Probability Notes 1 样本空间和概率

2014-04-11

1 样本空间和概率（Sample Space and Probability）

- 可以从两个角度理解概率：

1. 概率为：发生频率，反复进行同样实验出现某结果的频率
2. 概率为：主观信念，认为事件发生的可能性

1.1 概率模型的元素（Elements of a Probabilistic Model）

- 概率模型（probabilistic model）是对不确定情形的一个数学描述
- 概率模型包括如下两个元素

1. 样本空间（sample space Ω）：一个实验的所有可能结果（outcome）的集合
2. 概率规则（probability law）：赋予事件（event）A（若干可能结果的集合），一个代表信念或知识的非负数字$P(A)$（称为事件A的概率）的规则

- 正当（valid）的样本空间需要满足如下两个条件：

1. 样本空间内的元素是互相排斥（Mutually exclusive）的：若一个结果发生了，则其他结果均不发生
2. 样本空间必须是完全穷尽（Collectively Exhaustive）的：样本空间要包含所对应的实验的所有可能结果

1.2 概率公理（Probability Axioms）

- 概率模型中的概率规则需要满足如下公理：

1. 非负性（Nonnegativity）：
   $\forall A \subset \Omega,\ P(A) \geqslant 0$
2. 可加性（Additivity）：
   如果$A_1, A_2 \cdots$是相互独立的事件，则有：
   $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$
3. 归一性（Normalization）：
   $P(\Omega) = 1$

- 若将样本空间想象成一个单位质量的金属片，则事件A对应于金属片上的一个相应区域，$P(A)$对应于这个区域的金属的质量
- 由上述三公理可以推导出其他公式，如：
  $P(\varnothing) = 1 - P(\Omega) = 0$
  $1 = P(\Omega) = P(\Omega \cup \varnothing) = P(\Omega) + P(\varnothing) = 1 + P(\varnothing)$

1.3 离散概率规则（Discrete Probability Law）

- 如果样本空间包含有限个可能的结果，则：
  $P(\{s_1, s_2, \cdots, s_n\}) = P(\{s_1\}) + P(\{s_2\}) + \cdots + P(\{s_n\}) = P(s_1) + P(s_2) + \cdots + P(s_n)$

1.4 离散均匀概率规则（Discrete Uniform Probability Law）

- 样本空间包含n个可能结果，并且所有结果发生的可能性均相同，则事件A的概率为：
  $P(A) = \frac{\text{number of elements of } A}{n}$

1.5 概率规则的属性（Properties of Probability Laws）

- 若A, B, C为三个事件

1. $A \subset B \rightarrow P(A) \leqslant P(B)$
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
3. $P(A \cup B) \leqslant P(A) + P(B)$
4. $P(A \cup B \cup C) = P(A) + P(A^c \cap B) + P(A^c \cap B^c \cap C)$
5. $P(B) = P(A) + P(A^c \cap B) \geqslant P(A)$
6. $P(A \cup B) = P(A) + P(A^c \cap B)$
7. $P(B) = P(A \cap B) + P(A^c \cap B)$
8. $P(A) = P(A \cap B) + P(A \cap C) + P(A \cap B^c \cap C^c) + P(A \cap B \cap C)$
9. $P((A \cap B^c) \cup (A^c \cap B)) = P(A) + P(B) - 2P(A \cap B)$
10. $P(A \cap B) \geqslant P(A) + P(B) - 1$ （Bonferroni's inequality）
11. 若互斥事件$S_1, S_2, \cdots, S_n$构成样本空间的一个分划，则：
    $P(A) = \sum_{i=1}^{n} P(A \cap S_i)$

1.6 条件概率（Conditional Probability）

- 给定事件B，并知$P(B) > 0$，则事件A的条件概率$P(A|B)$定义为：
  $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- 条件概率帮助我们在掌握了部分信息的条件下理性地认识实验结果
- 条件概率的概率规则也满足概率规则的属性，例如：
  $P(A \cup B) \leqslant P(A) + P(B)$ 的条件概率版本为：
  $P(A \cup C|B) \leqslant P(A|B) + P(C|B)$

