概率论与数理统计学习笔记

Roger Young

编译时间： 2017 年 3 月 20 日

Contents

Chapter 1

LaTeX 基本使用说明

1.1 入门常识

LaTeX 会默认丢弃命令后的空白字符。如果希望在命令后的到一个空白字符，可以在命令后加上 {} 和一个空格，或者加上一个特殊的空白距离命令。LaTeX 源文件中可以使用% 来表示注释，LaTeX 在处理文件时，会忽略% 后的改行剩余文本。如果需要使用较长的注释（注释块），可以使用 verbatim 包。

1.2 页面布局

1.2.1 分页、分段、断行和断字

有必要时 LaTeX 会进行必要的分页、断行和断字。在特殊情况下，需要使用命令指示 LaTeX 进行分页、分段、断行和断字。这些命令包括：

- \\ ：连着的两个反斜线指示 LaTeX 另起一行，但不另起一段。
- *newline* 命令：等同于 \\；
- * 命令：强行断行后，还禁止分页；
- *newpage* 命令：另起一新页。

- $\backslash hyphenation\{wordlist\}$ ：可以用来断字；
- $\backslash -$：提示断字位置。
- $\backslash linebreak$：强制断行；
- $\backslash nolinebreak$：强制不断行；
- $\backslash linebreak[priority]$：建议断行，priority 可选值为 0、1、2、3、4；
- $\backslash nolinebreak[priority]$：建议不断行，priority 可选值为 0、1、2、3、4；
- $\backslash cleardoublepage$ ：开始新的偶数页；
- $\backslash clearpage$：开始新页，并且导致当前浮动的表格、图片等都输出。；
- $\backslash newpage$：开始新页；
- $\backslash enlargethispage\{size\}$：扩大当前页；
- $\backslash pagebreak$：强制分页；
- $\backslash nopagebreak$：强制不分页。

1.2.2 页眉设置

1.2.3 页脚设置

$\backslash footnote[number]\{text\}$	在当前页尾插入编号的脚标

1.3 字体设置

1.4 参考文献引用

1.5 LaTeX 环境简介

1.6 数学公式

1.7 作图功能

LaTeX 可以通过内置的 picture 环境来创建简单的图形。对于创建更复杂的图形可以尝试以下包：

- beamer
- pgf, Portable Graphics Format
- tikz, TikZ

一个 picture 环境可以通过“$\backslash begin\,\{picture\}\,(x,y)$ ”和“$\backslash end\,\{picture\}$ ”，或者“$\backslash begin\,\{picture\}\,(x,y)(x_0,y_0)$ ”和“$\backslash end\,\{picture\}$”来创建。其中 x、y、x_0、y_0 和 $\backslash unitlength$ 相关。(x,y) 指示 LaTeX 为 picture 预留的大小。可选的参数 (x_0,y_0) 是为 picture 预留的方形的的左下角的坐标。可以通过“$\backslash setlength$ ”命令来设置：

$$\backslash setlength\,\{\backslash unitlength\}\,\{1.2cm\}$$

绘图命名的格式一般如下：

$$\backslash put(x,y)\,\{object\}$$

或者

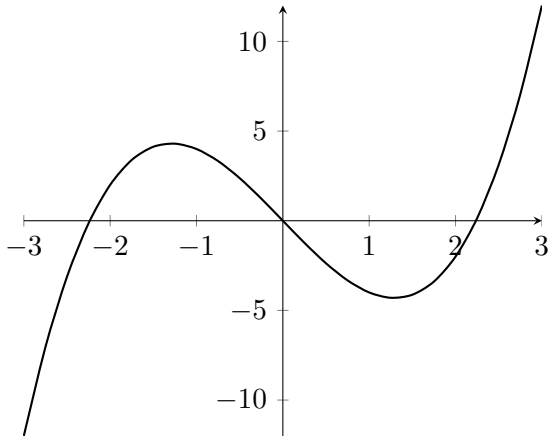
$$\backslash multiput(x,y)(\Delta x,\Delta y)\,\{n\}\,\{object\}$$

