第一章 数据挖掘与分析

任务目标

1、数据矩阵

2、属性

3、数据的几何和代数描述

4、数据：概率的观点

5、数据挖掘

6、补充阅读

7、习题

相关知识

1、线性代数、Python的Numpy和Pandas。

1.1 数据矩阵

数据经常用$n \times d$的矩阵表示，下面是n行d列的数据，其中行代表数据集中的数据，列代表数据中的特征或属性。

$$D = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline x_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ x_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array} \right)$$

其中x_i表示i行的一个d元组：

$$x_i = (x_{i1}, x_{i2}, \cdots x_{id})$$

而X_j表示j列的一个n元组：

$$X_j = (x_{1j}, x_{2j}, \cdots x_{nj})$$

1. 根据应用领域不同，数据矩阵的行还可以被称为实体、实例、样本、记录、事务、对象、数据点、特征向量、元组等。列可以被称为属性、性质、特征、维度、变量、域。
2. n被称为数据集的数据量，属性d被称为数据的维度。
3. 针对单个属性进行的分析，称为一元分析，针对两个属性进行的分析，称为二元分析，针对两个以上属性进行的分析，称为多元分析。

例1.1　表1-1列举了鸢尾花(iris)数据集的一部分数据。完整的数据集是一个150×5的矩阵。每一行代表一株鸢尾花，包含的属性有：萼片长度、萼片宽度、花瓣长度、花瓣宽度（以厘米计），以及该鸢尾花的类型。第一行是一个五元组：

$$x_1 = (5.9, 3.0, 4.2, 1.5, iris - versicolor)$$

表1-1 鸢尾花数据集的一部分数据

$$D = \left(\begin{array}{c|ccccc} & \text{萼片长度} & \text{萼片宽度} & \text{花瓣长度} & \text{花瓣宽度} & \text{类型} \\ & X_1 & X_2 & X_3 & X_4 & X_5 \\ \hline x_1 & 5.9 & 3.0 & 4.2 & 1.5 & iris - versicolor \\ x_2 & 6.9 & 3.1 & 4.9 & 1.5 & iris - versicolor \\ x_3 & 6.6 & 2.9 & 4.6 & 1.3 & iris - versicolor \\ x_4 & 4.6 & 3.2 & 1.4 & 0.2 & iris - versicolor \\ x_5 & 6.0 & 2.2 & 4.0 & 1.0 & iris - versicolor \\ x_6 & 4.7 & 3.2 & 1.3 & 0.2 & iris - versicolor \\ x_7 & 6.5 & 3.0 & 5.8 & 2.2 & iris - versicolor \\ x_8 & 5.8 & 2.7 & 5.1 & 1.9 & iris - versicolor \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ x_{149} & 7.7 & 3.8 & 6.7 & 2.2 & iris - versicolor \\ x_{150} & 5.1 & 3.4 & 1.5 & 0.2 & iris - versicolor \end{array}\right)$$

1.2 属性

1. 数值型属性

   数值型属性是在实数或整数域内取值。取值域为$\mathbb{N}$的属性Age(年龄)。花瓣长度也是一个数值型属性，其取值域为$\mathbb{R}^+$。

   - 区间标度类 温度
   - 比例标度类

2. 类别型属性

   类别型属性的定义域是由一个定值符号集合定义的。sex和Education都是类别型属性。

   $$domain(Sex) = \{Female, Male\}$$

   $$domain(Education) = \{HighSchool, BS, MS, PhD\}$$

   - 名义类
   - 次序类
   -

1.3 数据的几何和代数描述

若数据矩阵D的d个属性或维度都是数值型的，则每一行都可以看作一个d维空间的点：

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{id}) \in \mathbb{R}^d$$

或每一行可以等价地看作一个d维的列向量（所有向量都默认为列向量）：

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} = (x_{i1}, x_{i2}, \cdots, x_{id})^T \in \mathbb{R}^d$$

其中T是矩阵的转置算子。

d维笛卡尔坐标空间是由d个单位向量定义的，又被称为标准基，每个轴方向上一个。第j个标准基向量e_j是一个d维的单位向量，该向量的第j个分量是1，其他分量是0。

$\mathbb{R}^d$中的任何向量都可以由标准基向量的线性组合。

$$x_i = x_{i1}e_1 + x_{i2}e_2 + \cdots + x_{id}e_d = \sum_{j=1}^{d} x_{ij}e_j$$

例1.2　表1-1列举了鸢尾花(iris)数据集的一部分数据。完整的数据集是一个150×5的矩阵。每一行代表一株鸢尾花，包含的属性有：萼片长度、萼片宽度、花瓣长度、花瓣宽度（以厘米计），以及该鸢尾花的类型。第一行是一个五元组：

$$x_1 = 5.9e_1 + 3.0e_2 + 4.2e_3 = 5.9\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 3.0\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 4.2\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 5.9 \\ 3.0 \\ 4.2 \end{pmatrix}$$

1. 下面利用Python实现三个向量的相加。

   Python3.exe安装到D盘根目录下的Python37文件夹下（D:\Python37）目录，然后点击“我的电脑”-"属性"-“环境变量”-在系统环境变量中设置Path变量后面添加";D:\Python37;D:\Python37\Scripts"。

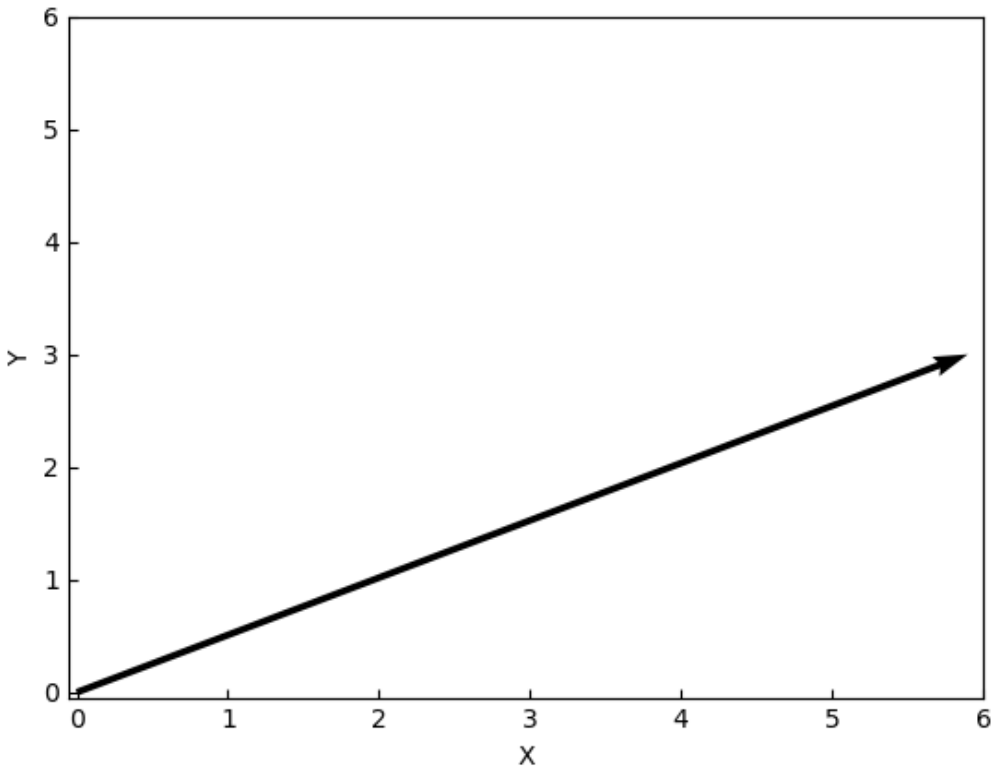
```
python3 -m pip install --upgrade pip --force-reinstall -i
https://pypi.tuna.tsinghua.edu.cn/simple #升级Python3的pip3工具
pip3 install numpy -i https://pypi.www.douban.com/simple
pip3 install pandas -i https://pypi.www.douban.com/simple
pip3 install matplotlib -i https://pypi.www.douban.com/simple
pip3 install scipy -i https://pypi.www.douban.com/simple
pip3 install ipython3 -i https://pypi.www.douban.com/simple
import numpy as  np
x1 = np.array([1,0,0])
x2 = np.array([0,1,0])
x3 = np.array([0,0,1])
b = 5.9*x1+3.0*x2+4.2*x3;
```

