**第一章 机器学习介绍**

**1.4 数学基础 概率论与数理统计**

#### 随机变量及其概率分布

**1.随机变量及概率分布**

取值带有随机性的变量，严格地说是定义在样本空间上，取值于实数的函数称为随机变量，概率分布通常指分布函数或分布律

**2.累计分布函数（Cumulative Distribution Function）的概念与性质**

定义：

性质：(1) (2)单调不减

(3)右连续 (4)

**3.离散型随机变量的概率分布**

**4.连续型随机变量的概率密度**

概率密度函数（Probability Density Function）非负可积，且:

(1)

(2) (3)为的连续点，则:

分布函数

**举例**：(1) 将个不同的球放入个盒子中，求事件A={每个盒子至多有一个球}的概率。

(2) 假定《Python和数据科学》班有个同学，则至少有二人生日相同的概率是多少（不考虑闰年）？

如果，。

否则，

|  |  |
| --- | --- |
|  | 至少2人生日相同的概率 |
|  | 0.03 |
|  | 0.12 |
|  | 0.41 |
|  | 0.97 |
|  | 1.00 |

**墨菲定理：Anything that can go wrong will go wrong.**

如果事情有变坏的可能，不管这种可能性有多小，它总会发生。

假设一件事情发生的概率是，次独立实验，则这件事情发生的概率是。

飞机发生空难的概率为。

|  |  |  |
| --- | --- | --- |
|  | 无空难（） | 有空难（） |
|  | 99.5% | 0.5% |
|  | 95.1% | 4.9% |
|  | 60.7% | 39.3% |
|  | 0.67% | 99.33% |
|  | 0 | 1 |

“墨菲定律”、“帕金森定律”和“彼德原理”被一些人称为二十世纪西方文化三大发现。

帕金森定律（Parkinson's Law）是官僚主义或官僚主义现象的一种别称，也可称之为“官场病”、“组织麻痹病”或者“大企业病”，源于英国著名历史学家诺斯古德·帕金森1958年出版的《帕金森定律》一书的标题。帕金森定律常常被人们转载传诵，用来解释官场的形形色色。帕金森在书中阐述了机构人员膨胀的原因及后果：一个不称职的官员，可能有三条出路，第一是申请退职，把位子让给能干的人；第二是让一位能干的人来协助自己工作；第三是任用两个水平比自己更低的人当助手。这第一条路是万万走不得的，因为那样会丧失许多权利；第二条路也不能走，因为那个能干的人会成为自己的对手；看来只有第三条路最适宜。于是，两个平庸的助手分担了他的工作，他自己则高高在上发号施令，他们不会对自己的权利构成威胁。两个助手既然无能，他们就上行下效，再为自己找两个更加无能的助手。如此类推，就形成了一个机构臃肿，人浮于事，相互扯皮，效率低下的领导体系。 帕金森得出结论：在行政管理中，行政机构会像金字塔一样不断增多，行政人员会不断膨胀，每个人都很忙，但组织效率越来越低下。这条定律又被称为“金字塔上升”现象。

《彼得原理》劳伦斯·彼得的著作。为我们清楚地描绘了职业晋升的瓶颈问题，他指出，每个人在层级组织里都会得到晋升，直到不能胜任为止。换句话说，一个人，无论你有多大的聪明才智，也无论你如何努力进取，总会有一个你胜任不了的职位在等待着你，并且你一定会达到那个位置。这就是著名的彼得原理。（百度百科）

#### 随机变量的数字特征

**1.数学期望 (Expectation)**

离散型：；

连续型：

性质：

(1)

(2)

(3) 若和独立，则

(4)

**2.方差 (Variance)**：

**3.标准差 (Standard Deviation)**：，

**4.离散型：**

**5.连续型：**

性质：

(1)

(2)与相互独立，则

(3)

(4) 一般有

(5)

(6)

**6.协方差**

**7.相关系数**  ,阶原点矩 ; 阶中心矩

总体（Population）：所要考察对象的全体叫做总体。

样本（Sample）：从总体中所抽取的一部分个体叫做总体的一个样本。

参数（Parameter）：反应总体特点的数字特征。

统计量（Statistic）：反映样本特点的数字特征。

样本均值（Sample Mean）：

样本方差（Sample Variance）：

为的无偏估计，即。

所以，,

平均数（Average）：均值（Mean）、中位数（Median）、众数（Mode）

均值即算术平均数。

均值的缺点：易受数据极端值的影响，对于偏态分布的数据，均值的代表性较差。

中位数：一组数据按递增或递减顺序排列后中间位置上的代表值。在一组排好序数据中，数据数量为奇数，则中值为中间的那个数。如果数据数量为偶数，则中值为中间的那两个数值的平均值。