贝塞尔曲线是个例外，其格式为：

$$\backslash qbezier(x_1, y_1)(x_2, y_2)(x_3, y_3)$$

常用的画图命令如下：

类型	命令	说明
线段	$\backslash put(x, y)\{\backslash line(x_1, y_1)\{length\}\}$	
箭头	$\backslash put(x, y)\{\backslash vector(x_1, y_1)\{length\}\}$	
圆	$\backslash put(x, y)\{\backslash circle\{diameter\}\}$	
文本	$\backslash put(x, y)\{\$\ text\ \$\}$	
公式	$\backslash put(x, y)\{\$\ formulas\ \$\}$	
椭圆形	$\backslash put(x, y)\{\backslash oval(x, y)\}$	
椭圆形	$\backslash put(x, y)\{\backslash oval(x, y)[position]\}$	
贝叶斯曲线	$\backslash qbezier(x_1, y_1)(x_2, y_2)(x_3, y_3)$	



Chapter 2

机器学习的数学基础

2.1 概述

我们知道，机器学习的特点就是：以计算机为工具和平台，以数据为研究对象，以学习方法为中心；是概率论、线性代数、数值计算、信息论、最优化理论和计算机科学等多个领域的交叉学科。所以本文就先介绍一下机器学习涉及到的一些最常用的的数学知识。

2.2 线性代数

2.2.1 标量

一个标量就是一个单独的数，一般用小写的的变量名称表示。

2.2.2 向量

一个向量就是一列数，这些数是有序排列的。用过次序中的索引，我们可以确定每个单独的数。通常会赋予向量粗体的小写名称。当我们需要明确表示向量中的元素时，我们会将元素排列成一个方括号包围的纵柱：

我们可以把向量看作空间中的点，每个元素是不同的坐标轴上的坐标。

2.2.3 矩阵

矩阵是二维数组，其中的每一个元素被两个索引而非一个所确定。我们通常会赋予矩阵粗体的大写变量名称，比如 **A**。如果一个实数矩阵高度为 m，宽度为 n，那么我们说 $A \in R^{m \times n}$。

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{2.1}$$

2.2.4 张量

几何代数中定义的张量是基于向量和矩阵的推广，通俗一点理解的话，我们可以将标量视为零阶张量，矢量视为一阶张量，那么矩阵就是二阶张量。

例如，可以将任意一张彩色图片表示成一个三阶张量，三个维度分别是图片的高度、宽度和色彩数据。将这张图用张量表示出来，就是最下方的那张表格：

其中表的横轴表示图片的宽度值，这里只截取 0 319；表的纵轴表示图片的高度值，这里只截取 0 4；表格中每个方格代表一个像素点，比如第一行第一列的表格数据为 [1.0,1.0,1.0]，代表的就是 RGB 三原色在图片的这个位置的取值情况（即 R=1.0，G=1.0，B=1.0）。

当然我们还可以将这一定义继续扩展，即：我们可以用四阶张量表示一个包含多张图片的数据集，这四个维度分别是：图片在数据集中的编号，图片高度、宽度，以及色彩数据。

张量在深度学习中是一个很重要的概念，因为它是一个深度学习框架中的一个核心组件，后续的所有运算和优化算法几乎都是基于张量进行的。

2.2.5 范数

有时我们需要衡量一个向量的大小。在机器学习中，我们经常使用被称为范数 (norm) 的函数衡量矩阵大小。Lp 范数如下：

$$\|x\|_p \left(\sum_i |x_i|^p \right)^{\frac{1}{p}} \tag{2.2}$$

所以：

L1 范数 $||x||$：为 x 向量各个元素绝对值之和；

L2 范数 $||x||_2$ ：为 x 向量各个元素平方和的开方。

2.2.6 特征分解

许多数学对象可以通过将它们分解成多个组成部分。特征分解是使用最广的矩阵分解之一，即将矩阵分解成一组特征向量和特征值。

方阵 A 的特征向量是指与 A 相乘后相当于对该向量进行缩放的非零向量 ν：

$A\nu = \lambda\nu$

标量 λ 被称为这个特征向量对应的特征值。

使用特征分解去分析矩阵 A 时，得到特征向量构成的矩阵 V 和特征值构成的向量 λ，我们可以重新将 A 写作：

$A = V diag(\lambda) V^{-1}$

2.2.7 奇异值分解（SVD）

除了特征分解，还有一种分解矩阵的方法，被称为奇异值分解（SVD）。将矩阵分解为奇异向量和奇异值。通过奇异分解，我们会得到一些类似于特征分解的信息。然而，奇异分解有更广泛的应用。