2. 通过matplotlib进行绘图。

```
from mpl_toolkits.mplot3d import axes3d;
import matplotlib.pyplot as plt;
import numpy as np;
plt.rcParams['xtick.direction'] = 'in' #x的刻度向内
plt.rcParams['ytick.direction'] = 'in' #y的刻度向内
if __name__=="__main__":
    plt.figure();
    ax = plt.gca();
    plt.quiver(0, 0, 5.9, 3.0, angles='xy', scale_units='xy', scale=1);
    plt.xticks(np.arange(0, 7));
    plt.yticks(np.arange(0, 7));
    plt.xlabel('X');
    plt.ylabel('Y');
```

```
    plt.show();
```

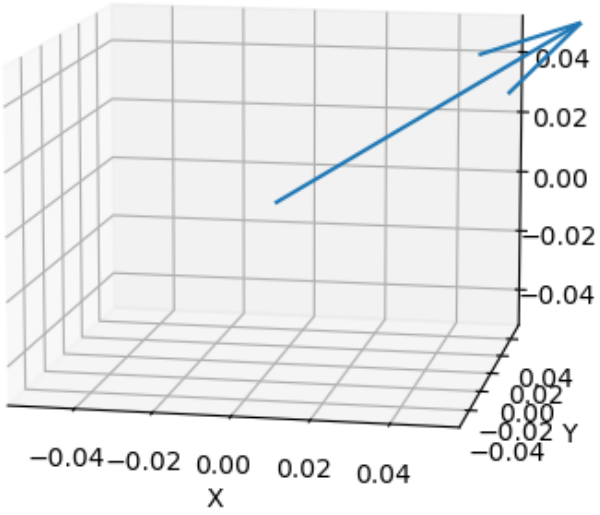
3. 绘制箭头




```
from mpl_toolkits.mplot3d import axes3d;
import matplotlib.pyplot as plt;
import numpy as np;

plt.rcParams['xtick.direction'] = 'in' #x的刻度向内
plt.rcParams['ytick.direction'] = 'in' #y的刻度向内

if __name__=="__main__":
    plt.figure();
    ax = plt.gca(projection='3d');
    plt.quiver(0,0,0,5.9,3.0,4.2,length=0.1, normalize=True);
    plt.xlabel('X');
    plt.ylabel('Y');
    plt.show();
```



每一个数值型的列或属性还可以看成n维空间$\mathbb{R}^n$中的一个向量：

$$X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

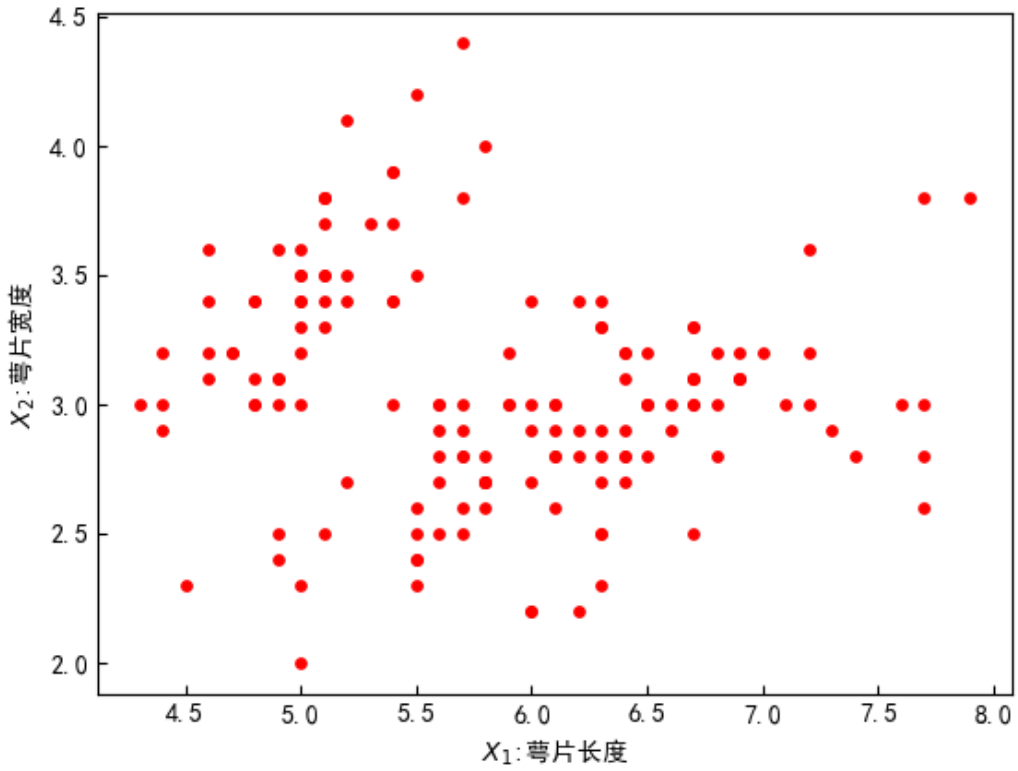
如果所有的属性都是数值型的，那么数据矩阵D事实上是一个$n \times d$的矩阵，可记作$D \in \mathbb{R}^{n \times d}$，如以下公式所示：

