线性代数方法
- 1 X* 为平面上向量
- 2 法向量r
- 3 列向量正交
- 4 求出 $X=(G^TG)^{-1}G^Tb$

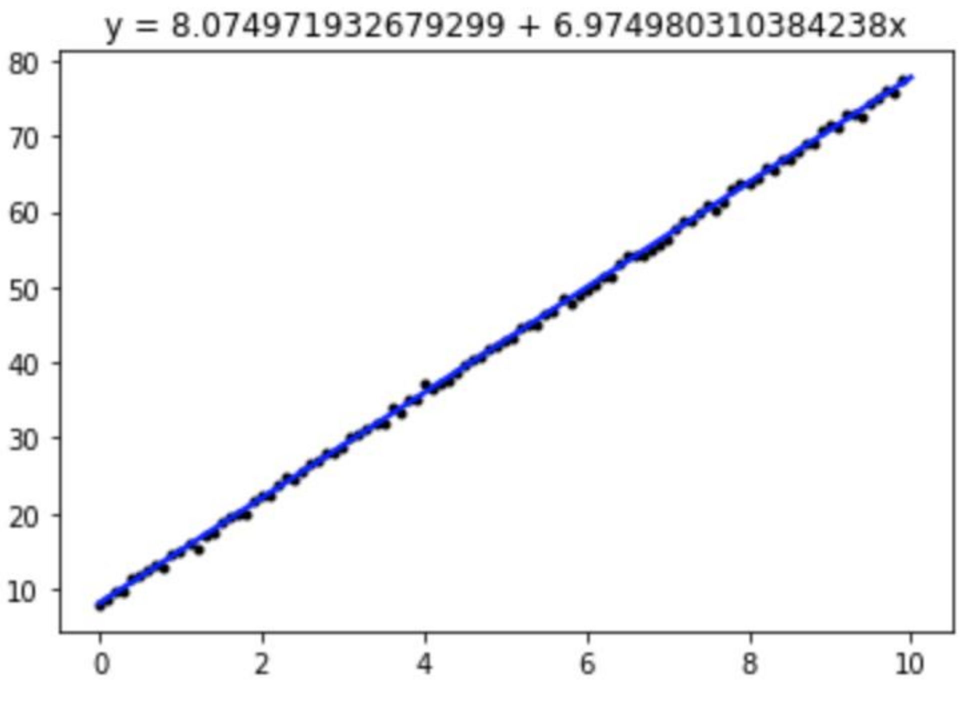
$$G=\begin{bmatrix}1 & 1\\ 2 & 3\\ 4 & 1\end{bmatrix}$$

$$X=\begin{bmatrix}x\\ y\end{bmatrix}$$

$$b=\begin{bmatrix}4\\ 7\\ 9\end{bmatrix}$$

$$G^T(b-GX^*)=0$$






```python
# 矩阵求法

bv = np.ones(len(X))

G=np.mat(np.array([X,bv]).T)

G.shape

b=np.matrix(Z).T

b.shape

(G.T*G).I*(G.T)*b
```

```
matrix([[7.],
        [8.]])
```