特点是不受数据极端值的影响。对于具有偏态分布的数据，中位数的代表性要比均值好。

众数：一组数据中出现次数最多的那个数。

优点：易于理解，不受极端值的影响。

缺点：具有不唯一性，对于一组数据可能有一个众数，也可能有两个或多个众数，也可能没有众数。

切尾均值(Trimmed Mean)：在一组数据中，去掉首位两端极端值后所计算的算术平均数。

按从小到大排列，则

优点：不受极端值的影响。

import numpy as np

from scipy import stats

data = np.array([4, 5, 1, 2, 7, 2, 6, 9, 3])

"""

计算均值（Mean）

"""

dt\_mean = np.mean(data)

print('Mean:', np.round(dt\_mean, 2))

"""

计算中位数（Median）

"""

dt\_median = np.median(data)

print('Median:', dt\_median)

"""

计算众数（Mode）

"""

dt\_mode = stats.mode(data)

print('Mode:', dt\_mode[0][0])

上述代码的输出结果是：

Mean: 4.33

Median: 4.0

Mode: 2

“”“

np.array()：创建一个多维数组；

np.zeros()：一次性创建全0数组；

np.ones()：一次性创建全1数组；

np.arange():

““”

np.zeros(10)

np.ones((3, 6))

np.arange(15)

#### 常见分布

(1) 0-1分布，伯努利分布（Bernoulli Distribution）: 。

证明：。

(2) 二项分布（Binomial Distribution）: ：。

证明：。

，相互独立均服从参数为的为0-1分布。

(3) 泊松分布（Poisson Distribution）: ：。

证明：。

指数函数泰勒级数展开式为

上式两边除以，得

令，有

因此，。

(4) 均匀分布（Uniform Distribution）：

(5) 指数分布（Exponential Distribution）:

证明：，。

所以，。

(6) 正态分布（Normal Distribution）: 。

**中心极限定理（Central Limit Theorem）**

设随机变量独立同分布（Independent Identical Distribution， IID），并且具有有限的数学期望和方差，则，即，其中 。

中心极限定理说明，在适当的条件下，大量相互独立同分布随机变量的均值经适当标准化后依分布收敛于标准正态分布。

**应用**：一个复杂系统由100个相互独立的元件组成，在系统运行时每个元件损坏的概率为0.1，为使系统正常工作，至少必须有85个元件工作，求系统的可靠度(正常工作的概率)。

用随机变量表示第个元件是否正常工作。如果第个元件正常工作，有，否则，。即服从0-1分布。，。

以表示100个元件中正常工作的元件数，即。由中心极限定理，

**1.5 Python介绍**

**Python的安装**

在Anaconda官网[https://www.anaconda.com/distribution/](https://www.anaconda.com/distribution/ )  上下载Python 3.7 version，然后安装。

Spyder是Anaconda集成的IDE（Integrated Development Environment, 集成开发环境）。

**1. Python的优缺点**

可用于数据分析的编程语言：Matlab、SAS（商业软件，价格昂贵）、R、Python（Free）

优点：1. 免费；2. 简单、易学；3. 解决“双语言”难题（一种语言用于对算法的研究、原型实现和测试，转换为另一种语言（Java、C++等）用于实际系统）。Python可适用于研究和原型实现，也适合搭建生产系统。

缺点：Python语言不足之处是的性能问题。Python程序运行的效率不如Java或C代码高。

**2 Python常用数据库**

NumPy

NumPy（[https://numpy.org/](http://www.baidu.com/link?url=NpnRbXk3hv4F4xkIsbuFIg5Oeff9sbDkO4acaDJn_AC&wd=&eqid=e9b987aa0008fb9c000000065d65fdd2)）是Numerical Python的简写，是Python数值计算的基石。

