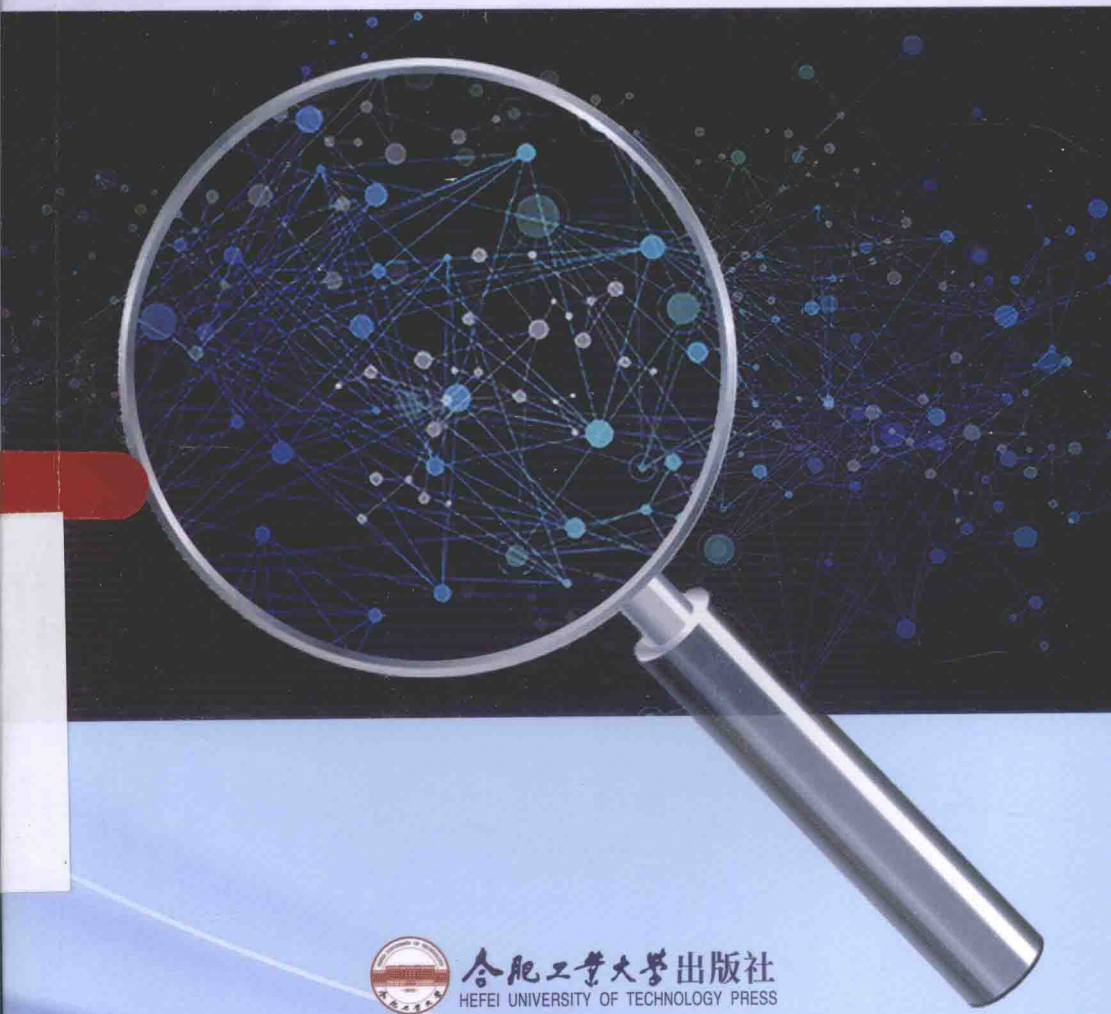
SHUJU FENXI JICHU JI MOXING

数据分析

基础及模型 >>>>>>>>

管 涛◎编著



合肥工业大学出版社
HEFEI UNIVERSITY OF TECHNOLOGY PRESS

管涛著

数据分析基础及模型

合肥工业大学出版社

内容提要

全书共分7章，主要内容包括：第1章介绍了相关的算子理论基本定义和概念；第2章介绍了数据分析的数理统计基础，包括常用统计分布、分布函数的变换、收敛性、MCMC采样、随机逼近；第3章引入了数据分析中常见的矩阵计算和分解的相关理论和方法；第4章阐述了一些数据分类模型；第5章介绍了流行的降维理论和方法；第6章论述了聚类分析相关理论和算法；第7章讨论了在线学习的理论和方法。本书亦注重算法的Matlab实现，书中包含了一些算法的Matlab程序。