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix} = \begin{pmatrix} -x_1^T- \\ -x_2^T- \\ \vdots \\ -x_n^T- \end{pmatrix} == \begin{pmatrix} | & | & & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & & | \end{pmatrix}$$

```
import pandas as pd
df = pd.read_csv('iris.csv',sep=',');
print(df.head(5))
   Unnamed: 0  Sepal.Length  ...  Petal.Width  Species
0           1           5.1  ...          0.2   setosa
1           2           4.9  ...          0.2   setosa
2           3           4.7  ...          0.2   setosa
3           4           4.6  ...          0.2   setosa
4           5           5.0  ...          0.2   setosa

[5 rows x 6 columns]
print(df['Sepal.Length'].mean());
5.843333333333334
print(df['Sepal.Length'].std());
0.828066127977863
import matplotlib.pyplot as plt;
```

```
plt.rcParams['font.sans-serif'] = ['SimHei'];
plt.rcParams['xtick.direction'] = 'in';
plt.rcParams['ytick.direction'] = 'in';
plt.scatter(df['Sepal.Length'],df['Sepal.Width'],s =15,color='red');
plt.xlabel('$X_1$:萼片长度');
plt.ylabel('$X_2$:萼片宽度');
plt.show()
```



1.3.1 距离和角度

将数据实例和属性用向量来描述或将整个数据集描述为一个矩阵，

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

1. 点乘

   a和b的点乘定义如下：

$$a^T b = \begin{pmatrix} a_1 & a_2 & \cdots & a_m \end{pmatrix} \times \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

$$= a_1 b_1 + a_2 b_2 + \cdots + a_m b_m$$

$$= \sum_{i=1}^{m} a_i b_i$$

2. 长度

向量$a \in \mathbb{R}^m$的欧几里得范数或长度定义为：

$$||a|| = \sqrt{a^T a} = \sqrt{a_1^2 + a_2^2 + \cdots + a_m^2} = \sqrt{\sum_{i=1}^{m} a_i^2}$$

a的方向上的单位向量定义为：

$$u = \frac{a}{||a||} = \left(\frac{1}{||a||}\right) a$$

根据定义，单位向量的长度为$||u|| = 1$，正则化向量。

3. 距离

$$\delta(a,b) = ||a-b|| = \sqrt{(a-b)^T(a-b)} = \sqrt{\sum_{i=1}^{m}(a_i - b_i)^2}$$

4. 角度

$$cos\theta = \frac{a^T b}{||a||||b||} = \left(\frac{a}{||a||}\right)^T \left(\frac{b}{||b||}\right)$$

柯西-施瓦茨（Cauchy-Schwartz）不等式描述了对于$\mathbb{R}^m$中的任意向量a和b，若满足如下关系：

$$|a^T b| \leqslant ||a|| \cdot ||b||$$

则根据柯西-施瓦茨不等式马上可以

$$-1 \leqslant cos\theta \leqslant 1$$

由于两个向量之间的最小角$\theta \in [0^\circ, 180^\circ]$且$cos\theta \in [-1, 1]$，余弦相似度取值范围为$+1$（对应$0^\circ$）到$-1$（对应$180^\circ$角，或是$\pi$弧度）。

5. 正交性

a和b向量是正交的，当且仅当$a^T b = 0$，这意味着$cos\theta = 0$，即两个向量之间的角度是90°或是弧度为$\frac{\pi}{2}$。

例1.3　两个向量是：

$$a = \begin{pmatrix} 5 \\ 3 \end{pmatrix} 和 b = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$

```
import numpy as np
a = np.array([5,3])
b = np.array([1,4])
print(a-b)
np.sqrt(np.sum(c**2))
ua = a /np.sqrt(np.dot(a,a))
```

1.3.2 均值和总方差

1. 均值

数据矩阵D的均值（mean）由所有点的向量取平均值。

$$mean(D) = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. 总方差

数据矩阵D的总方差由每个点到均值的均方差距离得到

$$var(D) = \frac{1}{n}\sum_{i=1}^{n}\delta(x_i,\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}||x_i-\mu||^2$$

化简公式（1.4），可以得到：

$$\begin{aligned} var(D) &= \frac{1}{n}\sum_{i=1}^{n}(||x_i||^2 - 2x_i^T\mu + ||\mu||^2) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}||x_i||^2 - 2n\mu^T\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right) + n||\mu||^2\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}||x_i||^2 - 2n\mu^T\mu + n||\mu||^2\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}||x_i||^2\right) - ||\mu||^2 \end{aligned}$$

因此，总方差即是所有数据点的长度平方的平均值减去均值长度的平方。

3. 居中数据矩阵

通常需要将矩阵居中，以使得矩阵的均值和数据空间的原点重合，居中数据可以通过将所有数据减去均值

$$Z = D - 1\cdot\mu^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} - \begin{pmatrix} \mu^T \\ \mu^T \\ \vdots \\ \mu^T \end{pmatrix} = \begin{pmatrix} x_1^T-\mu^T \\ x_2^T-\mu^T \\ \vdots \\ x_n^T-\mu^T \end{pmatrix} = \begin{pmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_n^T \end{pmatrix}$$

其中，$z_i = x_i - \mu$代表与x_i对应的居中数据点，$1\in\mathbb{R}^n$是所有元素都为1的n维向量，居中矩阵Z的均值是$0\in\mathbb{R}^d$正交投影

4. 正交投影

在数据挖掘中，经常需要将一个点或向量投影到另一个向量上。

a，$b\in\mathbb{R}^m$为两个m维的向量。向量b在向量a方向上的正交分解(orthogonal decomposition)，

$$b = b_{||} + b_{\perp} = p + r$$

其中$p = b_{||}$与a平行，$r = b_{\perp}$与a垂直（正交）。

向量p与向量a是平行的，因此对于某个标量c，有$p = ca$。因此，$r = b - p = b - ca$。由于p和r是正交的，有

$$p^Tr = (ca)^T(b-ca) = ca^Tb - c^2a^Ta = 0$$

$$c = \frac{a^Tb}{a^Ta}$$

因此，b在a上的投影为

$$p = b_{||} = ca = \left(\frac{a^Tb}{a^Ta}\right)a$$

例1.4　鸢尾花数据的前两个维度，萼片长度和萼片宽度，平均点：

$$mean(D) = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$$

5. 线性无关与维数

给定一个数据矩阵

$$D = (x_1 \quad x_2 \quad \cdots \quad x_n)^T = (X_1 \quad X_2 \quad \cdots \quad X_d)$$

在m维的向量空间$\mathbb{R}^m$中给定任意一组向量$v_1, v_2, \cdots, v_k$，它们的线型组合定义为：

$$c_1 v_1 + c_2 v_2 + \cdots + c_k v_k$$

其中，$c_i \in \mathbb{R}$ 是标量值。k个向量的所有可能的线性组合称为空间(span)，可表示为span $(v_1, \cdots, v_k)$，改空间是$\mathbb{R}^m$的一个子向量空间。若$span(v_1, \cdots, V_k) = \mathbb{R}^m$，则称$v_1, \cdots, v_k$为$\mathbb{R}^m$的一个生成集合(Spanning set)。