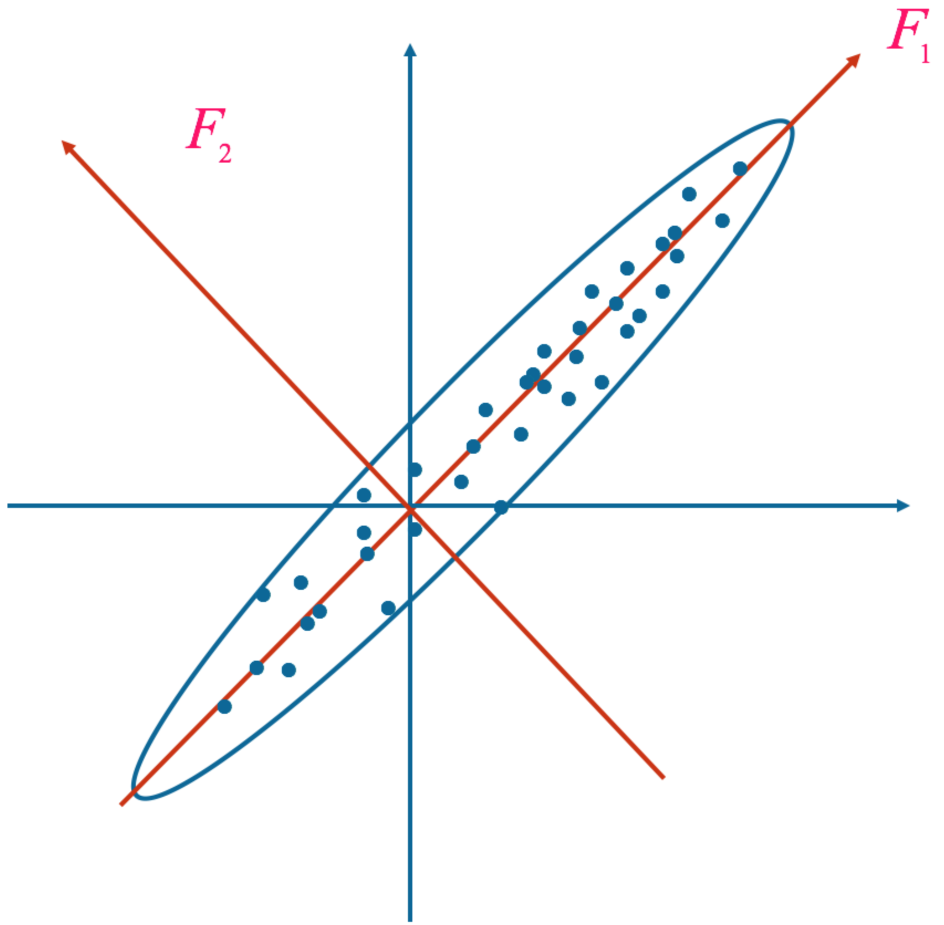
特征值和特征向量

- 特征值与特征向量：设A为n阶方阵，若数λ 和n维的非零列向量x，使关系式Ax=λx成立，则称数λ为方阵A的特征值，非零向量x称为A的对应与特征值λ的特征向量。

- 如何理解？

- 旋转变换的目的是为了使得n个样本点在F1轴方向上的离散程度最大，即F1的方差最大，变量F1代表了原始数据的绝大部分信息，在研究某经济问题时，即使不考虑变量F2也损失不多的信息。
- F1称为第一主成分，F2称为第二主成分。

- 特征值的大小代表了矩阵正交化之后所对应特征向量对于整个矩阵的贡献程度。特征向量就是旋转的新坐标体系。



特征值和特征向量

- 对称矩阵的特征值、特征向量
  - 协方差矩阵、转移概率矩阵等
  - 实对称矩阵的属于不同特征值的特征向量是正交的。
  - **$A * p1 = \lambda 1 * p1$**
  - **$A * p2 = \lambda 2 * p2$**
  - 如**p1**和**p2**正交，则必有**p1' * p2** = 0

特征值和特征向量

求法：

$$A = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \Rightarrow A\vec{v} = \lambda\vec{v} \Rightarrow Det(A - \lambda I) = 0.$$

$$Det\begin{pmatrix} 2-\lambda & 3 \\ 2 & 1-\lambda \end{pmatrix} = 0. \Rightarrow (2-\lambda)(1-\lambda) - 6 = 0$$

特征值和特征向量

$$\lambda_1 = -1,$$

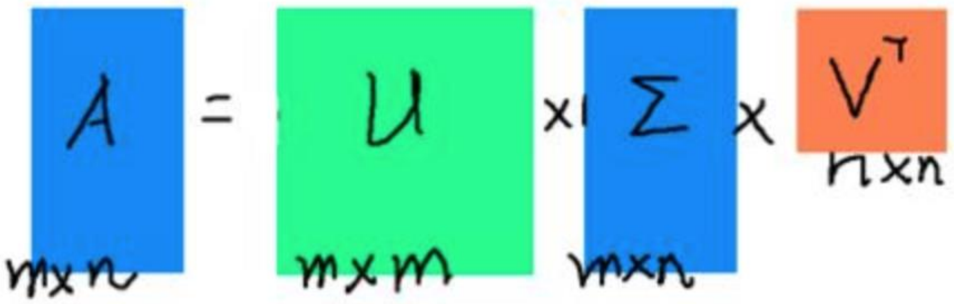
$$\lambda_2 = 4.$$

$$\rightarrow \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} = -1 \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} \rightarrow \vec{v}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} = 4 * \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} \rightarrow \vec{v}_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

SVD 分解

- 假设M是一个m×n阶矩阵（任意的），其中的元素全部实数。如此则存在一个分解使得：



$$A = \begin{bmatrix} u_1 & \cdots & u_k \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \end{bmatrix}$$

- 将上述矩阵展开

$$A = \sigma_1 u_1 v_1^{\mathrm{T}} + \sigma_2 u_2 v_2^{\mathrm{T}} + \ldots + \sigma_r u_r v_r^{\mathrm{T}}$$

SVD分解


前n个主成分：4
前n个主成分：8
前n个主成分：12
前n个主成分：16
前n个主成分：20
前n个主成分：24
前n个主成分：28
前n个主成分：32


你别吵

矩阵微积分

向量对标量：

$$\frac{\partial \mathbf{y}}{\partial x} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$$

标量对向量：

$$\frac{\partial y}{\partial \mathbf{x}} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}$$

矩阵微积分

向量对标量：

$$\frac{\partial \mathbf{y}}{\partial x} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$$

标量对向量：

$$\frac{\partial y}{\partial \mathbf{x}} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}$$

The derivative of $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$ with respect to $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

雅克比矩阵+零空间

$$y_1 = x_1^2 - 2x_2,\ y_2 = x_3^2 - 4x_2$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \frac{\partial y_1}{\partial x_3} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_2}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 2x_1 & -2 & 0 \\ 0 & -4 & 2x_3 \end{bmatrix}$$