本书可作为高等院校计算机科学与技术、电子商务、云计算、数据挖掘和分析、信息检索等专业的教学参考书，也可作为大数据领域工程人员的参考用书。

图书在版编目(CIP)数据

数据分析基础及模型/管涛著. —合肥：合肥工业大学出版社，2015.7
ISBN 978-7-5650-2317-0

Ⅰ.①数… Ⅱ.①管… Ⅲ.①统计数据—统计分析 Ⅳ.①O212.1

中国版本图书馆CIP数据核字(2015)第161754号

数据分析基础及模型

管 涛 著　　责任编辑 马成勋

出 版	合肥工业大学出版社	版 次	2015年7月第1版
地 址	合肥市屯溪路193号	印 次	2015年7月第1次印刷
邮 编	230009	开 本	710毫米×1000毫米 1/16
电 话	理工编辑部：0551—62903200	印 张	13.75
	市场营销中心：0551—62903198	字 数	200千字
网 址	www.hfutpress.com.cn	印 刷	安徽联众印刷有限公司
E-mail	hfutpress@163.com	发 行	全国新华书店

ISBN 978-7-5650-2317-0　　定价：28.00元

QIAN YAN 前言

大数据是随着信息技术的发展而诞生的一个科学领域，在学术界和工程界中得到了广泛的研究和应用，为社会的发展带来了收益、人们的生活带来了便利。大数据的研究得到了世界各国政府的重视，投入了不少的物力和财力。目前，不少企业开展着许多大数据的应用，例如，百度、google 的文本、图像搜索，淘宝、京东等电商的推荐系统，facebook、linkedin 等社交网站的用户连接模型，电子政务专家决策和咨询系统。大数据应用的本质在于面向用户的已知的或潜在的需求，建立恰当的数据分析和处理模型，得到结论并提供有效的服务或者参考，增加企业的利润。然而，在大数据上构建有效的数据分析模型并非易事。通常情况下，有效信息是隐藏在大量错误、冗余、异质、多源、高维、多尺度的复杂数据之中，数据之间也存在着众多的联系，那么，在这种情况下，我们该如何应对呢？围绕着这种需求，本书重点阐述了与复杂数据处理紧密相关的数学领域和原理，包括算子理论、统计分析方法、矩阵分析和计算方面的内容，同时，也介绍了当今流行的一些数据推断、分类、聚类模型和方法，包括贝叶斯推断、高斯混合模型、数据降维方法、自组织映射、竞争学习、在线学习。

本书与数据挖掘、机器学习、数学理论领域在分析问题的角度上有一定的区别。从内容上看，数据挖掘和机器学习书籍更侧重工程上的实践和应用，很少揭示算法蕴含的数学原理。而数学专业书籍多从理论上定义和证明，较少结合实际问题进行讲解，与应用脱节较大，计算机专业的研究人员入门较难。本书不探讨深刻的数学理论本身，而是结合数据分析需求和作者的工作，对相关的数学领域加以阐述和介绍，是对计算机领域的数据分析研究的一个有效的补充。

本书的撰写和出版得到了国家自然科学基金(No. 41171341)、河南省教育厅科学技术研究重点项目科技攻关计划(No. 14A520060)、郑州市普通科技攻关计划项目(No. 20130783)的支持，在此表示感谢，同时感谢家人的支持。

作　者

2015 年 3 月于郑州

MU LU 目录

符 号 表

符号	含义
Ω	概率空间
$\mathbb{R}^n$	n 维欧式空间
H	Hilbert 空间
$\emptyset$	空集
ρ	测度
Θ	参数空间
X	随机变量或样本或矩阵
$A,B,C,\cdots$	集合或矩阵
$\mathrm{tr}(A)$	矩阵 A 的迹
A^*	矩阵 A 的共轭转置
σ	矩阵奇异值
λ	矩阵特征值
$\lvert\cdot\rvert$	集合的基或元素的数目
$\in$	属于
$\subset$	包含于
$\ll$	远小于
$\propto$	正比于
∂	求偏导或微分