1、行空间和列空间

数据矩阵D有几个有趣的向量空间，其中两个是D的行空间与列空间。D的列空间，用$col(D)$表示，是d个点d个属性$x_j \in \mathbb{R}^n$的所有线性组合的集合。

$$col(D) = span(X_1, X_2, \cdots, X_d)$$

根据定义，$col(D)$是$\mathbb{R}^n$的一个子空间。D的行空间，用$row(D)$表示，是n个点$x_i \in \mathbb{R}^n$的所有线性组合的集合，

$$row(D) = span(x_1, x_2, \cdots, x_n)$$

根据定义，$row(D)$是$\mathbb{R}^d$的一个子空间。注意，D的行空间是D^T的列空间：

$$row(D) = col(D^T)$$

2、线性无关

给定一组向量$v_1, \cdots, v_k$，若其中至少有一个向量可以由其他向量线性表示，则称这些向量线性相关，$\iff$，若由k个标量$c_1, c_2, \cdots, c_k$，其中至少有一个不为0的情况下，可以使得

$$c_1 v_1 + c_2 v_2 + \cdots + c_k v_k = 0$$

成立，则这k个向量是线性相关的。

另一方面，这k个向量是线性无关的，当且仅当

$$c_1 v_1 + c_2 v_2 + \cdots + c_k v_k = 0$$
$$\iff c_1 = c_2 = \cdots = c_k = 0$$

3、维数和秩

假设S是$\mathbb{R}^m$的一个子空间，S的基(basis)是指S中一组线性无关的向量$v_1, \cdots, v_k$，这组向量生成S，即$span(v_1, \cdots, v_k) = S$。事实上，基是一个最小生成集。若基中的向量两两正交，则称该基是S的正交基（orthogonal basis），此外，若这些向量还是单位向量，则它们构成了S的一个标准正交基(orthogonal basis)。

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} \quad \cdots \quad e_m = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

S的任意两个基数都必须有相同数目的向量，该数目称作S的维数，表示为$dim(S)$。S是$\mathbb{R}^m$的一个子空间，因此$dim(S) \leqslant m$。

值得注意的是。任意数据矩阵的行空间和列空间的维数都是一样的，维数又可称为矩阵的秩。对于数据矩阵$D \in \mathbb{R}^{n \times d}$，即$rank(D) \leqslant min(n,d)$，因此列空间的维数最多是$d$，行空间的维数最多是$n$。因此，尽管表面上数据点是在一个$d$维的属性矩阵（外在维数）中，若$rank(D) < d$，则数据点实际上都处在比$\mathbb{R}^d$更低维的一个子空间里，$rank(D)$给出了数据点的内在维数。事实上，通过降维方法，可以用一个导出的数据矩阵$D' \in \mathbb{R}^{n \times k}$来近似数据矩阵$D \in \mathbb{R}^{n \times d}$的，这样维数会大大降低，即$k \ll d$。在这种情况下，$k$能够反映数据的真实内在维度。

例1.5　图1-5中的直线ℓ是由 $\ell = span((-2.15, 2.75)^T)$定义的，且$dim(\ell) = 1$。正则化后，可以获得$\ell$的标准正交基：

$$\frac{1}{\sqrt{12.19}} \begin{pmatrix} -2.15 \\ 2.75 \end{pmatrix} = \begin{pmatrix} -0.615 \\ 0.788 \end{pmatrix}$$

1.4 数据：概率观点

数据的概率观点假设每个数值型的属性X都是一个随机变量（random variable），是由一个实验（一个观察或测量的过程）的每个结果赋一个实数值的函数定义。形式化定义如下：X是一个函数$X : \mathcal{O} \to \mathbb{R}$，其中$\mathcal{O}$作为$X$的定义域，是实验所有可能结果的集合，又被称作样本空间(sample space)；$\mathbb{R}$代表X的值域，是实数集合。如果实验结果是数值型的，且与随机变量X的观测值相同，则$X : \mathcal{O} \to \mathcal{O}$就是恒等函数：$X(v) = v, v \in \mathcal{O}$。实验结果与随机变量取值之间的区别是很重要，在不同的情况下，可能会对观测值采用不同的处理方式。

例1.6　考虑表1-1中萼片长度属性(X_1)。该属性的所有$n = 150$个取值都列在表1-2中，落在$[4.3, 7.9]$的范围之内（测量单位是厘米），假设这个取值范围就是所有可能的实验结果$\mathcal{O}$。

默认情况下，认为X_1是一个连续随机变量，表示为$x_1(v) = v$，因为所有的结果（萼片长度值）都是数值型。

另一方面，如果要区分长萼片（大于等于7厘米）和短萼片的鸢尾花，可以定义一个离散随机变量A

$$A(v) = \begin{cases} 0 & v < 7 \\ 1 & v \geqslant 7 \end{cases}$$

在这个例子中，A的定义域是$[4.3, 7.9]$，值域是$\{0, 1\}$。因此A仅在离散值0和1上取非零概率。

1. 概率分布（质量）函数

   若X是离散的，则X的概率分布（质量)函数可定义为：

   $$f(x) = P(X = x), \quad x \in \mathbb{R}$$

   换句话说，函数f给出了随机变量X取值x的概率$P(X = x)$。概率质量函数（probability mass function）表示了概率聚集在X值域内的几个离散值上，其余值上的概率为0。f必须要遵循概率的基本规则，即f必须为非负：

   $$f(x) \geqslant 0$$

   而且所有概率的和必须等于1，

   $$\sum_x f(x\ ) = 1$$

表1-2 鸢尾花的数据集：萼片长度(厘米)

5.9	6.9	6.6	4.6	6.0	4.7	6.5	5.8	6.7	6.7	5.1	5.1	5.7	6.1	4.9
5.0	5.0	5.7	5.0	7.2	5.9	6.5	5.7	5.5	4.9	5.0	5.5	4.6	7.2	6.8
5.4	5.0	5.7	5.8	5.1	5.6	5.8	5.1	6.3	6.3	5.6	6.1	6.8	7.3	5.6
4.8	7.1	5.7	5.3	5.7	5.7	5.6	4.4	6.3	5.4	6.3	6.9	7.7	6.1	5.6
6.1	6.4	5.0	5.1	5.6	5.4	5.8	4.9	4.6	5.2	7.9	7.7	6.1	5.5	4.6
4.7	4.4	6.2	4.8	6.0	6.2	5.0	6.4	6.3	6.7	5.0	5.9	6.7	5.4	6.3
4.8	4.4	6.4	6.2	6.0	7.4	6.2	7.0	5.5	6.3	6.8	6..1	6.5	6.7	6.7
4.8	4.9	6.9	4.5	4.3	5.2	7.4	6.4	5.2	5.8	5.5	7.6	6.3	6.4	6.3
5.8	5.0	6.7	6.0	5.1	4.8	5.2	5.1	6.6	6.4	5.2	6.4	7.7	5.8	4.9
5.4	5.1	6.0	6.5	5.5	7.2	4.8	6.2	6.5	6.0	5.4	5.5	6.7	7.7	5.1

