数据科学基础

Foundations of Data Science

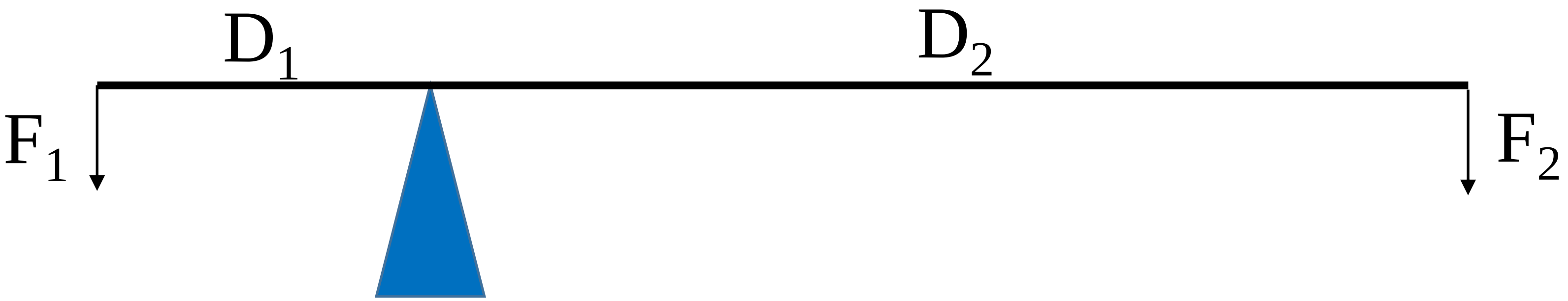
3.2 矩：数据动力学基础

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

矩（Moment）

矩：是物理学中的一个丰富概念，涉及质量、形状、空间、运动等各个方面。

力矩



- 力矩计算：$F_1 D_1$
- 反向力矩平衡：$F_1 D_1 = F_2 D_2$
- 同向力矩累加：$F_1 D_1 + F_2 D_2 + \ldots\ldots$

力　　　　权重

数据的原点矩

给定一批数据$x_1, x_2, \ldots x_n$，其k阶原点矩A_k (k=1,2,…) 定义为：

$$A_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

- 一阶矩是算术平均值，代表数据集的“重心”
- 二阶矩代表数据集的转动惯量
- 原点矩代表了数据集跟重量相关的动力度量

数据的中心矩

给定一批数据$x_1, x_2, \ldots x_n$，其k阶中心矩B_k (k=1,2,…) 定义为：

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - A_1)^k$$

- 一阶中心矩B_1为零、二阶中心矩B_2为方差
- 三阶中心矩刻画偏度、四阶中心矩刻画峰度
- 中心矩代表了数据集跟几何相关的动力度量

矩的计算

e	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
X	3	2	2	2	1	1	1	0

$$\mathrm{A}_1 = \frac{1}{8}(3 + 2 + \cdots + 0) = 1.5$$

$$\mathrm{A}_2 = \frac{1}{8}(3^2 + 2^2 + \cdots + 0^2) = 3$$

$$\mathrm{B}_2 = \frac{1}{8}\left((3 - 1.5)^2 + \cdots + (0 - 1.5)^2\right) = 0.75$$

数学期望

离散随机变量数学期望

设离散型随机变量X的概率分布为$P(X=x_k)=p_k$， $k=1,2,\cdots,n,\cdots$,

如果无穷级数$\sum_{k=1}^{\infty}x_{\mathrm{k}}p_k$绝对收敛，则称此级数的和为随机变量$X$的数学期望, 记作$E(X)$或$EX$, 即

$$E(X)=\sum_{k=1}^{\infty}x_kp_k$$

x_k数据，p_k概率/权重

数学期望

连续随机变量数学期望

设连续型随机变量X的概率密度为$f(x)$, 如果广义积分$\int_{-\infty}^{+\infty} xf(x)dx$

绝对收敛, 则称此积分值为随机变量X的数学期望, 记作 $E(X)$或EX,

即

$$E(X)=\int_{-\infty}^{+\infty} xf(x)dx$$

数学期望的性质

- 对于常数C, 有$E(C)=C$
- 对于常数C及随机变量X, 有$E(CX)=CE(X)$
- 设X和Y为两个随机变量, 则$E(X+Y)=E(X)+E(Y)$
- 设随机变量X和Y独立, 则$E(XY)=E(X)E(Y)$

方差

设X是一随机变量，如果数学期望$E[X-E(X)]^2$存在，则称之为X的方差，记作$Var(X)$或$D(X)$或。$\sqrt{Var(X)}$称为X的标准差, 记为$\sigma(X)$。

方差的计算式: $Var(X) = E(X^2) - (EX)^2$

方差的性质

- 设C是常数，则$Var(C)=0$;
- 设X是随机变量, C是常数, 则

  $Var(X+C)=Var(X)$，$Var(CX)=C^2Var(X)$.
- 设随机变量X与Y相互独立, 则

  $Var(X\pm Y)=Var(X)+Var(Y)$.

数学期望与方差的计算

X	3	2	1	0
P	1/8	3/8	3/8	1/8

$$E(X)=\sum_{k=1}^{\infty} x_k p_k = 3*\frac{1}{8}+2*\frac{3}{8}+1*\frac{3}{8}+0*\frac{1}{8}=1.5$$

$$E(X^2)=\sum_{k=1}^{\infty} {x_k}^2 p_k = 3^2*\frac{1}{8}+2^2*\frac{3}{8}+1^2*\frac{3}{8}+0^2*\frac{1}{8}=3$$

$$D(X)=E(X^2)-(EX)2=3-1.5^2=0.75$$

数学期望与方差的计算

$f(x)=\frac{x}{2}, 0<x<2;\ f(x)=0,$ 其他

$E(X)=\int_{-\infty}^{+\infty} xf(x)dx=\int_0^2 x\frac{x}{2}dx=\frac{x^3}{6}|_0^2=\frac{4}{3}$

$E(X^2)=\int_{-\infty}^{+\infty} x^2f(x)dx=\int_0^2 x^2\frac{x}{2}dx=\frac{x^4}{8}|_0^2=2$

$D(X)=E(X^2)-(EX)2=2-(\frac{4}{3})^2=\frac{2}{9}$

概率分布的矩-数字特征

随机变量X的矩定义如下：

- 若$E(X^k)$存在$(k=1,2,\cdots)$,
  称$E(X^k)$为X的k阶原点矩或k阶矩, 记为μ_k
- 若$E((X-EX)^k)$存在$(k=1,2,\cdots)$,
  称$E((X-EX)^k)$为X的k阶中点矩, 记为υ_k

中心矩的原点矩表示

$$\upsilon_k=\sum_{i=0}^{k} C_{\mathrm{k}}^{\mathrm{i}}\, \mu_i\,(-\mu_1)^{(k-i)}$$

$$\nu_2=\mu_2-\mu_1{}^2$$

$$\nu_3=\mu_3-3\mu_2\mu_1+2\mu_1{}^3$$

$$\nu 4=\mu_4-4\mu_3\mu_1+6\mu_2\mu_1{}^2-3\mu_1{}^4$$