$\mathrm{d}f$	对函数 f 求微分
∇	求梯度
Δ	Laplace 算子
$N(0,1)$	标准正态分布
Γ	Gamma 分布，伽玛分布
χ^2	卡方分布
μ	均值
θ	参数
$\sum$	方差
$K(\cdot,\cdot)$	核函数
$\wedge$	逻辑合取运算
$\vee$	逻辑析取运算
$\neg$	逻辑非
$\forall$	任意
$\exists$	存在
∞	无穷大
$\sim$	服从(分布)
$\rightarrow$	收敛于或映射
$\xrightarrow{F}$	依分布收敛于
$\xrightarrow{a.s.}$	几乎必然收敛于
ϕ	非线性映射
Φ	正交函数集

第1章 算子理论基础

函数定义为数集到数集的映射，算子推广了函数的概念，实现空间到空间的映射。一类特殊的算子称为泛函，从空间映射到数集。在人工智能和机器学习领域，许多方法使用样本学习和构造泛化的(非)线性逼近函数，如神经网络、支撑向量机、高斯混合模型、集成学习、回归和主成分分析等，完成数据的分类、判别、降维、回归和因子分析等。在模型结构或者模型复杂度确定的条件下，学习一个函数其实是根据样本确定最优的参数值。通常的做法为首先定义误差或者目标泛函，通过样本学习最小化该泛函，求得最优参数。

算子理论扩展了泛函的概念，在学习理论中占据了重要的地位，但目前的应用基础领域很少有这方面的文献。对数据的操作实际上就是对矩阵的运算，变换矩阵作为算子可以对矩阵进行旋转、投影、对角化。变换的结果由算子的性质决定，因此，研究算子的连续性、可微性、有界性、紧性、正定性、象空间、零空间等具有普遍意义。

1.1 基本概念和定义

1.1.1 空间

在数学分析中，空间是一种抽象的概念，定义为具有某种特殊属性的对象的集合。空间的种类很多，如欧式空间、概率空间、内积空间、拓扑空间、仿射空间等等。空间可以具有度量，由范数定义的度量空间称为赋范空间。空间在度量的定义下可以是完备的或者不完备的，如 Hilbert 空间在内积度量的定义下是完备的。完备的赋范空间称为 Banach 空间，

Hilbert 空间属于 Banach 空间。Hilbert 空间中的度量通过内积定义，其中包含了许多光滑与非光滑函数。它的一个光滑子空间为再生核 Hilbert 空间(Reproduce Kernel Hilbert Space，RKHS)，可以通过一个正定核函数生成。在工程中，通常考虑的空间具有三个基本要素，包括对象、结构和测度。对象指要研究的实体，可以是向量、矩阵、函数、子空间、算子等，结构指空间的构成方式，测度确定空间集合的大小。依据不同的研究对象，数据分析中的空间包括向量空间(如欧式空间)、泛函空间(如再生核 Hilbert 空间)。向量空间可以通过正交化的有限或无限个基向量张成，正交性通过向量内积定义。泛函空间可以通过正交基函数构造，如多项式基函数、正交三角函数基。函数的正交性定义为指定区域内两个函数乘积的积分。拓扑空间是一个集合 X 和其上定义的拓扑结构 τ 组成的二元组(X,τ)。依不同的作用，测度有多种定义方式，如勒贝格测度、概率测度等。概率测度定义了概率空间，用来度量空间中事件集的发生概率。

范数是定义在空间中对象上的一种度量，有向量范数、矩阵范数、算子范数等。给定实数域 $\mathbb{R}$，空间 $\mathbb{R}^n$ 中的向量 x 有如下形式的范数：

1 范数：$\| x \|_1 = \sum_{i=1}^{n} | x_i |$；

2 范数：$\| x \|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$；

∞ 范数：$\| x \|_\infty = \max\{ | x_1 |, | x_2 |, \cdots, | x_n | \}$；

更一般地，可以定义 p 范数，

p 范数：$\| x \|_p = (\sum_{i=1}^{n} | x_i |^p)^{\frac{1}{p}}$。

空间 $\mathbb{R}^{m\times n}$ 的元素为 $m \times n$ 阶的矩阵。若 $A \in \mathbb{R}^{m\times n}$，则 A 的范数有以下几种形式：

1 范数：$\| A \|_1 = \max\limits_{1\leqslant j\leqslant n} \sum\limits_{i=1}^{m} | a_{ij} |$，即列和最大为范数；