例1.7 （伯努利和二项分布） 在例子1.6中， A被定义为一个离散型的随机变量，用于表示长萼片的长度。根据表1-2的萼片长度数据，可以看到只有13朵鸢尾花的萼片长度大于等于7厘米。据此，可以估计A的概率分布（质量）函数：

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import os

def poly(m,k):
    n=k;
    prod = 1;
    while n>0:
        prod=prod*(m-n+1)/n;
        n-=1;
    return prod;

df = pd.read_csv('iris.csv',sep=',');
x1 = df[df['Sepal.Length']>=7];
x2 = df[df['Sepal.Length']<7];
x1.count()/df.count()
x2.count()/df.count()
y1 = []
for k in range(0,11,1):
    y1.append(poly(10,k)*np.power(0.087,k)*np.power(0.913,10-k))

plt.stem(range(0,len(y1)),y1);
plt.show();
```

$$f(1) = P(A = 1) = \frac{13}{150} = 0.087 = p$$

以及

$$f(0) = P(A = 0) = \frac{137}{150} = 0.913 = 1 - p$$

A服从伯努利分布，参数$p \in [0, 1]$，p代表成功的概率，即从鸢尾花的数据中随机挑选一朵长萼片的概率。$1 - p$是鸢尾花的数据中无法挑选出长萼片的概率。

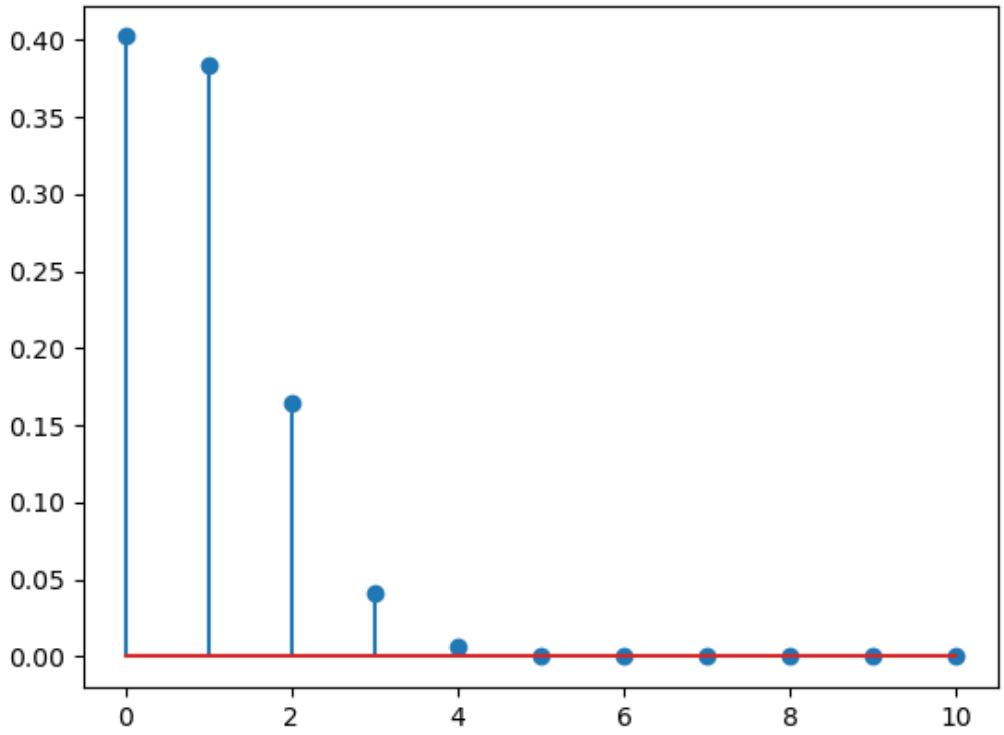
考虑另一个离散随机变量B，代表在m次相互独立的以p为成功概率的伯努利实验中挑选出长萼片鸢尾花的数目。B可以取$[0, m]$中的离散值，其概率质量函数可由伯努利分布给出：

$$f(k) = P(B = k) = \binom{m}{k} p^k (1-p)^{m-k}$$

以上公式可以按照如下理解，共有$\binom{m}{k}$种方法从m次尝试中挑选出k朵长萼片花。k次选择都成功的概率为p^k，而其余$m-k$次都失败的概率为$(1-p)^{m-k}$。由于$p = 0.087$，在$m = 10$次尝试中正好观察$k = 2$次长萼片的概率为：

$$f(2) = P(B = 2) = \binom{10}{2} (0.087)^2 (0.913)^8 = 0.164$$

图1-6显示了在$m = 10$的时候，对应不同的k值的概率质量函数。由于p值很小，在有限的几次尝试中获得k次成功的概率随着k的增大而迅速减小，当$k \geqslant 6$时，几乎为0。



2. 概率密度函数

若X是连续的，则其取值范围是整个实数集合$\mathbb{R}$。取任意特定值x的概率是1除以无穷多种可能，即对所有多种可能。即对所有$x \in \mathbb{R}$，$P(X = x) = 0$。

$$P(x \in [a, b]) = \int_a^b f(x)dx$$

考虑表1-1中萼片长度属性(X_1)。该属性的所有$n = 150$个取值都列在表1-2中，落在$[4.3, 7.9]$的范围之内（测量单位是厘米），假设这个取值范围就是所有可能的实验结果$\mathcal{O}$。

$$f(x) \geqslant 0, \quad x \in \mathbb{R}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P(X \in [x-\epsilon, x+\epsilon]) = \int_{x-\epsilon}^{x+\epsilon} f(x)dx \simeq 2\epsilon \cdot f(x)$$

$$f(x) \simeq \frac{P(X \in [x-\epsilon, x+\epsilon])}{2\epsilon}$$

尽管概率密度函数$f(x)$不能确定概率$P(X=x)$，但是

$$\frac{P(X \in [x_1-\epsilon, x_1+\epsilon])}{P(X \in [x_2-\epsilon, x_2+\epsilon])} \simeq \frac{2\epsilon \cdot f(x_1)}{2\epsilon \cdot f(x_2)} = \frac{f(x_1)}{f(x_2)}$$

若$f(x_1)$比$f(x_2)$大，则X的值靠近x_1的概率要大于靠近x_2的概率，反之亦然。

例1.8 （正态分布） 再次考虑鸢尾花数据集中的萼片长度值。假设那些值服从高斯或者正态密度函数。如下所示：

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

正态密度分布一共有两个参数，平均值μ和方差σ^2（这两个参数的含义在第2章讨论）。

尽管$f(x=\mu) = f(5.84) = \frac{1}{\sqrt{2\pi \cdot 0.681}} exp\{0\} = 0.483$,

$$P(X=\mu) \simeq 2\epsilon \cdot f(\mu) = 2\epsilon \cdot 0.483 = 0.967\epsilon$$

由于$\epsilon \to 0$，我们有$P(X=\mu) \to 0$。然而，根据公式（1.9）,可以观察到值靠近平均值$\mu = 5.84$的概率是观察到值靠近$x = 7$的概率的2.69倍。

$$\frac{f(5.84)}{f(7)} = \frac{0.483}{0.18} = 2.69$$

3. 累积分布函数

对于任意的随机变量X，无论是离散的还是连续的，可以定义累积分布函数(CDF)
$F: \mathbb{R} \to [0,1]$，该函数给出了观察到的最大值为某个给定点x的概率：

$$F(x) = P(X \leqslant x), \quad -\infty < x < \infty$$

当X为离散型时，F可以表示如下：

$$F(x) = P(X \leqslant x) = \sum_{u \leqslant x} f(u)$$

当X为连续型时，F可以表示如下：

$$F(x) = P(X \leqslant x) = \int_{-\infty}^{x} f(u)du$$

4. 二元随机变量

对一组属性X_1和X_2，除了将每个属性当作一个随机变量之外，可以把它们当作一个二元随机变量进行成对的分析

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

$X: \mathcal{O} \to \mathbb{R}^2$是对样本空间中的每个结果都赋予一对实数，或二维向量$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2$。和一元的情况相似，若结果值为数值型，则默认$X$为恒等函数。

（1） 联合概率质量函数

若X_1和X_2同为离散型随机变量，则X的联合概率质量函数如下所示：

$$f(x)=f(x_1,x_2)=P(X_1=x_1,X_2=x_2)=P(X=x)$$

f必须满足如下两个条件：

$$f(x)=f(x_1,x_2)\geqslant 0,\quad -\infty<x_1,x_2<\infty$$
$$\sum_{\infty}f(x)=\sum_{x_1}\sum_{x_2}f(x_1,x_2)=1$$

（2）联合概率密度函数

若X_1和X_2同为连续型随机变量，则X的联合概率密度函数如下所示：

$$P(X\in W)=\int\int_{x\in W}f(x)dx=\int\int_{(x_1,x_2)^T\in W}f(x_1,x_2)dx_1dx_2$$

其中，$W\subset\mathbb{R}^2$是二维实空间的子集。f同样必须满足如下两个条件：

$$f(\boldsymbol{x})=f(x_1,x_2)\geqslant 0,\quad -\infty<x_1,x_2<\infty$$
$$\int_{\mathbb{R}^2}f(\boldsymbol{x})dx=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x_1,x_2)dx_1dx_2=1$$

与一元的例子一样，对于任意的特定点$\boldsymbol{x}$，概率质量$P(\boldsymbol{x})=P((x_1,x_2))^T=0$。然而，可以用$f$来计算$x$处的概率密度。考虑方形区域$W=([x_1-\epsilon,x_1+\epsilon][x_2-\epsilon,x_2+\epsilon])$，即一个以$\boldsymbol{x}=(x_1,x_2)^T$为中心的宽度为$2\epsilon$的二维窗口。$x$处的概率密度可以近似地表示为

$$\begin{aligned}P(\boldsymbol{X}\in W)&=P(\boldsymbol{X}\in([x_1-\epsilon,x_1+\epsilon],[x_2-\epsilon,x_2+\epsilon]))\\&=\int_{x_1-\epsilon}^{x_1+\epsilon}\int_{x_2-\epsilon}^{x_2+\epsilon}f(x_1,x_2)dx_1dx_2\\&\simeq 2\epsilon\cdot 2\epsilon\cdot f(x_1,x_2)\end{aligned}$$

这意味着：

$$f(x_1,x_2)=\frac{P(\boldsymbol{X}\in W)}{(2\epsilon)^2}$$

因此，(a_1,a_2)对(b_1,b_2)的相对概率可以通过概率密度函数计算如下：

$$\frac{P(\boldsymbol{x}\in([a_1-\epsilon,a_1+\epsilon],[a_2-\epsilon,a_2+\epsilon]))}{P(\boldsymbol{X}\in([b_1-\epsilon,b_1+\epsilon],[b_2-\epsilon,b_2+\epsilon]))}\simeq\frac{(2\epsilon)^2\cdot f(a_1,a_2)}{(2\epsilon)^2\cdot f(b_1,b_2)}=\frac{f(a_1,a_2)}{f(b_1,b_2)}$$

例子1.10（二元分布），考虑鸢尾花数据集中的萼片长度和萼片宽度属性。用A表示和长萼片长度（大于等于7厘米）对应的伯努利随机变量，如例1.7中所定义。

定义另一个伯努利随机变量B对应长萼片宽度（比如大于等于3.5厘米）。令$\boldsymbol{X}=\begin{pmatrix}A\\B\end{pmatrix}$为一个离散型二元随机变量，则$\boldsymbol{X}$的联合概率质量函数可以用如下数据估算出来：

$$\begin{aligned}f(0,0)=P(A=0,B=0)&=\frac{116}{150}=0.773\\f(0,1)=P(A=0,B=1)&=\frac{21}{150}=0.140\\f(1,0)=P(A=1,B=0)&=\frac{10}{150}=0.067\\f(1,1)=P(A=1,B=1)&=\frac{3}{150}=0.020\end{aligned}$$

图1-10画出了以上概率密度质量函数。

将鸢尾花数据集（见表1-1）中的属性$\boldsymbol{X_1}$和$\boldsymbol{X_2}$当作连续型随机变量，可以定义一个连续型二元随机变量$\boldsymbol{X}=\begin{pmatrix}X_1\\X_2\end{pmatrix}$。假设$\boldsymbol{X}$服从二元正态分布，则其联合概率密度函数可以用下式：

$$f(x|\mu,\Sigma)=\frac{1}{2\pi\sqrt{|\Sigma|}}exp\left\{-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right\}$$

这里μ和Σ是二元正态分布的参数，分别代表二维的均值向量和协方差矩阵。$|\Sigma|$代表Σ的行列式。二元正态密度可见图1-11，其中，均值为

$$\mu=(5.843,3.054)^T$$

协方差矩阵为

$$\Sigma=\begin{pmatrix}06.81 & -0.039\\ -0.039 & 0.187\end{pmatrix}$$

有一点需要强调：函数$f(x)$仅针对x处的概率密度，且$f(x)\neq P(X=x)$。如前所述，有$P(X=x)=0$。

（3）联合累积分布函数

两个随机变量X_1和X_2的联合累积分布函数定义为函数F，其中，对于所有的x_1，$x_2\in(-\infty,\infty)$，有：

$$F(x)=F(x_1,x_2)=P(X_1\leqslant x_1,X_2\leqslant x_2)=P(X\leqslant x)$$

（4）统计独立性

若对于任意的$W_1\subset\mathbb{R}$和$W_2\subset\mathbb{R}$，都有：

$$P(X_1\in W_1,X_2\in W_2)=P(X_1\in W_1)\cdot P(X_2\in W_2)$$

则称为随机变量X_1和X_2是（统计）独立的。此外，若X_1和X_2是独立的，则如下两个条件也需要满足：

$$F(x)=F(x_1,x_2)=F(x_1)\cdot F(x_2)$$
$$f(x)=f(x_1,x_2)=f(x_1)\cdot f(x_2)$$

其中F_i是累积分布函数，f_i是随机变量X_i的概率质量函数或概率密度函数。

1.4.2 多元随机变量

d维多元随机变量$X=(X_1,X_2,\cdots,X_d)^T$又称为向量随机变量，定义为给样本空间中的每一个结果都赋一个实数向量的函数，即$X:\mathcal{O}\to\mathbb{R}^d$。$X$的值域可以用向量$x=(x_1,x_2,\cdots,x_d)^T$表示。若所有$X_j$都是数值型的，则$X$默认为恒等函数。换句话说，若所有属性都是数值型的，可以将样本空间里面的每个结果（数据矩阵里的每一个点）当作一个向量随机变量。另一方面，若有的属性是非数值型的，则X将结果映射为其值域上的数值型向量。

若所有的X_j都是离散的，则X是联合离散的，其联合概率质量函数f可以定义如下：

$$f(x)=P(X=x)$$
$$f(x_1,x_2,\cdots,x_d)=P(X_1=x_1,X_2=x_2,\cdots,X_d=x_d)$$

若所有的X_j都是连续的，则X是联合连续的，其联合概率密度函数如下：

$$P(X\in W)=\int\cdots\int_{x\in W}f(x)dx$$
$$P((X_1,X_2,\cdots,X_d)^T\in W)=\int\cdots\int_{(x_1,x_2,\cdots,x_d)^T\in W}f(x_1,x_2,\cdots,x_d)dx_1dx_2\cdots dx_d$$

其中$W\subseteq\mathbb{R}^d$。

概率的基本规则必须要满足，即$f(x) \geqslant 0$且X值域内所有x的$f(x)$之和要为1。对每一个点$x \in \mathbb{R}$，则$X = (x_1, \cdots, x_d)^T$的联合累积分布函数为：

$$F(x) = P(X \leqslant x)$$
$$F(x_1, x_2, \cdots, x_d) = P(X_1 \leqslant x_1, X_2 \leqslant x_2, \cdots, X_d \leqslant x_d)$$

$X_1, X_2, \cdots, X_d$是独立的随机变量，当且仅当对于任意的区域$W_i \subset \mathbb{R}$，有：

$$P(X_1 \in W_1, X_2 \in W_2, \cdots, X_d \in W_d)$$
$$P(X_1 \in W_1) \cdot P(X_2 \in W_2) \cdot \cdots \cdot P(X_d \in W_d)$$

若$X_1, X_2, \cdots, X_d$是独立的，则下述条件成立：

$$F(x) = F(x_1, \cdots, x_d) = F_1(x_1) \cdot F_2(x_2) \cdot \cdots \cdot F_d(x_d)$$
$$f(x) = f(x_1, \cdots, x_d) = f_1(x_1) \cdot f_2(x_2) \cdot \cdots \cdot f_d(x_d)$$

其中F_i是累积分布函数，f_i是随机变量X_i的概率质量函数或概率密度函数。

1.4.3. 随机抽样和统计量

随机变量X的概率质量函数或概率密度函数可能遵循某种已知的形式，可也能是未知的（数据分析中经常出现这种情况）。当概率函数未知的时候，根据所给数据的特点，假设其服从某种已知分布可能会有好处。然后，即使在这种假设情况下，分布的参数依然是未知的。因此，通常需要根据数据来估计参数或者是整个分布。

在统计学中，总体（population）通常用于表示所研究的所有实体的集合。通常我们对整个总体的特定特征或是参数感兴趣（比如美国所有计算机专业学生的平均年龄）。然而，检视整个总体有时候不可行或代价太高。因此，通过对总体进行随机抽样、针对抽样到的样本计算合适的统计量来对参数进行推断，从而对总体的真实参数作出近似估计。

1. 一元样本

给定一个随机变量X，对X的大小为n的随机抽样样本定义为一组n的个独立同分布（independent and identically distributed，IID）的随机变量$S_1, S_2, \cdots, S_n$，即所有S_i之间是相互独立的，概率质量或概率密度函数与X是一样的。

若将X当作一个随机变量，则可以将X的每一个观察值$x_i (1 \leqslant i \leqslant n)$本身当作一个恒等随机变量，并且每一个观察到的数据都可以假设为从X中随机抽样到的一个样本。因此，所有的x_i都是相互独立的，而且与X是同分布的。根据公式（1.11）,联合概率函数可以给出：

$$f(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_X(x_i)$$

其中f_X是X的概率质量或概率密度函数。

2. 多元样本

对于多元参数估计，n个数据点$x_i (1 \leqslant i \leqslant n)$构成一个从向量随机变量$X = (X_1, X_2, \cdots, X_d)$中取得的$d$维的多元随机样本。假定$x_i$是独立同分布的，且它们的联合分布如下所示：

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f_X(X_i)$$

其中f_X是X的概率质量函数或概率密度函数。

估计一个多元联合概率分布的参数通常比较困难而且很耗费计算资源。为了简化，一种常见的假设是d个属性$X_1, X_2, \cdots, X_d$在统计上是独立的。然而，没有假设它们是同分布的，因此那几乎从不会发生。在这种属性独立假设下，公式（1.12）可以重新定义：

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \prod_{j=1}^{d} f_{X_j}(x_{ij})$$

3. 统计量

可以通过一个合适的样本统计量来估计总体的一个参数，统计量通常定义为样本的函数。令$\{S_i\}_{i=1}^m$代表从多元随机变量X中取出的m个随机样本。统计量$\hat{\theta}$是一个函数：$\hat{\theta}:(S_1, S_2, \cdots, S_m) \to \mathbb{R}$。该统计量是对总体参数$\theta$的估计。$\hat{\theta}$本身也是一个随机变量。若使用一个统计量的值来估计一个总体参数，则该值称作对参数的点估计，该统计量被称作对参数的一个估计量。

例1.11（样本均值）考虑鸢尾花数据集中的属性--萼片长度（X_1），相关值在表1-2中。假设X_1的均值未知。可以假定观测到的值$\{x_i\}_{i=1}^n$构成一个从X_1得到的随机样本。

样本均值是一个统计量，定义为

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

输入表1-2的数值，可以得到：

$$\hat{\mu} = \frac{1}{150}(5.9 + 6.9 + \cdots + 7.7 + 5.1) = \frac{876.5}{150} = 5.84$$

$\hat{\mu} = 5.84$是对未知总体参数μ（随机变量X_1的真实平均值）的点估计。

1.5 数据挖掘

数据挖掘是由一系列能够帮助从大数据中获得洞见和知识的核心算法构成的。是一门融合了数据库系统、统计学、机器学习和模式识别等领域的交叉学科。数据挖掘是知识发现过程中的一环，知识发现包括数据提取、数据清洗、数据融合、数据简化和特征构建等预处理过程。

1.5.1 探索性数据分析

探索性数据分析是分别或者一起研究数据的数值型属性和类别型属性，希望以统计学提供的集中度、离散度等信息来提取数据样本的关键特征。

1.5.2 频繁模式挖掘

频繁模式挖掘是指从巨大又复杂的数据集中提取富含信息的有用模式。模式由重复出现的属性值的集合（项集）或者复杂的序列集合（考虑显式的先后位置和时序关系）或图的集合（考虑点之间的任意关系）构成。核心目标是发现在数据中隐藏的趋势和行为，以更好地理解数据点和属性之间的关系。

1.5.3 聚类

聚类是指将数据点划分为若干簇，并使得簇内的点尽可能相似，而簇间点尽可能区分开的任务。根据数据和所有的簇的特征，有不同的聚类方法。基于代表（representative-based）、层次式的（hierarchical）、基于密度的（density-based）、基于图的（graph-based）和谱聚类。

1.5.4 分类

分类是为一个未添加标注的数据点预测其标签或类。分类器就是一个模型或者函数M，对于给定的输入x，能够预测其类标签$\hat{y}$，即$\hat{y} = M(x)$，$\hat{y} \in \{c_1, c_2, \cdots, c_k\}$，其中每个$c_i$代表一个类标签（一个类属性值）。为建立这样的模型，需要一组带有正确类标签的点，成为训练集。学到模型M后，对于任意新给定的点都可以自动预测其类。主要包括决策树、概率型分类器、支持向量机等。

习题

1. 说明公式（1.5）中的居中数据矩阵Z的均值为0.

$$Z = D - 1 \cdot \mu^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} - \begin{pmatrix} \mu^T \\ \mu^T \\ \vdots \\ \mu^T \end{pmatrix} = \begin{pmatrix} x_1^T - \mu^T \\ x_2^T - \mu^T \\ \vdots \\ x_n^T - \mu^T \end{pmatrix} = \begin{pmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_n^T \end{pmatrix}$$

证明：

$$Z = D - 1 \cdot \mu^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} - \begin{pmatrix} \mu^T \\ \mu^T \\ \vdots \\ \mu^T \end{pmatrix} = \begin{pmatrix} x_1^T - \mu^T \\ x_2^T - \mu^T \\ \vdots \\ x_n^T - \mu^T \end{pmatrix}$$

$$mean(Z) = \frac{1}{n}\sum_{i=1}^{n}(x_i^T - \mu^T) = \frac{1}{n}\sum_{i=1}^{n}(x_i^T) - \frac{1}{n}\sum_{i=1}^{n}(\mu^T) = \mu - \mu = 0$$

