中国科学技术大学
University of Science and Technology of China

《人工智能数学原理与算法》
第2章：机器学习基础

2.3 概率统计基础

冯福利
fengfl@ustc.edu.cn

目录

随机变量

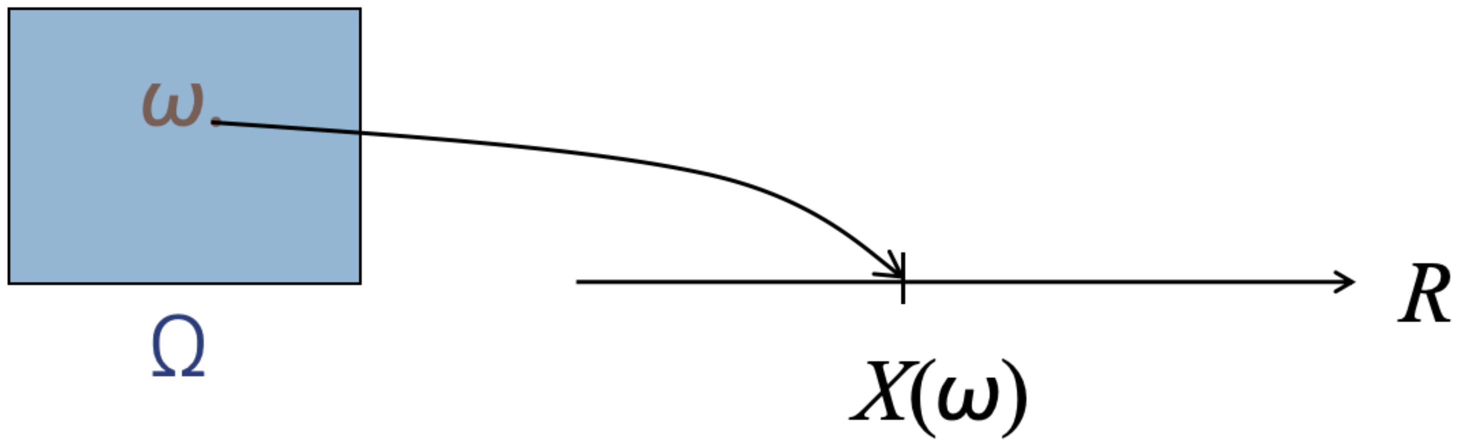
随机变量

- 在实际问题中，随机试验的结果可以用数值来表示，由此产生了随机变量的概念。
- 有些试验结果本身与数值有关，因此可用一个变量来表示试验的各种结果；
  - 抛一枚骰子出现的点数；
  - 任取10个产品中的次品数
- 有些试验结果与数值无关，但可以把结果数值化，即引入一个变量来表示试验的各种结果
  - 抛一枚硬币，正面向上用1表示，反面向上用0表示
  - 课上任选一名同学，用其学号表示
  - 火箭返回地球的落点位置可以用坐标或经纬度表示

随机变量

随机变量的定义

- 我们可以把**Ω**中的每一个样本点**ω**与一个实数**X**(**ω**)相对应， **X**(**ω**)可看成**ω**的实值函数



- 称实值函数**X**: **Ω** → ***R***为随机变量(random variable)，简记为r.v.

随机变量

随机变量的特点

- 它随试验结果的不同而取不同的值，因而取值具有随机性。在试验之前只知道它可能取值的范围，而不能预先确定取哪个值。

- 由于试验结果的出现具有一定的概率，因此随机变量取某个值或者某个范围内的值也有一定的概率。

随机变量在某范围的取值表示随机事件

随机变量

例子

- 抛一枚骰子，令X =出现的点数，则X是一r.v.
  - X的取值为1,2,3,4,5,6
  - $\{X \le 4\}$表示事件：点数不超过4
  - $\{X$为偶数$\}$表示事件：点数为偶数
- 观察某电子元件的寿命，令Z: 该元件的寿命，则Z是一r.v.
  - Z的取值为所有非负实数
  - $\{Z \le 500\}$表示事件：该元件寿命不超过500h

随机变量

随机变量的意义

- 引入随机变量后，可以利用随机变量来描述随机现象，对事件及事件概率的研究扩大为对随机变量及其取值规律的研究。

- 有了随机变量，可以使用更多的数学工具。

随机变量

随机变量的类型

通常分为两类：

- 离散型随机变量：所有可能取值可以一一列举
  - 例：取到次品的个数，抛骰子的点数

- 连续型随机变量：所有取值不可一一列举，可取一个区间内的所有值
  - 例：元件的寿命，到达车站的时刻

离散型随机变量

- 若随机变量X的取值是有限个或者可列无穷个，则称X为离散型随机变量。
- 设离散型随机变量X的所有可能取值为

$$x_1, x_2, \ldots, x_n, \ldots$$

并设$P(X = x_n) = p_n \ (n = 1, 2, \ldots)$，则称上式为$X$的分布律，常表示为

X	x_1	x_2	$\cdots,$	x_n	$\cdots$
P	p_1	p_2	$\cdots,$	p_n	$\cdots$

随机变量

离散型随机变量性质

- 对任意的自然数n，有$p_n \geq 0$
- $\sum_n p_n = 1$

例：设随机变量X的分布律为

$$P(X = n) = \frac{c}{4^n},\ n = 1, 2, \ldots$$

求常数c.

解：据分布律性质知

$$1 = \sum_n P(X = n) = \sum_n \frac{c}{4^n} = \frac{c}{3}$$

故$c = 3$.

随机变量

例

- 假设姚明罚球的命中率为0.9，求他两次独立罚球命中次数X的分布律。

解：X的取值为0,1,2.

$$P(X=0)=(1-0.9)\times(1-0.9)=0.01$$

$$P(X=1)=2\times0.9\times(1-0.9)=0.18$$

$$P(X=2)=0.9\times0.9=0.81$$

X	0	1	2
P	0.01	0.18	0.81

数学期望

- 设X为离散型随机变量，分布律为

$$P(X=x_i)=p_i,\ i=1,2,\ldots$$

若级数$\sum_{i=1}^{\infty}x_ip_i$绝对收敛，则称级数$\sum_{i=1}^{\infty}x_ip_i$的和为$X$的数学期望，记为$E(X)$,即

$$E(X)=\sum_{i=1}^{\infty}x_ip_i\,.$$

若级数$\sum_{i=1}^{\infty}|x_i|p_i$发散，则称$X$的数学期望不存在

例

- 有4个盒子，编号为1,2,3,4。现将3个球随机放入4只盒子。用X表示有球盒子的最小号码，求$E(X)$.

解：先求X的分布律。

$$P(X=1)=\frac{37}{64},\ P(X=2)=\frac{19}{64}$$

$$P(X=3)=\frac{7}{64},\ P(X=4)=\frac{1}{64}$$

所以$E(X)=1\times\frac{37}{64}+2\times\frac{19}{64}+3\times\frac{7}{64}+4\times\frac{1}{64}=\frac{25}{16}$

随机变量

条件期望

- 类似于条件概率，常需要在某事件发生的条件下，对随机变量进行分析

- 在事件A的条件下，随机变量X的条件期望可定义为

$$E[X \mid A] = \sum_{x} xP(X = x \mid A)$$

其中，求和是对X所有可能取值而言。这里，$P(X = x \mid A)$被称为事件A的条件下的X的条件分布律。

条件期望

- 特别地，

$$E[X \mid Y=y]=\sum_x xP(X=x \mid Y=y)$$

- 例：随机抛骰子两次。X_1:第一个点数；X_2:第二个点数；X: 点数和

$$E[X_1|X=5]=\sum_{x=1}^{4} xP(X_1=x \mid X=5)=\sum_{i=1}^{4} x\frac{1}{4}=\frac{5}{2}$$

$$E[X|X_1=2]=\sum_{x=3}^{8} xP(X=x|X_1=2)=\sum_{x=3}^{8} x\frac{1}{6}=\frac{11}{2}$$

方差

克服期望/均值的局限性

- 均值在很多场合的分布情况刻画不准确

张村有个张千万，
隔壁九个穷光蛋。
平均起来算一算，
人人都是张百万。

- 设X为随机变量,对于$m \in R$，若

$$P(X \geq m) \geq \frac{1}{2} \text{ 且 } P(X \leq m) \geq \frac{1}{2},$$

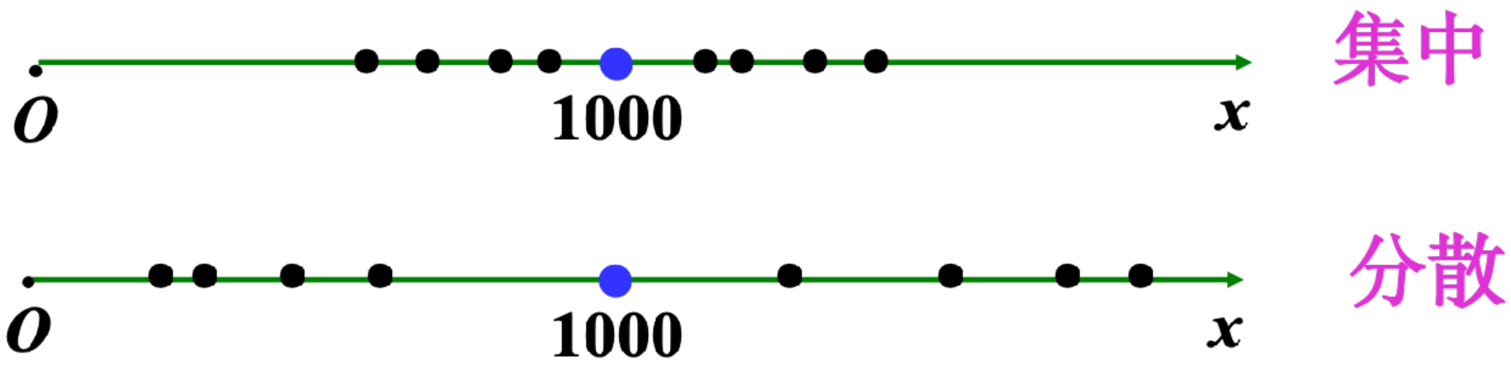
则称m为X的中位数

- 刻画随机变量取偏离期望的值的概率.
  - 尾部分布：$P(X \geq t), P(X \leq -t)$

方差

□ 方差的引入

例：有两批灯泡，其平均寿命都是1000小时，

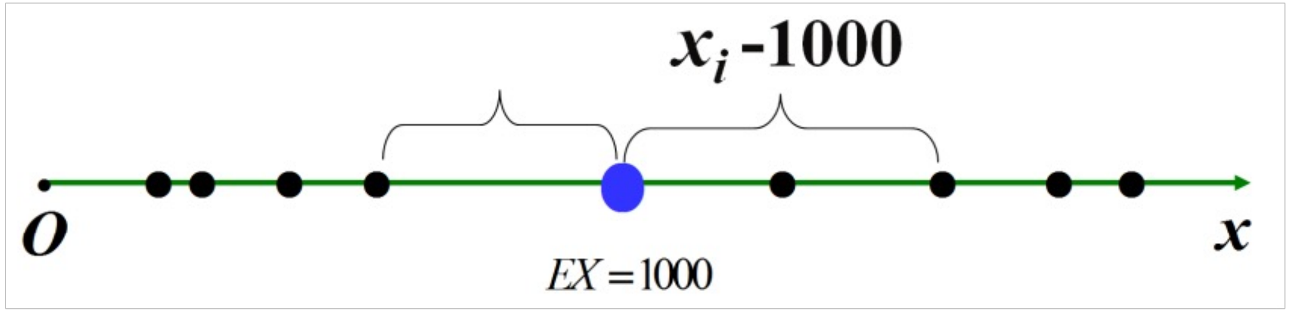


用**方差**来反映随机变量取值的离散程度。

方差

□ 方差的引入

X	x_1	x_2	$\cdots,$	x_n	$\cdots$
P	p_1	p_2	$\cdots,$	p_n	$\cdots$



用$\sum_k (x_k - E[X])^2 p_k$定义离散型随机变量X的方差

方差

□ 方差的定义

- 设X是一个随机变量，若$E[(X-E[X])^2]$存在，则称$E[(X-E[X])^2]$为X的方差(variance)，记为$D(X)$或者$\mathrm{Var}(X)$，即

$$D(X)=\mathrm{Var}(X)=E[(X-E[X])^2].$$

- 称$\sqrt{D(X)}$为标准差或均方差（standard deviation），记为$\sigma(X)$.

方差

- 例

  - 假设随机变量X为常数，则$D(X) = 0.$

  - 假设随机变量X依概率$1/k$取值$kE[X]$,依概率$1-1/k$取值0,则$D(X) = (k-1)(E[X])^2$

  - 直觉上而言，当随机变量取值越接近期望，方差越小；反之方差越大。

常见分布

0-1分布

- 如果随机试验只有两个结果：A与$\bar{A}$，则称该试验为伯努利(Bernoulli)试验。
- 定义随机变量

$$X=\begin{cases}1 & \text{若}A\text{发生}\\ 0 & \text{若}A\text{不发生}\end{cases}$$

记$P(A)=p$，则称X服从**0–1分布**

Jacob Bernoulli

X	0	1
P	$1-p$	p

常见分布

0-1分布的特点

- 若X服从参数为p的0–1分布，则

$$E(X)=p=P(X=1).$$

- 对随机事件A，可以定义指示变量

$$X_A=\begin{cases}1 & \text{若}A\text{发生}\\ 0 & \text{若}A\text{不发生}\end{cases}$$

则X_A服从0–1分布。

引入指示变量是简化问题分析的有效手段

N重伯努利实验

- 有一类独立重复试验概型，具有如下特点：
  - 每次试验只有两种结果：A与$\bar{A}$
  - 试验进行n次，每次试验结果相互独立

则称该独立重复试验为**n重伯努利试验**。

记$P(A)=p, P(\bar{A})=1-p=q.$

- 设X为n重伯努利试验中事件A发生的次数，

$$P(X=i)=\binom{n}{i}p^i(1-p)^{n-i},\ i=0,1,\ldots,n$$

二项分布

常见分布

二项分布

- 若随机变量X的分布律为

$$P(X=i)=\binom{n}{i}p^i(1-p)^{n-i},\ i=0,1,\ldots,n$$

则称X**服从参数为n,p的二项分布**(其中n为自然数，$0\le p\le 1$为参数),记作

$$X\sim B(n,p)$$

- 分布律的验证
  - $P(X=i)\ge 0$
  - $\sum_{i=0}^{n}P(X=i)=1$

常见分布

◻ 高斯分布

- 高斯分布友其方差和期望定义

$$E(X)=\mu = \int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

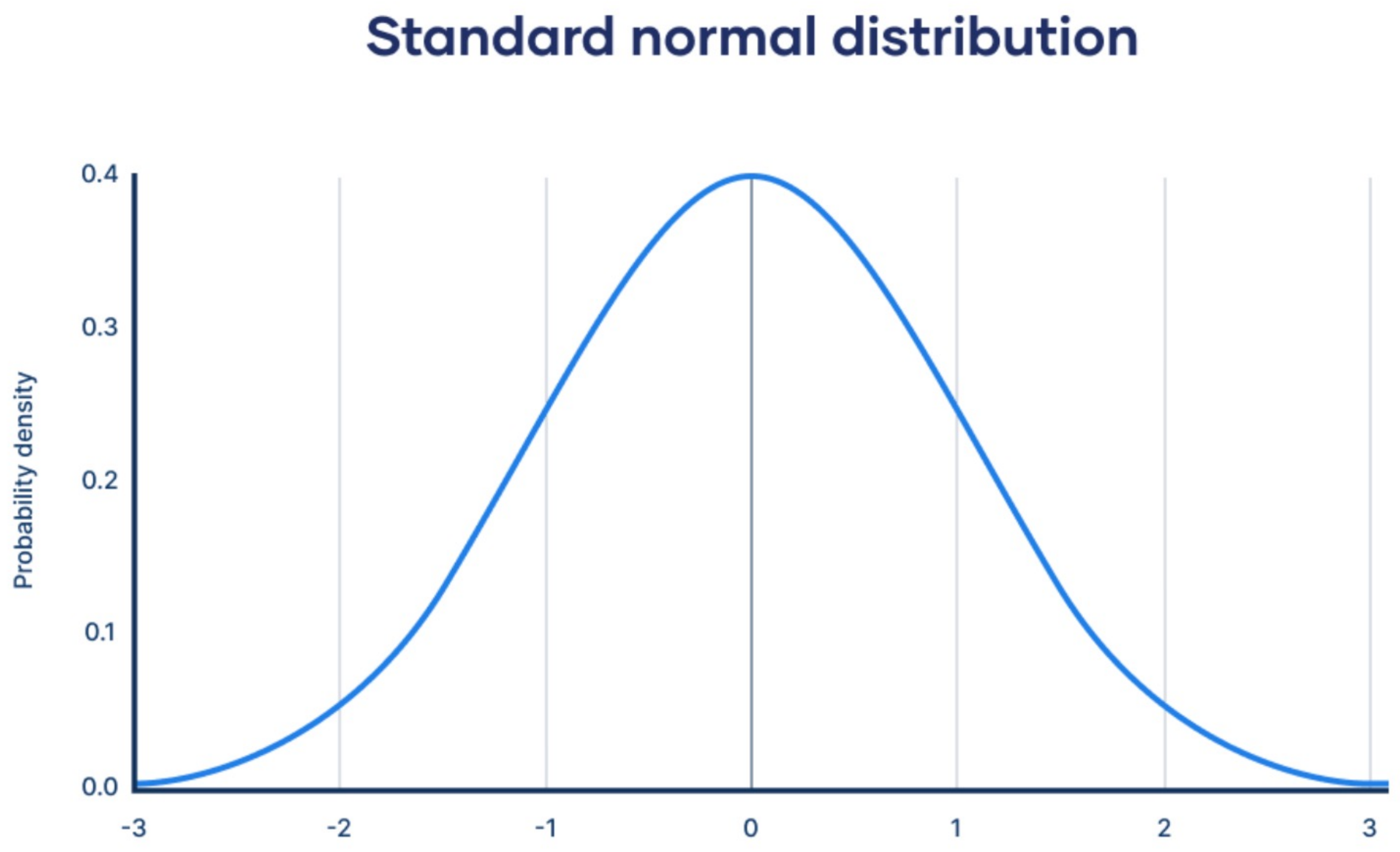
$$Var(X)=\sigma^2 = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx) - \mu^2$$

Standard Deviation(X)=σ

常见分布

高斯分布

- 均值取0，方差为1时，则称为标准高斯分布



概率悖论

□ 常见概率悖论

辛普森悖论

📌 **什么是辛普森悖论（Simpson's Paradox）？**
辛普森悖论是一种统计现象，**当数据被分组分析时，某个趋势可能出现；但当数据合并后，趋势可能完全相反。**
它揭示了**分组变量（隐藏变量）** 可能如何影响数据解释，导致误导性结论。
🎯 **举例 1：大学录取率的性别偏见**
📊 **问题**
某大学被指控在招生中歧视女性，理由是：
•**男性的录取率：45%**
•**女性的录取率：30%**
这似乎表明男性比女性更容易被录取。然而，当按学院分类后，发现：

学院	男性申请人数	男性录取率	女性申请人数	女性录取率
A	100	**80%**	10	**90%**
B	400	**30%**	200	**35%**

🔍 **观察**

- 学院 **A** 录取率较高，女性的录取率（90%）高于男性（80%）。
- 学院 **B** 录取率较低，女性录取率（35%）仍然高于男性（30%）。
- 但总体来看，女性的总录取率（30%）低于男性（45%）。

概率悖论

□ 常见概率悖论

辛普森悖论

🎯 **举例 2：医学治疗效果**

📊 **问题**

某种新药用于治疗肾结石，研究人员比较了**新药与传统治疗**的成功率：

治疗方法	总成功率
新药	78%
传统方法	83%

看起来**传统方法更有效**，但当按“结石大小”分组后：

结石大小	新药成功率	传统方法成功率
小结石	93%	87%
大结石	73%	69%

📌 **解释**

•**新药在每种结石大小的子组中都比传统方法更有效。**

•但由于**新药更多用于治疗难治的“大结石”**，而传统方法更多用于治疗较容易的“小结石”，导致总体数据看起来传统方法更有效。

概率悖论

□ 常见概率悖论

📌 **幸存者偏差（Survivorship Bias）**
幸存者偏差是一种**选择偏差**，当我们只关注“成功者”或“幸存者”时，往往会忽略那些失败或被淘汰的案例，从而得出错误的结论。
🎯 **经典案例 1：二战轰炸机的装甲加固问题**
📊 **问题**
二战期间，盟军希望减少轰炸机被敌人击落的概率。研究人员分析了**成功返航的轰炸机**，发现机身上弹孔分布如下：
•机翼：弹孔较多
•机身：弹孔较多
•引擎：弹孔较少
结论？ 研究人员最初认为**应该在弹孔最多的地方加厚装甲。**

📌 **现实分析**
统计学家 **亚伯拉罕·瓦尔德（Abraham Wald）** 指出，这种方法是错误的！
•**返航的轰炸机** 代表的是 **幸存者**，它们在机翼和机身中弹后仍然能飞回来。
•**真正被击落的轰炸机** 可能是**引擎中弹**，但它们没有返回，因此数据中没有它们的弹孔分布。

✅ **正确做法：应该加强引擎的装甲，而不是机翼和机身！**

概率悖论

☐ 常见概率悖论

🎯 **森伯恩悖论（Sleeping Beauty Paradox）**
📊 **问题**
1.睡美人被催眠，参加一个实验。
2.周日抛一枚**公平硬币**：
   1. 如果**正面**，她周一醒来一次，然后实验结束。
   2. 如果**反面**，她周一醒来一次，然后被催眠，周二再醒来一次。
3.每次醒来，她都不知道今天是周一还是周二，被问：**“你认为硬币是正面的概率是多少？”**
🤔 **直觉 vs. 现实**
•**直觉**：硬币公平，概率应该是 **1/2**。
•**数学计算**：
- 可能的情况：
  - **正面 & 周一（1 次唤醒）**
  - **反面 & 周一（1 次唤醒）**
  - **反面 & 周二（1 次唤醒）**
- 由于“反面”导致她醒来的次数是“正面”的 2 倍，**贝叶斯定理计算后，硬币正面的概率是 1/3！**

✅ **最佳答案：硬币是正面的概率是 1/3**，不是 1/2。

概率悖论

☐ 常见概率悖论

🎯 **生日悖论（Birthday Paradox）**

📊 **问题**

在一个 **23 人的房间里**，至少有两人生日相同的概率是多少?

🤔 **直觉 vs. 现实**

•**直觉**：365 天，有 23 个人，似乎很难碰上相同生日。

•**实际概率**：

- 计算后发现，**相同生日的概率 ≈ 50.7%** 🎂🎉
- **只需要 23 个人，概率就超过 50%**！

✅ **这一悖论让人低估了生日重复的可能性，因为我们错误地假设“匹配自己生日”的概率，而不是“任意两人匹配”。**

□ 常见概率悖论

🎯 **贝特朗悖论（Bertrand's Paradox）**
📊 **问题**
假设你在一个**单位圆内随机选择一条弦**，这条弦**超过正三角形一边长度的概率是多少?**
🤔 **直觉 vs. 现实**
•**不同的“随机”方法**，会导致完全不同的答案：

- **方法 1**：随机选择两个点并连线，概率 ≈ 1/3
- **方法 2**：随机选择弦的中点，概率 ≈ 1/2
- **方法 3**：随机选择弦的角度，概率 ≈ 1/4

✅ **悖论点：不同的“随机化”定义会影响概率分布，说明概率问题中“如何定义随机”至关重要!**