1.7 乘法规则（Multiplication Rule）

- 假设所有的条件事件的概率均为正，则有：
  $P(\cap_{i=1}^{n} A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n | \cap_{i=1}^{n-1} A_i)$

1.8 全概率定理（Total Probability Theorem）

- 若互斥事件$A_1, \cdots, A_n$构成样本空间的一个分划（partition）（即任何一个可能的结果都唯一地属于$A_1, \cdots A_n$中的一个事件），并且假设$\forall i,\ P(A_i) > 0$则对于任意事件B，有：
  $$P(B) = P(A_1 \cup B) + \cdots + P(A_n \cup B)$$
  $$= P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)$$
- 全概率定理为我们提供了一个分而治之地计算各种事件概率的方法

1.9 贝叶斯法则（Bayes' Rule）

- 若互斥事件$A_1, \cdots A_n$构成样本空间的一个排列，并且假设$\forall i,\ P(A_i) > 0$则对于任意满足$P(B) > 0$的事件B，有：

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)}$$
$$= \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1)+\cdots+P(A_n)P(B|A_n)}$$

- 称$P(A_i)$是先验概率，而称$P(A_i|B)$是事件B发生后的后验概率

1.10 独立性（Independence）

- 两个事件的独立性
- 若$P(A \cap B) = P(A)P(B)$ 则称事件A与事件B独立 如果额外地有$P(B) > 0$，则：
  $P(A|B) = P(A)$
  $P(A \cap B) = P(A)P(B)$
- 若事件A, B独立，则事件B^c, A也独立

- 条件独立性
- 已知事件C满足$P(C) > 0$，若$P(A \cap B|C) = P(A|C)P(B|C)$，则称$A, B$在事件$C$下相互独立 如果额外地有$P(B \cap C) > 0$，则$P(A|B \cap C) = P(A|C)$
- 独立性并不蕴含条件独立性，条件独立性也不蕴含独立性

- 多个事件间的独立性
- 若$P(\cap_{i \in S} A_i) = \prod_{i \in S} P(A_i)$， for every subset S of $\{1, 2, \cdots, n\}$

1.11 数数（Counting）

- n个对象的排列（Permutation）：
  $n!$
- n个对象中选出k个的排列（K-Permutation）：
  $\frac{n!}{(n-k)!}$
- n个对象中k个对象对象的组合（Combination）：
  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- n个对象分划（Partition）为r组，第i组有n_i个对象：
  $\binom{n}{n_1, n_2, \cdots, n_r} = \frac{n!}{n_1!n_2!\cdots n_r!}$

Probability Notes 2 离散随机变量

2014-04-12

2 离散随机变量（Discrete Random Variable）

2.1 随机变量（Random Variable）

- 随机变量（Random Variable）是实验中结果的一个实值函数
- 随机变量的函数也是随机变量
- 一个随机变量可以条件于其他事件或其他随机变量
- 一个随机变量可以独立于其他事件或其他随机变量

2.2 离散随机变量（Discrete Random Variable）

- 离散随机变量（Discrete Random Variable）是实验结果的一个实值函数，并且函数的取值是有限或可数的
- 离散随机变量有概率质量函数（Probability Mass Function,PMF），函数描述了随机变量每个可能取值的概率
- 离散随机变量的函数也是一个离散随机变量，其概率质量函数可以从原本离散随机变量的概率质量函数获得

2.3 概率质量函数（Probability Mass Function）

- X 是一个离散随机变量
- $\{X = x\}$ 表示该离散随机变量取值为x的事件（实验中所有导致X 取值为x的结果的集合）
- $P(\{X = x\})$ 表示上述事件发生的概率
- 对于特定的x,$p_X(x) = P(\{X = x\})$ 是取值x的概率质量（Probability Mass）
- $p_X(x)$是离散随机变量X 的概率质量函数