∞ 范数：$\| A \|_\infty = \max\limits_{1\leqslant i\leqslant n} \sum\limits_{j=1}^{n} | a_{ij} |$，即行和最大为范数；

p 范数：$\| A \|_p = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} | a_{ij} |^p \right)^{1/p}$，$p \in \mathbb{N}$，$p > 2$；

F 范数：$\| A \|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} | a_{ij} |^2 \right)^{1/2} = \sqrt{\mathrm{trace}(A^* A)}$，其中 A^* 为 A 的共轭矩阵。

矩阵范数具有如下一些性质：对于 $A, B \in \mathbb{R}^{m \times n}$，$C, D \in \mathbb{R}^{m \times n}$，$\alpha \in \mathbb{R}$，有

$\| A \| \geqslant 0$，仅当 $A = 0$ 时等号成立；

$\| \alpha A \| = | \alpha | \, \| A \|$；

$\| A + B \| = \| A \| + \| B \|$；

$\| CD \| \leqslant \| C \| \, \| D \|$；

$\| C \| = \| C^* \|$，C^* 为 C 的共轭矩阵。

例 1.1 L_p 空间及范数

$L^P (1 \leqslant p < \infty)$ 空间是泛函分析中一类特殊的 Banach 空间，要求可测函数 $f: \Omega \to \mathbb{R}$ 在测度空间 $(\Omega, \sum, \mu)$ 中具有有限 p－可积性。L^P 范数定义为：

$$\| f \|_p = \left(\int_{\Omega} | f |^p \mathrm{d}\mu \right)^{1/p} < \infty,$$

特别地，L^2 空间（或记为 L_2）在数据分析模型中经常用到。p－范数扩展到有限长度离散向量产生了 l^p 空间，对于向量 x 具有如下的定义：

$$\| x \|_p = \left(\sum_{i=1}^{n} | x_i |^p \right)^{1/p} < \infty$$

1.1.2 紧集与算子

紧集是空间中一类特殊的集合，如欧式空间 $\mathbb{R}^n$ 中的有界闭集。为什么需要考虑紧集？在无限维空间中，许多良好的数学性质未必成立，从而难以研究点列的极限行为、空间的性质。如果能通过有限的结构来探讨无限情况下的性质，那么问题就容易解决，而具有这样有限结构的集合即为紧集。紧集提供了一种利用有限结构研究无限结构的方法。紧集的定

义可以从不同的角度描述，如序列、拓扑结构。

拓扑空间定义方法：紧集是拓扑空间内的一类特殊点集，它们的任何开覆盖都有有限子覆盖。

序列定义方法：紧集的任意序列有收敛子列，且子列的极限点属于该集合。下面是紧集所具有的一些常用性质：

(1)在有限维空间紧集即为有界闭集；

(2)实数空间中的紧集具有最大和最小元素；

(3)紧集在连续函数下的像仍是紧集；

(4)定义在紧集上的连续实值函数有界，且有最大和最小值。

紧集是一种良好的结构，而紧算子如何定义，有何用途呢？算子表示了数学物理中的一种功能，是一种作用于对象之上的运算。算子可以表现为函数、矩阵或者其他数学、物理变换。不同的矩阵，产生的作用不同，如投影、旋转、平移等。算子有不同的种类，如线性算子、共轭算子、拉普拉斯算子、积分算子等。下面介绍一些可能用到的算子概念。

紧算子通过集合定义，将有界集映射为相对紧集(relatively compact set)[Conway,1990]。给定算子 $T:X\to Y$，X，Y 为赋范向量空间，如果对于 X 的任意有界序列 $\{x_n\}_{n=1}^{\infty}$，都存在子序列 $\{x_{n_k}\}_{k=1}^{\infty}$ 使得 $\{T(x_{n_k})\}_{k=1}^{\infty}$ 在 Y 中有界，称 T 为紧的。

拉普拉斯算子(Laplacian)L 是一种微分算子，记为 Δ 或 ∇^2。在 n 维欧氏空间，L 定义为如下的二阶微分算子：

$$Lf=\Delta f=\nabla^2 f=\nabla\cdot\nabla f,$$

其中：f 为二阶可微函数。拉普拉斯算子在谱聚类算法中多有应用。

在物理学中，$f=f(x,y,z,t)$ 表示关于空间和时间的温度函数，空间的热传导方程表示为：

$$\frac{\partial f}{\partial t}=k\left(\frac{\partial^2 f}{\partial x^2}+\frac{\partial^2 f}{\partial y^2}+\frac{\partial^2 f}{\partial z^2}\right)=k\,\nabla^2 f,$$