```
import numpy as np;
import pandas as pd;
x = np.random.rand(10)*10;
mu = x.mean()
(x-mu).mean()
x1 = pd.DataFrame({'label':x});
df = x1[x1['label']>6]
```

2. 证明对于公式（1.2）中的L_p的距离，有：

$$\delta(x,y) = \lim_{p\to\infty}\delta_p(x,y) = \max_{i=1}^{d}\{|x_i - y_i|\}, \quad \mathbf{x},\mathbf{y} \in \mathbb{R}^d$$

证明：公式（1.2）

$$\delta_p(a,b) = \| a-b \|_p \ (p=2)$$

$$\begin{aligned}\delta_\infty(x,y) &= \lim_{p\to\infty}\delta_p(x,y) = \lim_{p\to\infty}\| x-y\|_p \\ &= \lim_{p\to\infty}\sqrt[p]{\sum_i^n (x_i-y_i)^p} \\ &= \lim_{p\to\infty}(x-y)_{max}\cdot\left((\frac{\|x_1-y_1\|}{(x-y)_{max}})^p + (\frac{\|x_2-y_2\|}{(x-y)_{max}})^p + \cdots + (\frac{\|x_n-y_n\|}{(x-y)_{max}})^p\right)^{\frac{1}{p}} \\ &= (x-y)_{max}\cdot\lim_{p\to\infty}\left((\frac{\|x_1-y_1\|}{(x-y)_{max}})^p + (\frac{\|x_2-y_2\|}{(x-y)_{max}})^p + \cdots + (\frac{\|x_n-y_n\|}{(x-y)_{max}})^p\right)^{\frac{1}{p}}\end{aligned}$$

$\because 1 \leqslant \sum_{i=1}^{n}\left(\frac{\|x_i-y_i\|}{(x-y)_{max}}\right)^p \leqslant n$，$\therefore \lim_{p\to\infty}1^{\frac{1}{p}} = 1$，$\lim_{p\to\infty}n^{\frac{1}{p}} = \lim_{p\to\infty}e^{\frac{ln(n)}{p}} = 1$。

$\because$ 夹逼原理（squeeze theorem），

$$\lim_{p\to\infty}\left((\frac{\|x_1-y_1\|}{(x-y)_{max}})^p + (\frac{\|x_2-y_2\|}{(x-y)_{max}})^p + \cdots + (\frac{\|x_n-y_n\|}{(x-y)_{max}})^p\right)^{\frac{1}{p}} = 1$$

从而$\lim_{p\to\infty}\| (x-y) \|_p = (x-y)_{max}$。

3. 登录http://www.sse.com.cn/market/sseindex/diclosure/c/c_20161230_4223361.shtml,下载附件4（中国战略新兴产业成份指数样本股列表）数据，根据样本股“证券中文简称”，构建样本股的特征数据（节能环保、新一代信息技术产业、生物产业、高端装备制造、新能源产业、新材料产业、新能源汽车、数字创意产业、高技术服务业），总共有100个指标股票，孙婵子给每个同学分配两支股票，完成后由孙婵子统一汇总。下面是样例数据。

股票简称	2019年申请的专利	2019年研发投入	2019年收入	2019年市值	所在地	所属产业
海康威视	1339	54.84亿	576.58亿	3000亿	杭州	新一代信息技术产业