2.4 概率质量函数的计算（Calculation PMF of a Random Variable X）

- 对于离散随机变量X 的所有可能取值x

1. 找到所有属于事件$\{X = x\}$ 的结果
2. 将这些结果的概率相加获得$p_X(x)$

2.5 离散随机变量举例

1. 伯努利随机变量（Bernoulli Random Variable）

   实验：扔1次硬币

   结果：正面朝上（head）或反面朝上（tail）

   概率规则：

   $$P(\text{head}) = p$$

   $$P(\text{tail}) = 1 - p$$

   事件：

   $$X = \begin{cases} 1 & \text{if a head} \\ 0 & \text{if a tail} \end{cases}$$

   概率质量函数：

$$p_X(k) = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases}$$

2. 二项随机变量（Binomial Random Variable）
   实验：扔n次硬币
   结果：共有2^n种可能结果，每个结果都是一个长度为n的正面朝上和反面朝上的序列
   概率规则：设某一结果a_i中出现i次正面朝上，出现$n-i$次反面朝上，则
   $P(a) = p^i \times (1-p)^{n-i}$
   事件：$\{X = k\}$ 扔n次硬币，其中出现k次正面朝上
   概率质量函数：

$$p_X(k) = P(X = k) = \binom{n}{k} = p^k (1-p)^{n-k} \quad , \quad k = 0, 1, \cdots, n$$

   二项随机变量的属性：

$$\sum_{k=0}^{n} p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = 1$$

3. 几何随机变量（Geometric Random Variable）
   实验：不断地扔硬币
   事件：$\{X = k\}$ 扔n次硬币，在第k次时首次出现正面朝上
   概率质量函数：

$$p_X(k) = (1-p)^{k-1} p, \quad k = 1, 2, \cdots,$$

4. 泊松随机变量（Poisson Random Variable）
   概率质量函数：

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \cdots,$$

   泊松随机变量的属性：

$$\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = 1$$

5. 离散均匀随机变量（Discrete Uniform Random Variable）
   $X \sim unif[a, b]$
   概率质量函数：

$$p_X(k) = \begin{cases} 1/b - a + 1 & \text{if } k = a, a+1, \cdots, b \\ 0 & \text{otherwise} \end{cases}$$

2.6 离散随机变量的函数（Function of Discrete Random Variable）

- 设X是一个离散随机变量，考虑X的函数：$Y = g(X)$，则Y也是一个离散随机变量
- 若已知X的概率质量函数$p_X(x)$，则可以获得Y的概率质量函数：
  $p_Y(y) = \sum_{\{x | g(x) = y\}} p_X(x)$

2.7 离散随机变量的期望（Expectation of Discrete Random Variable）

- 定义离散随机变量X的期望为：
  $E[X] = \sum_x x p_X(x)$
- 期望的本质是：X所有可能取值的加权平均数

2.8 离散随机变量的函数的期望法则（Expected Value Rule for Functions of Discrete Random Variables）

- X 是离散随机变量，令$g(X)$ 是X 的一个函数，则随机变量$g(X)$ 的期望为：
  $E[g(X)] = \sum_x g(x) p_X(x)$

2.9 矩（Moment）

- 定义离散随机变量X 的n 次矩为：
  $E[X^n] = \sum_x x^n p_X(x)$

2.10 方差（Variance）

- 定义离散随机变量X 的方差为随机变量$(X - E[X])^2$ 的期望，即：
  $$\begin{aligned} var(X) &= E[(X - E[X])^2] \\ &= \sum_x (x - E[X])^2 p_X(x) \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

2.11 标准差（Standard Deviation）

- 定义离散随机变量X 的标准差为：
  $\sigma_X = \sqrt{var(X)}$

2.12 期望与方差的属性（Properties of Mean and Variance）

- X 为随机变量，$Y = aX + b$ 是X 的一个线性函数，a，b均为标量，则：
  $E[Y] = aE[X] + b$
  $var(Y) = a^2 var(X)$

2.13 常见离散随机变量的期望与方差举例

1. 伯努利随机变量
   $E[X] = p$
   $E[X^2] = p$
   $var(X) = E[X^2] - (E[X])^2 = p(1-p)$
2. 离散均匀随机变量
   $E[X] = \frac{a+b}{2}$
   $var(X) = \frac{(b-a)(b-a+12)}{12}$
3. 泊松随机变量
   $E[X] = \lambda$
   $var(X) = \lambda$