矩阵对标量

$$\frac{\partial \mathbf{Y}}{\partial x} \overset{\text{def}}{=} \begin{bmatrix} \frac{\partial Y_{11}}{\partial x} & \frac{\partial Y_{12}}{\partial x} & \cdots & \frac{\partial Y_{1n}}{\partial x} \\ \frac{\partial Y_{21}}{\partial x} & \frac{\partial Y_{22}}{\partial x} & \cdots & \frac{\partial Y_{2n}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Y_{m1}}{\partial x} & \frac{\partial Y_{m2}}{\partial x} & \cdots & \frac{\partial Y_{mn}}{\partial x} \end{bmatrix}$$

标量对矩阵：梯度矩阵

$$\frac{\partial y}{\partial \mathbf{X}} \overset{\text{def}}{=} \begin{bmatrix} \frac{\partial y}{\partial X_{11}} & \frac{\partial y}{\partial X_{21}} & \cdots & \frac{\partial y}{\partial X_{m1}} \\ \frac{\partial y}{\partial X_{12}} & \frac{\partial y}{\partial X_{22}} & & \frac{\partial y}{\partial X_{m2}} \\ & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial X_{1n}} & \frac{\partial y}{\partial X_{2n}} & \cdots & \frac{\partial y}{\partial X_{mn}} \end{bmatrix}$$

$\mathbf{y}$	$\dfrac{\partial \mathbf{y}}{\partial \mathbf{x}}$
$\mathbf{Ax}$	$\mathbf{A}$
$\mathbf{x}^T\mathbf{A}$	$\mathbf{A}^T$
$\mathbf{x}^T\mathbf{x}$	$2\mathbf{x}^T$
$\mathbf{x}^T\mathbf{Ax}$	$\mathbf{x}^T\mathbf{A} + \mathbf{x}^T\mathbf{A}^T$

链式法则

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \text{and } \mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_r \end{bmatrix}$$

$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}} \overset{\text{def}}{=} \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} & \cdots & \frac{\partial z_1}{\partial x_n} \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} & & \frac{\partial z_2}{\partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial z_r}{\partial x_1} & \frac{\partial z_r}{\partial x_2} & \cdots & \frac{\partial z_r}{\partial x_n} \end{bmatrix}, \quad \text{where } \frac{\partial z_i}{\partial x_j} = \sum_{k=1}^{m} \frac{\partial z_i}{\partial y_k} \frac{\partial y_k}{\partial x_j} \quad \begin{cases} i = 1,2,\cdots,r \\ j = 1,2,\cdots,n \end{cases}$$

$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \begin{bmatrix} \sum \frac{\partial z_1}{\partial y_k}\frac{\partial y_k}{\partial x_1} & \sum \frac{\partial z_1}{\partial y_k}\frac{\partial y_k}{\partial x_2} & \cdots & \sum \frac{\partial z_1}{\partial y_k}\frac{\partial y_k}{\partial x_n} \\ \sum \frac{\partial z_2}{\partial y_k}\frac{\partial y_k}{\partial x_1} & \sum \frac{\partial z_2}{\partial y_k}\frac{\partial y_k}{\partial x_2} & & \sum \frac{\partial z_2}{\partial y_k}\frac{\partial y_k}{\partial x_n} \\ \vdots & & \ddots & \vdots \\ \sum \frac{\partial z_r}{\partial y_k}\frac{\partial y_k}{\partial x_1} & \sum \frac{\partial z_r}{\partial y_k}\frac{\partial y_k}{\partial x_2} & \cdots & \sum \frac{\partial z_r}{\partial y_k}\frac{\partial y_k}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial z_1}{\partial y_1} & \frac{\partial z_1}{\partial y_2} & \cdots & \frac{\partial z_1}{\partial y_m} \\ \frac{\partial z_2}{\partial y_1} & \frac{\partial z_2}{\partial y_2} & & \frac{\partial z_2}{\partial y_m} \\ \vdots & & \ddots & \vdots \\ \frac{\partial z_r}{\partial y_1} & \frac{\partial z_r}{\partial y_2} & \cdots & \frac{\partial z_r}{\partial y_m} \end{bmatrix} \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & & \frac{\partial y_2}{\partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} = \frac{\partial \mathbf{z}}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{x}}$$

$$E(\mathbf{w}) = \frac{1}{N}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \frac{1}{N}(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y})$$
$$= \frac{1}{N}(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y})$$

$\mathbf{y}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$
$\mathbf{A}\mathbf{x}$	$\mathbf{A}$
$\mathbf{x}^T\mathbf{A}$	$\mathbf{A}^T$
$\mathbf{x}^T\mathbf{x}$	$2\mathbf{x}^T$
$\mathbf{x}^T\mathbf{A}\mathbf{x}$	$\mathbf{x}^T\mathbf{A} + \mathbf{x}^T\mathbf{A}^T$

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

部分参考资料：

http://en.wikipedia.org/wiki/Matrix_calculus

http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf