数据科学基础

Foundations of Data Science

2.1 概率的定义

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

当一件事情你还不能用数学符号描述，

那么说明你还没有想清楚这件事情！

元素与数据的映射

映射

$\{1, 0\}$

样本点e

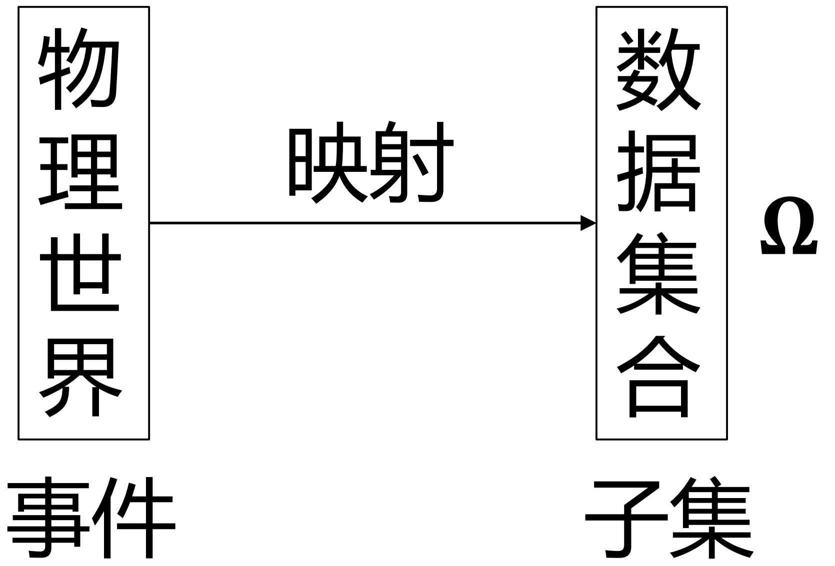
映射

$\{0,1,2\ldots\}$

样本空间Ω

为什么要映射成集合?

事件的集合表示



- 样本空间Ω的任意子集A称为（随机）事件
- 观察到样本点e，若$e \in A$则称这一事件发生

事件的集合表示

- 基本事件：由一个样本点组成的单点集
- 复合事件：由两个或两个以上样本点组成的集合
- 必然事件：全集Ω
- 不可能事件：空集∅

事件的集合运算

- 包含：$A \subseteq B$，即事件A发生必然导致事件B发生
- 相等：$A = B$，即$A \subseteq B$ 且 $B \subseteq A$
- 和：$A \cup B$，即A和B至少一个发生
- 差：$A - B$，即事件 A 发生且事件 B 不发生。
- 积：$A \cap B$，也记作AB，即事件A和B都发生
- 互不相容：$AB = \emptyset$，即A和 B不能同时发生
- 互逆：$A \cup B = \Omega$且 $AB = \emptyset$，A和B互逆，通常B记为$\bar{A}$

更多详情，请咨询课程助教。

复杂事件的集合运算表示

- A发生而B与C都不发生表示为：$A\bar{B}\bar{C}=A-B-C=A-(B\cup C)$
- A与B 都发生而C不发生表示为： $AB\bar{C}=AB-C=AB-ABC$
- 三个事件都发生表示为：ABC
- 三个事件恰好发生一个表示为：$A\bar{B}\bar{C}+\bar{A}B\bar{C}+\overline{AB}C$
- 三个事件恰好发生两个表示为：$AB\bar{C}+\bar{A}BC+A\bar{B}C$
- 三个事件至少发生一个表示为：$A\cup B\cup C$

事件发生的频率

- 重复观察n次事件A发生的次数n_A称为A的频数
- 比值$\frac{n_A}{n}$称为事件A发生的频率，并记成$f_n\ (A)$

频率具有以下性质：

1. $0 \le f_n(A) \le 1$;
2. $f_n(\Omega) = 1$
3. 若A_1， $\cdots A_k$两两互不相容，则

$$f_n(A_1 \cup \cdots \cup A_k) = (A_1) + \cdots + f_n\ (A_k)$$

频率的收敛性

观察、实验、验证

当n足够大时，$f_n(A)$收敛于某个常数

如何证明?

古典概率

若 Ω 是**有限样本空间**，其样本点为 e_1， $\cdots$， e_n ，在有限样本空间中引进概率。
1/n称为事件$\{e_i\}$的概率，记为$P(\{e_i\})$。

$$P(\{e_1\}) + \cdots + P(\{e_n\}) = P(\Omega) = 1$$

有限样本空间

等可能

十九世纪初数学家（拉普拉斯）定义的概率

古典概率计算

设样本空间为$\Omega = \{e_1, \cdots, e_n\}$，根据等概率，有

$$P(\{e_1\}) = \cdots = P(\{e_n\})$$

且基本事件是两两不相容的，于是

$$1 = P(\Omega) = P(\{e_1\} \cup \cdots \cup \{e_n\})$$

$$= P(\{e_1\}) + \cdots + P(\{e_n\}) = nP(\{e_i\})$$

$$P(\{e_i\}) = \frac{1}{n}, \quad i = 1, \cdots, n$$

若事件A包含k个基本事件，则有

$$P(A) = \sum_{i=1}^{k} P(\{e_{i_k}\}) = \frac{k}{n}$$

几何概率

数据表示如何从有限集合到无限集合推广？

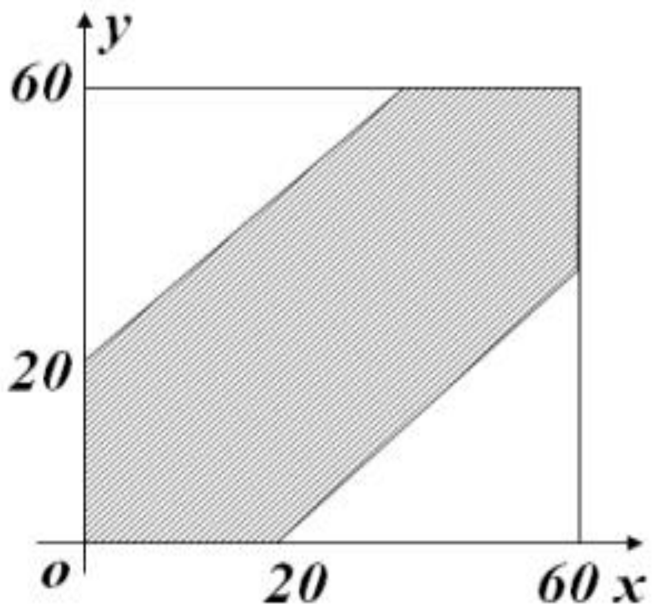
- 某人午觉醒来，发觉表停了，他打开收音机，想听电台报时，求他等待的时间短于10分钟的概率。
- 在400毫升自来水中有一个大肠杆菌，今从中随机抽出2毫升水样放到显微镜下观察，求发现大肠杆菌的概率。

一种相当自然的答案是认为例1所求的概率等于$\frac{1}{6}$，例2所求的概率等于$\frac{1}{200}$。

在求这些概率时，我们采纳了某种几何特性的等可能假设。

几何概率示例

[约会问题] 两人相约7点到8点在某地会面，先到者等候另一人20分钟，过时就可离去，试求这两人能会面的概率。



解：以x，y分别表示两人到达时刻，则会面的充要条件为$|x-y|\le 20$. 这是一个几何概率问题，可能的结果全体是边长为60的正方形里的点，能会面的点的区域用阴影标出，所求概率为

$$P(A)=\frac{60^2-40^2}{60^2}=\frac{5}{9}$$

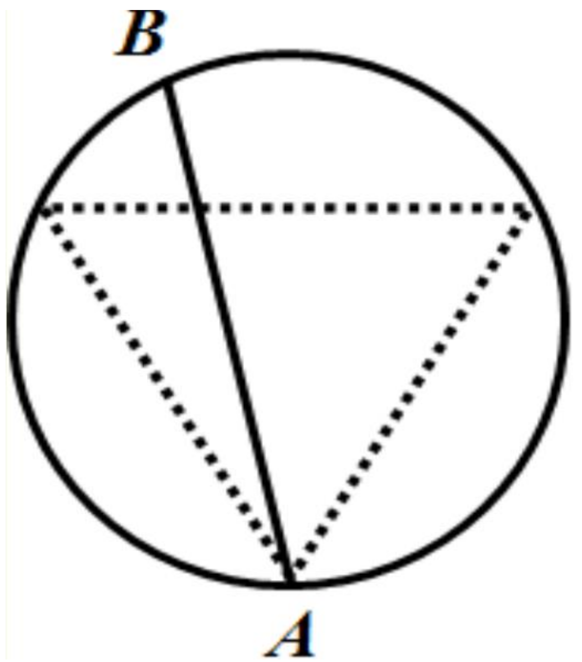
几何概率练习

在半径为1的圆内随机地取一条弦，

问弦长超过$\sqrt{3}$的概率等于多少?

几何概率练习1

[解法一] 任何弦交圆周二点，不失一般性，先固定其中一点于圆周上，以此点为顶点作一等边三角形，显然只有落入此三角形内的弦才满足要求，这种弦的另一端跑过的弧长为整个圆周的$\frac{1}{3}$，故所求概率等于$\frac{1}{3}$。

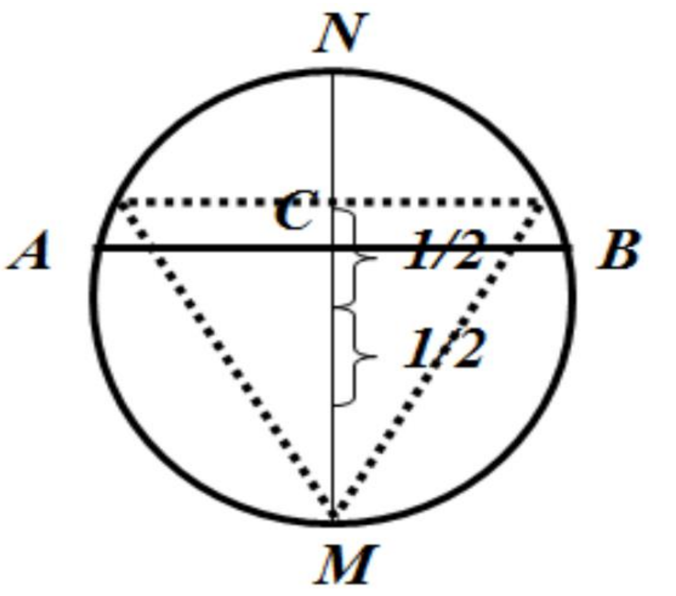


几何概率练习2

[解法二] 弦长只跟它与圆心的距离有关，而与方向无关，因此可以假定它垂直于某一直径。

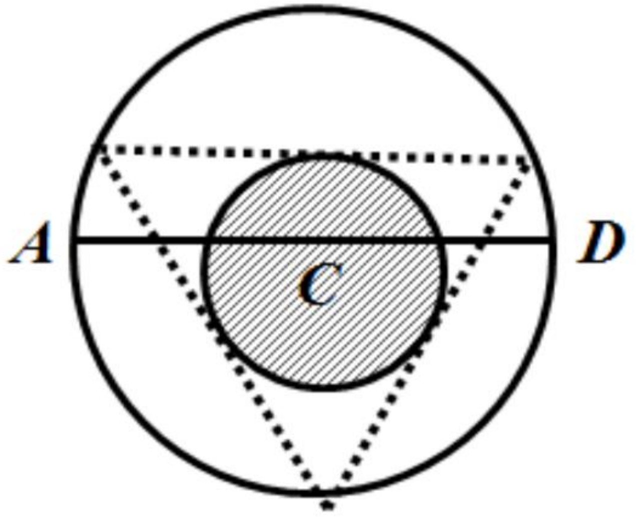
当且仅当它与圆心的距离小于$\frac{1}{2}$时，才满足要求，

因此所求概率为$\frac{1}{2}$。



几何概率练习3

[解法三] 弦被其中点唯一确定，当且仅当其中点属于半径为1/2的同心圆内时，才满足要求，此小圆面积为大圆面积的 $\frac{1}{4}$，故所求概率等于 $\frac{1}{4}$。



小结

在数据映射中，我们需要遵循物理世界到数据集合的某种**结构保持。**

概率的公理化定义

Ω为样本空间，对于每一事件A赋予一实数$P(A)$，

若$P(.)$满足下列条件则称为概率：

(1) 非负性：$0 \le P(A) \le 1$；

(2) 规范性：$P(\Omega) = 1$

(3) 可加性：A和B互不相容则$P(A \cup B) = P(A) + P(B)$

我把概率想清楚了！

更多详情，请咨询课程助教。

概率的性质

- 定理1：$\mathrm{P}(\emptyset) = 0$
- 定理2：$P(\bar{A}) = 1 - P(A)$
- 定理3：若$A \subset B$，则有

  $P(A) \leq P(B)$，$P(B - \mathrm{A}) = \mathrm{P}(B) - P(A)$
- 定理4：对于任意两个事件A和B有

  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$