2.14 多个离散随机变量的联合概率质量函数（Joint PMF of Multiple Discrete Random Variables）

- X 和Y 为与同一个实验相关的两个随机变量，则X 与Y 的联合概率质量函数$p_{X,Y}$ 定义为：
  $p_{X,Y}(x, y) = P(\{X = x\} \cap \{Y = y\}) = P(X = x, Y = y)$
- 可以由联合概率质量函数获得边际概率质量函数（marginal PMF）：
  $p_X(x) = \sum_y p_{X,Y}(x, y)$
  $p_Y(y) = \sum_x p_{X,Y}(x, y)$
- X 与Y 的函数$g(X, Y)$ 定义了一个随机变量，有：

$E[g(X,Y)] = \sum_x \sum_y g(x,y) p_{X,Y}(x,y)$

- 如果g是线性函数，并有形式：$aX+bY+c$ 则：
  $E[aX+bY+c] = aE[X]+bE[Y]+c$

2.15 条件于事件的离散随机变量（Conditioning a Discrete Random Variable on an Event）

- 给定事件A，且知$P(A)>0$，离散随机变量X 的条件概率质量函数为：
  $p_{X|A}(x) = P(X=x|A) = \frac{P(\{X=x\} \cap A)}{P(A)}$
  因为：
  $P(A) = \sum_x P(\{X=x\} \cap A)$
  因此有：
  $\sum_x p_{X|A}(x) = 1$
- 若事件$A_1, A_1 \cdots, A_n$ 是独立事件，构成样本空间的一个划分，且知$\forall i, P(A_i)>0$, 则：
  $p_X(x) = \sum_{i=1}^n P(A_i) p_{X|A_i}(x)$
  这是全概率定理的一个特例，若进一步知$\forall B, \forall i, P(A_i \cap B)>0$, 则有：
  $p_{X|B}(x) = \sum_{i=1}^n P(A_i|B) p_{X|A_i \cap B}(x)$

2.16 条件于随机变量的离散随机变量（Conditioning one Discrete Random Variable on Another）

- X, Y 是两个离散随机变量，给定$Y=y$ 则，联合概率质量函数为：
  $p_{X,Y}(x,y) = p_Y(y) p_{X|Y}(x|y)$
- 可以利用上面的联合概率质量函数计算变量X 的边际概率质量函数：
  $p_X(x) = \sum_y p_Y(y) p_{X|Y}(x|y)$

2.17 离散随机变量的条件期望（Conditional Expectations of Discrete Random Variable）

- 给定事件$A, P(A)>0$，离散随机变量X 的条件期望定义为：
  $E[X|A] = \sum_x x p_{X|A}(x)$
- 给定事件$A, P(A)>0$，离散随机变量X 的函数$g(X)$ 的条件期望定义为：
  $E[g(X)|A] = \sum_x g(x) p_{X|A}(x)$
- 若事件$A_1, A_1 \cdots, A_n$ 是独立事件，构成样本空间的一个划分，且知$\forall i, P(A_i)>0$, 则：
  $EX = \sum_{i=1}^n P(A_i) E[X|A_i]$
- 给定离散随机变量Y 取值为y，则离散随机变量X 的条件期望为：
  $E[X|Y=y] = \sum_x p_{X|Y}(x|y)$
  若进一步知$\forall B, \forall i, P(A_i \cap B)>0$, 则有：
  $E[X|B] = \sum_{i=1}^n P(A_i|B) E[X|A_i \cap B]$
- $E[X] = \sum_y p_Y(y) E[X|Y=y]$

2.18 离散随机变量的独立性（Independence of Discrete Random Variable）

- X 和Y 为与同一个实验相关的两个随机变量，事件A 满足$P(A)>0$，则：
  若$\forall x$ $p_{X|A}(x) = p_X(x)$ 则称X 独立于事件A
  若$\forall x, y$ $p_{X,Y}(x,y) = p_X(x) p_Y(y)$

- X, Y 是两个随机变量，$g(X)$ 和$h(Y)$分别是X 和Y 的函数，则：
  $E[XY] = E[X]E[Y]$
  $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$
  $var(X+Y) = var(X) + var(Y)$

附：常见离散随机变量小结

1. 离散均匀随即变量，区间$[a, b]$
   概率质量函数：
   $$p_X(k) = \begin{cases} 1/b-a+1 & \text{if } k = a, a+1, \cdots, b \\ 0 & \text{otherwise} \end{cases}$$
   期望：
   $$E[X] = \frac{a+b}{2}$$
   方差：
   $$var(X) = \frac{(b-a)(b-a+12)}{12}$$
2. 伯努利随机变量，参数p （扔一次硬币出现正面朝上）
   概率质量函数：
   $$p_X(k) = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases}$$
   期望：
   $$E[X] = p$$
   方差：
   $$var(X) = E[X^2] - (E[X])^2 = p(1-p)$$
3. 二项，参数p, n （扔n 次硬币，出现正面朝上的次数）
   概率质量函数：
   $$p_X(k) = P(X = k) = \binom{n}{k} = p^k(1-p)^{n-k}, \quad k = 0, 1, \cdots, n$$
   期望：
   $$E[X] = np$$
   方差：
   $$var(X) = np(1-p)$$
4. 几何随机变量，参数p （连续扔硬币直至第一次出现正面朝上）
   概率质量函数：
   $$p_X(k) = (1-p)^{k-1}p, \quad k = 1, 2, \cdots,$$
   期望：
   $$E[X] = \frac{1}{p}$$
   方差：
   $$var(X) = \frac{1-p}{p^2}$$
5. 泊松随机变量，参数 λ（在n 大，p小时用于近似二项随机变量， $\lambda = np$）
   概率质量函数：
   $$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \cdots,$$
   期望：
   $$E[X] = \lambda$$
   方差：

$\mathrm{var}(X) = \lambda$

Probability Notes 3 一般随机变量

2014-04-15

3 一般随机变量（General Random Variable）

3.1 连续随机变量与概率密度函数（Continuous Random Variable and Probability Density Function）

- 称一个随机变量X为连续随机变量，则存在一个非负函数f_X即概率密度函数（PDF for short），该函数对于任意实数轴（real line）上的子集B都满足：
  $P(X \in B) = \int_B f_X(x)dx$
- 若$B = [a, b]$，上式即可写作：
  $P(a \leqslant X \leqslant b) = \int_a^b f_X(x)dx$

3.2 概率密度函数的属性（Properties of PDF）

- 令X是一个连续随机变量，f_X是其概率密度函数

1. $f_X(x) \geqslant 0 \quad for\ all\ x$
2. $\int_{-\infty}^{+\infty} f_X(x)dx = 1$
3. 对于非常小的δ
   $P([x, x+\delta]) \approx f_X(x) \cdot \delta$
4. 对于任意实数轴上的子集B
   $P(X \in B) = \int_B f_X(x)dx$

3.3 连续随机变量的期望（Expectation of a Continuous Random Variable）

- 连续随机变量X，其概率密度函数为f_X，其期望定义为：
  $E[X] = \int_{-\infty}^{+\infty} xf_X(x)dx$

3.4 连续随机变量的期望法则（Expected Value Rule for Functions of Random Variables）

- X是连续随机变量，令$g(X)$是X的一个函数，则$g(X)$的期望为：
  $E[g(X)] = \int_{-\infty}^{+\infty} g(x)f_X(x)dx$

3.5 连续随机变量的方差（Variance of Continuous Random Variable）

- 连续随机变量X的方差定义为：
  $var(X) = E[(X - E[X])^2] = \int_{-\infty}^{+\infty} (x - E[X])^2 f_X(x)dx$
- 并且：
  $0 \leqslant var(X) = E[X^2] - (E[X])^2$

3.6 期望与方差的属性（Properties of Mean and Variance）

- X为随机变量，$Y = aX + b$是X的一个线性函数，a，b均为标量，则：
  $E[Y] = aE[X] + b$
  $var(Y) = a^2 var(X)$

3.7 累积分布函数（Cumulative Distribution Functions）

- 随机变量X 的累积分布函数F_X 定义为：
  $F_X(x) = P(X \leqslant x), \quad for\ all\ x$

3.8 累积分布函数的属性（Properties of CDF）

- 累积分布函数是单调非减函数（monotonically non-decreasing）：
  $if\quad x \leqslant y,\ then\quad F_X(x) \leqslant F_X(y)$
- $\lim_{x \to -\infty} F_X(x) = 0$
  $\lim_{x \to +\infty} F_X(x) = 1$

- 若X 为离散随机变量，则其概率质量函数$F_X(x)$是x的分段常数函数（piecewise constant function）
- 若X 为连续随机变量，则其概率密度函数$F_X(x)$是x的连续函数

- 若X 为离散随机变量，并且X 取值为整数值，则其概率质量函数与累积分布函数可以相互推导
  对概率质量函数求和获得累积分布函数：
  $F_X(k) = \sum_{i=-\infty}^{k} p_X(i)$
  对累积分布函数求差分获得概率质量函数：
  $p_X(k) = P(X \leqslant k) - P(X \leqslant k-1) = F_X(k) - F_X(k-1)$
- 若X 为连续随机变量，则其概率密度函数与累积分布函数可以相互推导
  对概率密度函数积分获得累积分布函数：
  $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$
  对累积分布函数微分获得概率密度函数：
  $f_X(x) = \frac{dF_X}{dx}(x)$

3.9 多个连续随机变量的联合概率密度函数(Joint PDFs of Multiple Continuous Random Variable)

- 连续随机变量X, Y 具有联合概率密度函数$f_{X,Y}$ 可以用来计算事件B 的概率：
  $P((X, Y) \in B) = \iint_{(x,y)\in B} f_{X,Y}(x, y)\,dxdy$
- 可以从联合概率密度函数获得X 或Y 的边际概率密度函数：
  $f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\,dy$
  $f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\,dx$
- 连续随机变量X, Y 的联合累积分布函数定义为：
  $F_{X,Y}(x, y) = P(X \leqslant x, Y \leqslant y)$
- 连续随机变量X, Y 的联合累积分布函数可以通过联合概率密度函数获得：
  $f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x, y)$
- 令g是连续随机变量X 与Y 的一个函数，则$g(X, Y)$ 也是一个连续随机变量，其期望为：
  $E[g(X, Y)] = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{+\infty} g(x, y) f_{X,Y}(x, y)\,dxdy$
  若g是线性函数，具有形式$aX + bY + c$ ，则：
  $E[aX + bY + c] = aE[X] + bE[Y] + c$

3.10 条件于事件的连续随机变量（Conditioning a Continuous Random Variable on an Event）

- 给定事件 A ，且知 P(A)>0 ，连续随机变量 X 的条件概率密度函数满足：
  $P(X \in B|A) = \int_B f_{X|A}(x)dx$
- 若A 是实数轴的一个子集，并且$P(X \in A) > 0$ ，则：
  $f_{X|\{X\in A\}}(x) = \frac{f_X(x)}{P(X\in A)}, \quad if\ x \in A$
  $f_{X|\{X\in A\}}(x) = 0, \quad otherwiese$
- 若事件$A_1, A_1 \cdots, A_n$ 是独立事件，构成样本空间的一个划分，且知$\forall i, P(A_i) > 0$, 则：
  $f_X(x) = \sum_{i=1}^{n} P(A_i) f_{X|A_i}(x)$
  这是全概率定理的一个特例

3.11 条件于随机变量的连续随机变量（Conditioning a Continuous Random Variable on an Random Variable）

- X, Y 是两个连续随机变量，给定$Y = y$ 则，联合概率密度函数为：
  $f_{X,Y}(x, y) = f_Y(y) f_{X|Y}(x|y)$
- 可以利用上面的联合概率密度函数计算变量 X 的边际概率密度函数：
  $f_X(x) = \int_{-\infty}^{+\infty} f_Y y f_{X|Y}(x|y)dy$
- 另外有：
  $P(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx$