或者

$$\frac{\partial f}{\partial t}=k\Delta f$$

式中∇^2或Δ表示Laplace算子，f表示热量在空间的分布，k为常数。该方程也称为扩散方程，是Fokker-Plank方程的原型，描述了粒子或者热量在空间随时间变化的分布。该方程的解析解表达为热核函数$G(x,t)$（即高斯分布）与f的卷积的形式，即对$x \in \mathbb{R}^3$，

$$H(f)(x) = G(t,x) * f = \int G(t,x,\tau) f(\tau) \mathrm{d}\tau.$$

积分算子T. 在再生核Hilbert空间(RKHS)，积分算子定义如下：

$$T(f)(x) = \int_{\Omega} K(x,y) f(y) \mathrm{d}\mu(y)$$

式中：Ω表示积分域；$K(\cdot,\cdot)$核函数；μ表示测度，如概率分布P。

在RKHS中，积分算子特征系统与谱聚类、核函数具有紧密的联系。

迹类算子. Hilbert空间中的紧对称算子T称为迹类的[Weidman, 1990]，如果它的特征值$\{\lambda_n\}$满足$\sum_{n=1}^{\infty} |\lambda_n| < \infty$。

算子范数. 如果X和Y是Banach空间，T是$X \to Y$的线性算子，$\|T\|$定义为：$\|T\| = sup\{\|Tx\| : \|x\| \leqslant 1\}$。

算子的迹定义为：

$$tr(T) = \sum_{n=1}^{\infty} (Te_n, e_n),$$

其中：$\{e_n\}$是H中标准正交集。

算子的迹范数

$$\|T\|_{tr} = \sum_{n=1}^{\infty} |\lambda_n|.$$

1.1.3 正交函数基

在泛函空间中，两个函数之间的正交性通过内积来定义。在离散空间中，函数在某范围内的离散值构成了向量，正交性通过内积表示。因此，一个定义域内函数之间的正交性也可以通过内积来定义，即两个函数在指定域内的乘积的积分。当该积分为零时，表示两个函数正交，否则，相关。给定一组函数，若对于任意两个不同的函数，内积为零；而对于相

同的函数，内积为常数，则称该函数集满足正交性。进而，若在该集合之外不存在函数与集合内的所有函数正交，则该集合构成了完备正交函数基。完备正交函数基可能具有有限或者无限个元素。就有限情况而言，形式化定义可以描述如下，给定定义域 Ω、测度 μ，函数集 $\{f_i \mid i \leqslant N\}$ 在 Ω 上的正交性定义为：

$$\int_{\Omega} f_i \circ f_j \mathrm{d}\mu = \begin{cases} 0, & i \neq j \\ k, & i = j \end{cases}, \quad k \neq 0$$

无限情况下具有类似的定义形式。下面介绍一些常见的正交函数基。

例 1.2 复指数函数基

$\Phi = \{\mathrm{e}^{\mathrm{j}n\omega_0 t} \mid n = 0, \pm 1, \pm 2, \cdots\}$ 为闭区间 $[0, T]$ 上完备正交集，其中：$\omega_0 = 2\pi/T$ 为基频。

复指数函数基已经应用于周期信号的傅利叶级数展开，分析信号的频率成分。

例 1.3 勒让德(Legendre) 多项式

$$\Phi = \{p_n(x) \mid n \geqslant 0\}, \text{其中 } P_n(x) = \frac{1}{n!\ 2^n}\left(\frac{\mathrm{d}}{\mathrm{d}x}\right)^n (x^2 - 1)^n,$$

构成了区间 $[-1, 1]$ 上的正交多项式函数集。函数 $f(x)$ 的勒让德多项式级数形式为：

$$f(x) = \sum_{n=0}^{\infty} c_n P_n(x),$$

其中：$c_n = \dfrac{2n+1}{2}\displaystyle\int_{-1}^{1} f(x) p_n(x) \mathrm{d}x$。

例 1.4 切比雪夫多项式

n 阶切比雪夫多项式 $T_n(x)$ 定义为：

$$T_n(x) = \begin{cases} \cos(n\arccos x) & |x| \leqslant 1 \\ \mathrm{ch}(n\,\mathrm{arcch}x) & |x| > 1 \end{cases}$$