SciPy：科学计算库，依赖于NumPy，提供高效的数值计算。

pandas

提供了高级数据结构和函数

matplotlib

matplotlib（[www.**matplotlib**.org/](http://www.baidu.com/link?url=cvQjaKC7S4lVe-VxJCxZ2JvaafYiDopF10keW_1uI-ilU1eJyxWD9LFBHUFskENu&wd=&eqid=d3ab1ab80006ce02000000065d65fec9)）用于制图及其它二位数据可视化的Python库。

scikit-learn

scikit-learn（[https://**scikit-learn**.org/](http://www.baidu.com/link?url=Ql1gf2oRaaYTiSzlW0LPmogm93mTeTLr1h6ZJJqdnq9Nb20Q5z7rjTGTgUsL96ZN&wd=&eqid=bc2142f500072307000000065d65fe87)）是一个机器学习工具包。

statsmodels

一个统计分析包。

**3. 常用数据结构**

（1）列表（list）

列表的长度是可以变的，它所包含的内容也是可以修改的。

可以用中括号[ ]或者list函数来定义列表。

（2）字典（dict）

字典是拥有灵活尺寸的键值对集合，其中键和值都是Python对象。用大括号{ }是创建字典的一种方式，在字典中用逗号将键值对分离。

（3） 集合（set）

集合是一种无序且元素唯一的容器。可以认为集合是一个只有键没有值的字典。集合有两种创建方式：通过set函数或者用大括号{ }。

集合支持数学上的集合操作，例如并集、交集、差集等。

**4. 蒙特·卡罗方法（Monte Carlo method）**

蒙特·卡罗方法（Monte Carlo method），也称统计模拟方法，是二十世纪四十年代中期由于科学技术的发展和电子计算机的发明，而被提出的一种以概率统计理论为指导的一类非常重要的数值计算方法。(百度百科)

是使用随机数（或更常见的伪随机数）来解决很多计算问题的方法。

蒙特卡罗方法的三个主要步骤：

* 构造或描述概率过程；
* 实现从已知概率分布抽样；
* 建立各种估计量。

**5. Q-Q plot**

Q，Quantile，分位数，亦称分位点。

用于直观验证一组数据是否来自某个分布，或者验证某两组数据是否来自同一（族）分布。经常用于检验数据是否来自于正态分布。

检验一组数据（）是否来自分布函数为的某个分布的步骤如下：

1. 将数据从小到大排序：。的概率分别是：；
2. 由算出；
3. 画散点图；
4. 在图上添加一条直线做参考。看散点是否落在这条线的附近。

注：也有一些研究人员选择用来替换。

**示例**：已知，用蒙特卡罗方法验证。

1. 由计算机从分布 随机产生个数据这里，我们选取。
2. 计算。
3. 绘制Q-Q图（图1所示）。由Q-Q图可以看到所有的散点都在一条直线附近。所以，我们可以得出是成立的。

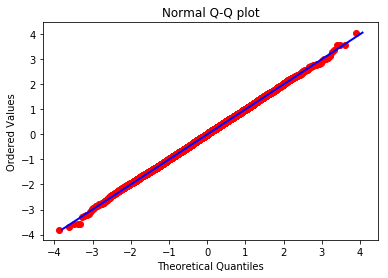


图1. Q-Q图

Python代码如下：

import numpy as np

from scipy import stats

from scipy.stats import norm

import matplotlib.pyplot as plt

n = 10000

mu = 1

sigma = 2

np.random.seed(123)

x = np.random.normal(mu, sigma, n)

"""

np.random.normal()： 正态分布

np.random.uniform()： 均匀分布

np.random.binomial()： 二项分布

np.random.exponential()： 指数分布

np.random.poisson()： 泊松分分布

"""

y = (x-mu)/sigma

y.sort()

prob = (np.arange(n)+1/2)/n

q = norm.ppf(prob, 0, 1)

count, bins, ignored = plt.hist(y, 300, density = True) #直方图

plt.show()

"""

直方图(Histogram)，又称质量分布图，是一种统计报告图，

由一系列高度不等的纵向条纹或线段表示数据分布的情况。

一般用横轴表示数据类型，纵轴表示分布情况。

直方图可以被归一化以显示“相对”频率。

显示属于几个类别中的每个类别的比例，其高度等于1。

"""

"""

Q-Q plot: Method 1

"""

plt.scatter(x = q, y = y, color = 'red')

# Add a 45-degree reference line

plt.plot([y[0], y[n-1]], [y[0], y[n-1]], color = 'blue', linewidth = 2)

plt.xlabel('Theoretical Quantiles')

plt.ylabel('Ordered Values' )

plt.title('Normal Q-Q plot')

plt.show()

"""

Q-Q plot: Method 2

"""

stats.probplot(y, dist = "norm", plot = plt)

plt.title("Normal Q-Q plot")

plt.show()