数据科学基础 I (Matlab)

东北大学

数据科学中的数学基础与运算基础

函数

- 函数表示量与量之间的关系

$$y = f(x)$$

因变量 | 自变量

量

- 标量
- 向量
- 矩阵
- 张量

更广泛的意义：映射，即从x所在的“定义域”到y所在的“值域”的映射。

函数

```
>> x=4
x =
    4
>> y=sqrt(x)    %sqrt是平方根函数
y =
    2
>> z=sin(y)     %sin是正弦三角函数
z =
  0.9093
```

```
>> x=[4 5 6]; y=[1 2 3];

>> z=dot(x,y) %向量的内积函数
z =
    32
```

向量的长度

- 设向量 $x = [\ x_1\ \ x_2\ \ \dots\ \ x_n\ ]$
- 则向量长度 $\|x\| = \sqrt{[x,x]} = \sqrt{\sum_{i=1}^{n} x_i^2}$

$\Downarrow$

向量自身内积

向量的范数(norm)

- 设向量 $x = [\ x_1\ \ x_2\ \ \dots\ \ x_n\ ]$
- 则有:
  - 该向量的1范数为各个元素的绝对值之和

$$\|x\|_1 = \sum_{i=1}^{N} |x_i|$$

向量的范数(norm)

- 设向量 $x = [\ x_1\ \ x_2\ \ \ldots\ \ x_n\ ]$
- 则有：
  - 该向量的2范数为各个元素的平方和的平方根

$$\|x\|_2 = \sqrt{\sum_{i=1}^{N} |x_i|^2}$$

向量的范数(norm)

- 设向量 $x = [\ x_1\ \ x_2\ \ \ldots\ \ x_n\ ]$
- 则有：
  - 该向量的负无穷范数为所有元素的绝对值中最小的

$$\|x\|_{-\infty} = min|x_i|$$

向量的范数(norm)

- 设向量 $x = [\ x_1\ \ x_2\ \ \ldots\ \ x_n\ ]$
- 则有：
  - 该向量的正无穷范数为所有元素的绝对值中最大的

$$\|x\|_{+\infty} = max|x_i|$$

向量的范数(norm)

- 设向量 $x = [\ x_1\ \ x_2\ \ \dots\ \ x_n\ ]$
- 则有:
  - 该向量的p范数为p次方和的p次方根

$$L_p = \|x\|_p = \sqrt[p]{\sum_{i=1}^{N} |x_i|^p}$$

矩阵范数

- 对任意矩阵$A_{m\times n}$，其元素为a_{ij}。
  - 矩阵的1-范数（列和范数）：矩阵的每一列上的元素绝对值先求和，再从中取最大值。（列和最大）

$$\|A\|_1 = \max_{1\le j\le n}\sum_{i=1}^{m}|a_{ij}|$$

矩阵范数

- 对任意矩阵$A_{m\times n}$，其元素为a_{ij}。
  - 矩阵的无穷范数（行范数）：矩阵的每一行上的元素绝对值先求和，再从中取最大值。（行和最大）

$$\|A\|_\infty = \max_{1\le j\le m} \sum_{j=1}^{n} |a_{ij}|$$

矩阵范数

- 对任意矩阵$A_{m\times n}$，其元素为a_{ij}。
  - 矩阵的2-范数（欧几里德范数，谱范数）：矩阵A^TA的最大特征值的平方根。

$$\|A\|_2=\sqrt{\lambda_{max}(A^TA)}$$

矩阵范数

```
>> A=[-1 0 5; 2 4 6; 0 4 5]
A =
    -1     0     5
     2     4     6
     0     4     5
```

```
>> n=norm(A,1)
n =
    16
>> n=norm(A,2)
n =
    10.5625
>> n=norm(A,inf)
n =
    12
```

矩阵范数

- 对任意矩阵$A_{m\times n}$，其元素为a_{ij} 。
  - 矩阵的L0范数：非0元素的个数，通常用它来表示稀疏，L0范数越小0元素越多，也就越稀疏。
  - 矩阵的L1范数：矩阵中的每个元素绝对值之和。

矩阵范数

- 对任意矩阵$A_{m\times n}$，其元素为a_{ij}。
  - 矩阵的F范数：矩阵的各个元素平方之和再开平方根，它通常也叫做矩阵的L2范数。

$$\|A\|_F=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2}$$

矩阵的秩

- 指矩阵中所有行向量中极大线性无关组的元素个数

```
>> A=[-1 0 5; 2 4 6; 0 4 5]
A =
    -1     0     5
     2     4     6
     0     4     5
>> rank(A)
ans =
     3
```

```
>> A=[1 2 3; 2 4 6; 0 4 5]
A =
     1     2     3
     2     4     6
     0     4     5
>> rank(A)
ans =
     2
```

线性相关

矩阵的特征向量与特征值

- 设A是n阶方阵，如果数λ和n维非零列向量x使关系式$Ax=\lambda x$成立，那么λ称为矩阵A特征值，非零向量x称为A的对应于特征值λ的特征向量。

- 把每个特征向量看做是一个坐标轴，特征值是对应坐标轴（即特征向量）的坐标值。简单来说，就是用特征值（坐标）与特征向量（坐标轴）来表示原矩阵。

矩阵的特征向量与特征值

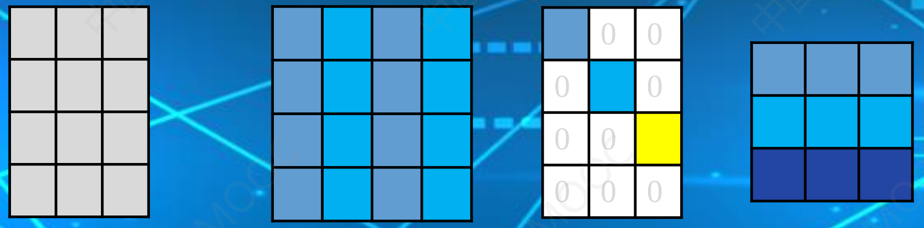
```
>> A = [2 1 0;1 3 1;0 1 4]
A =
     2     1     0
     1     3     1
     0     1     4
```

```
>> [V,D] = eig(A)
V =
    0.7887    -0.5774    0.2113
   -0.5774    -0.5774    0.5774
    0.2113     0.5774    0.7887
```

第一组特征值与相应特征向量　第二组特征值与相应特征向量　第三组特征值与相应特征向量

```
D =
    1.2679         0         0
         0    3.0000         0
         0         0    4.7321
```

特征向量应用：SVD分解



$M = U\ \Sigma\ V^*$

m×n　m×m　m×n　n×n

特征向量矩阵

特征值构成的对角阵

SVD分解用于图像压缩

原图：

298x749=223202

取100个特征：

U:298x100；

S:100x100；

V:100x749；

合计：114700

SVD分解用于图像压缩

```
a=imread(‘neu.jpg');
a = a(:,:,1);
imshow(mat2gray(a))
[m, n]=size(a);
a=double(a);
r=rank(a);
[U, S, V]=svd(a);
k = 100; %取前k个特征
re=U(:,1:k)*S(1:k,1:k)*V(:,1:k)
';
figure;
imshow(mat2gray(re));
```