每个实数矩阵都有一个奇异值分解，但不一定都有特征分解。例如，非方阵的矩阵没有特征分解，这时我们只能使用奇异值分解。奇异分解与特征分解类似，只不过这回我们将矩阵 A 分解成三个矩阵的乘积：

$$A = UDV^T$$

假设 A 是一个 $m \times n$ 矩阵，那么 U 是一个 $m \times m$ 矩阵，D 是一个 $m \times n$ 矩阵，V 是一个 $n \times n$ 矩阵。

这些矩阵每一个都拥有特殊的结构，其中 U 和 V 都是正交矩阵，D 是对角矩阵（注意，D 不一定是方阵）。对角矩阵 D 对角线上的元素被称为矩阵 A 的奇异值。矩阵 U 的列向量被称为左奇异向量，矩阵 V 的列向量被称右奇异向量。

SVD 最有用的一个性质可能是拓展矩阵求逆到非方矩阵上。另外，SVD 可用于推荐系统中。

2.2.8 Moore-Penrose 伪逆

对于非方矩阵而言，其逆矩阵没有定义。假设在下面问题中，我们想通过矩阵 A 的左逆 B 来求解线性方程：

$$Ax = y \tag{2.3}$$

等式两边同时左乘左逆 B 后，得到：

$$x = By \tag{2.4}$$

是否存在唯一的映射将 A 映射到 B 取决于问题的形式。

如果矩阵 A 的行数大于列数，那么上述方程可能没有解；如果矩阵 A 的行数小于列数，那么上述方程可能有多个解。

Moore-Penrose 伪逆使我们能够解决这种情况，矩阵 A 的伪逆定义为：

但是计算伪逆的实际算法没有基于这个式子，而是使用下面的公式：

其中，矩阵 U，D 和 V 是矩阵 A 奇异值分解后得到的矩阵。对角矩阵 D 的伪逆 D^+ 是其非零元素取倒之后再转置得到的。

2.2.9 几种常用的距离

设有两个 n 维变量 $A = [x_{11}, x_{12}, ..., x_{1n}]$ 和 $B = [x_{21}, x_{22}, ..., x_{2n}]$，则下面可以定义一些常用的距离公式：

Chapter 3

基本概念汇总

3.1 基本概念

- **确定性现象**：在一定条件下，必然发生的现象。
- **统计规律性**：在大量重复试验或观察中所呈现出的固有的规律性。
- **随机试验**：具有以下三个特点的试验称为随机试验。
  - 可以再相同的条件下重复地进行；
  - 每次试验的可能结果不止一个，并且能事先明确试验的所有可能结果；
  - 进行一次试验前，不能确定哪一个结果会出现。

  随机试验一般可以使用大写斜体字母表示，比如 E 。
- **样本空间**：我们将随机试验 E 的所有可能结果组成的集合称为 E 的样本空间，记为 S 。样本空间中的元素，即 E 的每个结果，称为**样本点**。
- 随机事件：简称事件，是指试验 E 的样本空间 S 的子集。再每次试验中，当且仅当这一子集的恶一个样本点出现时，称这一事件发生。

- 频率：在相同条件下，进行了 n 次试验，在这 n 次试验中，事件 A 发生的次数 n_A 称为事件 A 发生的频数。比值 $\frac{n_A}{n}$ 称为事件 A 发生的频率，并记成 $f_n(A)$ 。

- 概率：设 E 为随机试验，S 是它的样本空间，对于 E 的每一事件 A 赋予一个实数，记为 $P(A)$ ，称为事件 A 的概率。概率是表示事件在一次试验中发生的可能性的大小的数。

- 等可能概型：具有以下两个特点的随机试验 E 称为等可能概型。

  – 随机试验的样本空间只包括两个元素；

  – 试验中每个基本事件发生的可能性相同。

  等可能概型是概率论发展初期研究的主要对象，所以也称为古典概型。

- 条件概率：设 A ，B 是两个事件，且 $P(A) > 0$ ，称

$$P(B|A) = \frac{P(AB)}{P(A)}$$

  为事件 A 发生的条件下事件 B 发生的条件概率。

- 随机变量：设随机试验的样本空间为 $S = e$ ，$X = X(e)$ 是定义在样本空间 S 上的实值单值函数，对于任意实数 x ，集合 $e|X(e) \leq x$ 有确定的概率，则称 $X = X(e)$ 为随机变量。

- 离散型随机变量：可能取到的值是有限多个或者可列无限多个的随机变量。

- 离散型随机变量的分布律：设离散型随机变量 X 所有可能取的值为 $x_k(k = 1, 2, \dots)$,X 取各个可能值的概率，即事件 $X = x_k$ 的概率，为

$$P\{X = x_k\} = p_k, k = 1, 2, \dots$$

则称该式为离散型随机变量 X 的分布律。分布律也可以用表格的形式表示：

X	x_1	x_2	...	x_n	...
p_k	p_1	p_2	...	p_n	...

3.2 常用数学公式

$$C_a^r = \binom{a}{r} = \frac{a!}{r!(a-r)!} = \frac{a(a-1)\dots(a-r+1)}{r!} \tag{3.1}$$

3.3 重要公式的证明

Chapter 4

随机变量及其分布

4.1 离散型随机变量及其分布律

常用分布	分布律	分布函数
(0-1) 分布	$P\{X=k\}=p^k(1-p)^{1-k}, \quad k=0,1 \quad (0<p<1)$	

Team	P	W	D	L	F	A	Pts
Manchester United	6	4	0	2	10	5	12
Celtic	6	3	0	3	8	9	9
Benfica	6	2	1	3	7	8	7
FC Copenhagen	6	2	1	3	5	8	7

4.2 连续型随机变量及其概率密度

4.3 多维随机变量及其分布

4.4 随机变量的数学特征

Chapter 5

大数定律及中心极限定律

Chapter 6

参数估计

统计推断的基本问题可以分为两类，一类是估计问题，另一类是假设检验问题。

6.1 点估计

定义 1

设总体 X 的分布函数的形式已知，但它的一个或多个参数未知，借助于总体 X 的一个样本来估计总体未知参数的值的问题，称为参数的点估计问题。

点估计问题的一般提法如下：设总体 X 的分布函数 $F(x;\theta)$ 的形式已知。θ 为待估参数，X_1，X_2，…，X_n 是 X 的一个样本。x_1，x_2，…，x_n 为相应的一个样本值。点估计的问题就是要构建一个适当的统计量 $\hat{\theta}(X_1, X_2, \ldots, X_n)$，用它的观测值 $\hat{\theta}(x_1, x_2, \ldots, x_n)$ 作为未知参数 θ 的近似值，我们称 $\hat{\theta}(X_1, X_2, \ldots, X_n)$ 为 θ 的估计量，称 $\hat{\theta}(x_1, x_2, \ldots, x_n)$ 为 θ 的估计值。

由于估计量是样本的函数，因此对于不同的样本值，θ 的估计值一般是不相同的。同时，对于同一参数，用不同的估计方法求出的估计量可能也

不相同。可以通过无偏性、有效性和相合性来评价估计量。

6.1.1 无偏性

定义 2

假设 X_1，X_2，…，X_n 是总体 X 的一个样本，$\theta \in \Theta$ 是包含在通体 X 的分布中的待估参数（）Θ 是 θ 的取值范围）。若估计量 $\hat{\theta}(X_1, X_2, \ldots, X_n)$ 为 θ 的数学期望 $E(\hat{\theta})$ 存在，且对于任意 $\theta \in \Theta$ 有，

$$E(\hat{\theta}) = \theta$$

则称 $\hat{\theta}$ 是 θ 的**无偏估计量**。

在科学技术中将 $E(\hat{\theta}) - \theta$ 称为以 $\hat{\theta}$ 作为 θ 的估计的系统误差。无偏估计的实质意义就是无系统误差。

设总体 X 的均值为 μ，方差为 $\sigma^2 > 0$ 均未知，不论总体 X 服从什么分布，样本均值 $\overline{x}$ 都是总体均值 μ 的无偏估计，样本方差 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ 是总体方差的无偏估计，但估计量 $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ 却不是 σ^2 的无偏估计。

6.1.2 有效性

方差 ssh 是随机变量取值与其数学期望的偏离程度的度量，所以无偏估计以方差小者为好。进而引入估计量有效性的概念。

定义 3

设 $\hat{\theta}_1(X_1, X_2, \ldots, X_n)$ 与 $\hat{\theta}_2(X_1, X_2, \ldots, X_n)$ 都是 θ 的无偏估计量，若对于任意 $\theta \in \Theta$，有

$$D(\hat{\theta}_1) \leq D(\hat{\theta}_2)$$

则至少对于某一个 $\theta \in \Theta$，上式中的不等号成立，则称 $D(\hat{\theta}_1)$ 较 $D(\hat{\theta}_2)$ 有效。

6.1.3 相合性

定义 4

设 $\hat{\theta}(X_1, X_2, \ldots, X_n)$ 为参数 θ 的估计量，若对于任意 $\theta \in \Theta$，当 $n \to \infty$ 时，$\hat{\theta}(X_1, X_2, \ldots, X_n)$ 以概率收敛于 θ 则称有 $\hat{\theta}$ 为 θ 的相合估计量。

相合性意味着随着样本量的增大，一个估计量的值是否可以稳定于待估参数的真值。相合性时对一个估计量的基本要求，若估计量不具有相合性，那么无论将样本量取得多大，都不能将参数 θ 估计的足够准确。

构建估计量的方法通常有两种：**矩估计法**和**最大似然估计法**。

6.1.4 矩估计法

设 X 为概率密度为 $f(x; \theta_1, \theta_2, \ldots, \theta_n)$ 的连续型随机变量，或者 X 为分布律为 $P\{X = x\} = p(x; \theta_1, \theta_2, \ldots, \theta_n)$ 的离散型随机变量，其中的 $\theta_1, \theta_2, \ldots, \theta_n$ 为待估参数，$X_1, X_2, \ldots, X_n$ 是来自 X 的样本。假设总体 X 的前 k 阶矩

$$\mu_1 = E(X^l) = \int -\infty^{\infty} x^l f(x; \theta_1, \theta_2, \ldots, \theta_n) dx \quad X \text{ 为连续型}$$
$$\mu_1 = E(X^l) = \sum_{x \in R_x} x^l p(x; \theta_1, \theta_2, \ldots, \theta_n) \quad X \text{ 为离散型}$$

($l = 1, 2, \ldots, k$， R_x 是 X 可能的取值范围) 存在。一般来说，它们是 $\theta_1, \theta_2, \ldots, \theta_n$ 的函数，基于样本矩

$$A_l = \frac{1}{n} \sum_{i=1}^{n} X_i^l$$

依概率收敛于相应的总体矩阵 $\mu_l (l = 1, 2, \ldots, k)$, 样本矩的连续函数依概率收敛于相应的总体矩的连续函数，我们就用样本矩作为相应的总体矩的估计量，以样本矩的连续函数作为相应的总体矩的连续函数的估计量，这种估计方法称为矩估计法。

矩估计法的具体做法如下：

Chapter 7

假设检验

7.1 假设检验问题的 P 值法

例 7.1.1. 设总体 $X \sim N(\mu, \sigma^2)$，μ 未知，$\sigma^2 = 100$，现有样本 x_1，x_2，…，x_{52}，算得 $\overline{x} = 62.75$。现在来检验假设：

$$H_0 : \mu \leq \mu_0 = 60,\ H_1 : \mu > 60。$$

Proof. 采用 Z 检验法，检验统计量为：

$$Z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

以数据代入，得 Z 的观察值为：

$$Z = \frac{62.75 - 60}{\frac{10}{\sqrt{52}}} = 1.983$$

概率

$$P\{Z \geq z_0\} = P\{Z \geq 1.983\} = 1 - \Phi(1.983) = 0.0238$$

称之为 Z 检验法的右边检验的 p 值。 □

定义 5

假设检验问题的 p 值（probability value）是由检验统计量的样本观测值得出的原假设可能被拒绝的最小显著性水平。在现在计算机统计软件中，一般都会给出检验问题的 p 值。按照 p 值的定义，对于任意指定的显著性水平 α，就有：

1. 若 p 值 $\leq \alpha$，则在显著性水平 α 下拒绝 H_0；
2. 若 p 值 $> \alpha$，则在显著性水平 α 下接受 H_0；

这种利用 p 值来确定是否拒绝 H_0 的方法，称为 p 值法。

Chapter 8

方差分析

Chapter 9

回归分析