特殊情况下，

$$T_0(x)=1,$$

$$T_1(x)=x,$$

$$T_2(x)=2x^2-1,$$

……

$$T_n(x)=2xT_{n-1}(x)-T_{n-2}(x),$$

在信号处理中，切比雪夫多项式广泛应用于滤波器设计。

1.2 常用不等式

不等式是算法理论的基础，在概率论、矩阵分析、泛函分析、数值优化领域应用广泛。在机器学习领域，不等式发挥着重要的作用。具体来讲，不等式在算法领域的作用体现在两个方面：第一、算法的理论验证。不论是经典算法，还是新构造的算法，人们都需要证实算法的可靠性，保证算法在数学上不存在原理性错误。同时，还要通过推导，论证算法的收敛性、误差分析，计算算法的收敛速度，以保证算法的有效性。第二、算法往往具有一些重要的性质，推导和衍生这些性质需要利用不等式简化。比如，采用不等式估算算法的计算复杂性、空间复杂性。在不同个数学分支中，不等式有很多种类。常见不等式在不同的空间（如欧式空间、概率空间、函数空间）表述形式有一定的差异，但是有些公式在本质上是一致的。下面介绍五个常用的不等式，前三个是关于概率的不等式，后两者为关于期望的不等式。

1.2.1 Markov 不等式

Markov 不等式是由俄国数学家马尔可夫提出的，描述了非负随机变量大于某一正数的概率的上界[Stein & Shakarchi, 2005]。假设 X 为一非负随机变量，$E(X)$ 存在，则对于任意的 $\varepsilon>0$，

$$P(X>\varepsilon)\leqslant\frac{E(X)}{\varepsilon} \tag{1-1}$$

证明：由于 X 为非负随机变量，则

$$E(X)=\int_0^{\infty} xf(x)\mathrm{d}x=\int_0^{\varepsilon} xf(x)\mathrm{d}x+\int_{\varepsilon}^{+\infty} xf(x)\mathrm{d}x$$

$$\geqslant \varepsilon\int_{\varepsilon}^{+\infty} xf(x)\mathrm{d}x=\varepsilon P(X>\varepsilon)$$

因此，$P(X>\varepsilon)\leqslant\dfrac{E(X)}{\varepsilon}$

1.2.2 Chebyshev 不等式

Chebyshev 不等式适用于所有的分布形式，描述了随机变量的取值超出某一值的比例。假设随机变量 X 的均值和方差均存在，令 $\mu=E(X)$，$\sigma^2=V(X)$，则对于任意的 $\varepsilon>0$，Chebyshev 不等式定义为[现代应用数学手册编委会，2000]：

$$P(|X-\mu|\geqslant\varepsilon)\leqslant\frac{\sigma^2}{\varepsilon^2},\tag{1-2}$$

或者

$$P(|Z|\geqslant\varepsilon)\leqslant\frac{1}{\varepsilon^2},$$

其中：$Z=\dfrac{X-\mu}{\sigma}$.

证明：仅考虑连续密度函数情况。设 $P(x)$ 的密度函数为 $f(x)$，则

$$P(|X-\mu|\geqslant\varepsilon)=\int_{|x-\mu|\geqslant z} p(x)\mathrm{d}x=\int_{\mu+z}^{\infty} p(x)\mathrm{d}x+\int_{-\infty}^{\mu-z} p(x)\mathrm{d}x$$

$$=\int_{x-\mu\leqslant -z}\frac{|x-\mu|^2}{\varepsilon^2}p(x)\mathrm{d}x+\int_{x-\mu\geqslant z}\frac{|x-\mu|^2}{\varepsilon^2}p(x)\mathrm{d}x$$

$$\leqslant\int_{-\infty}^{\mu-z}\frac{|x-\mu|^2}{\varepsilon^2}p(x)\mathrm{d}x+\int_{\mu+z}^{\infty}\frac{|x-\mu|^2}{\varepsilon^2}p(x)\mathrm{d}x$$

$$+\int_{\mu-z}^{\mu+z}\frac{|x-\mu|^2}{\varepsilon^2}p(x)\mathrm{d}x$$

$$=\int_{-\infty}^{\infty}\frac{|x-\mu|^2}{\varepsilon^2}p(x)\mathrm{d}x=\frac{\sigma^2}{\varepsilon^2}$$

1.2.3 Hoeffding 不等式

Hoeffding 不等式描述了一组随机变量的均值偏离它的期望的程度