3.12 连续随机变量的条件期望（Conditional Expectation of Continuous Random Variable）

- 定义：
- 给定事件$A, P(A) > 0$ ，连续随机变量X 的条件期望定义为：
  $E[X|A] = \int_{-\infty}^{+\infty} x f_{X|A}(x)dx$
- 给定随机变量Y 取值y ，则连续随机变量X 的条件期望定义为：
  $E[X|Y = y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)dx$
- 期望值法则：
- 给定事件$A, P(A) > 0$ ，连续随机变量X 的函数$g(X)$ 的条件期望定义为：
  $E[g(X)|A] = \int_{-\infty}^{+\infty} g(x) f_{X|A}(x)dx$
- 给定随机变量Y 取值y ，连续随机变量X 的函数$g(X)$ 的条件期望定义为：
  $E[g(X)|Y = y] = \int_{-\infty}^{+\infty} g(x) f_{X|Y}(x|y)dx$
- 全期望定理（Total expectation theorem）：
- 若事件$A_1, A_1 \cdots, A_n$ 是独立事件，构成样本空间的一个划分，且知$\forall i, P(A_i) > 0$, 则：
  $E[X] = \sum_{i=1}^{n} P(A_i) E[X|A_i]$
  $E[X] = \int_{-\infty}^{+\infty} E[X|Y = y] f_Y(y)dy$
- 另外有：
  $E[g(X, Y)|Y = y] = \int g(x, y) f_{X|Y}(x|y)dx$

$$E[g(X,Y)] = \int E[g(x,y)|Y=y] f_Y(y)\,dy$$

3.13 连续随机变量的独立性（Independence of Continuous Random Variable）

- 令X 与Y 是两个连续随机变量，称X 与Y 独立，需要：
  $f_{X,Y}(x,y) = f_X(x) f_Y(y), \quad \text{for all } x,y$
- 若两个连续随机变量X 与Y 相互独立，则：
  $E[XY] = E[X]E[Y]$
  $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$
  $\text{var}(XY) = \text{var}(X) + \text{var}(Y)$

3.14 连续随机变量的贝叶斯法则（Bayes' Rule for Continuous Random Variables）

- 令Y 是一个连续随机变量
- 若X 是一个连续随机变量，则：
  $f_Y(y) f_{X|Y}(x|y) = f_X(x) f_{Y|X}(y|x)$

  因而有如下贝叶斯法则

$$f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{f_Y(y)} = \frac{f_X(x) f_{Y|X}(y|x)}{\int_{-\infty}^{+\infty} f_X(t) f_{Y|X}(y|t)\,dt}$$

- 若N 是一个离散随机变量，则：
  $f_Y(y) P(N=n|Y=y) = p_N(n) f_{Y|N}(y|n)$

  因而有如下贝叶斯法则：

$$P(N=n|Y=y) = \frac{p_N(n) f_{Y|N}(y|n)}{f_Y(y)} = \frac{p_N(n) f_{Y|N}(y|n)}{\sum_i p_N(i) f_{Y|N}(y|i)}$$

- 以及：

$$f_{Y|N}(y|n) = \frac{f_Y(y) P(N=n|Y=y)}{p_N(n)} = \frac{f_Y(y) P(N=n|Y=y)}{\int_{-\infty}^{+\infty} f_Y(t) P(N=n|Y=t)\,dt}$$

- 类似地$P(A|Y=y)$ 和$f_{Y|A}(y)$也有相应贝叶斯法则

附：常见连续随机变量小结

1. 连续均匀随机变量，区间$[a,b]$

   概率密度函数：

   $f_X(x) = \frac{1}{b-a}, \quad \text{if } a \leqslant x \leqslant b$

   $f_X(x) = 0, \quad \text{otherwise}$

   期望：

   $E[X] = \frac{a+b}{2}$

   方差：

   $\text{var}(X) = \frac{(b-a)^2}{12}$

2. 指数随机变量，参数 λ

   概率密度函数：

$f_X(x) = \lambda e^{-\lambda x}, \quad if \ x \geqslant 0$
$f_X(x) = 0, \quad otherwise$

累积分布函数：

$F_X(x) = 1 - e^{-\lambda x}, \quad if \ x \geqslant 0$
$F_X(x) = 0, \quad otherwise$

期望：

$E[X] = \frac{1}{\lambda}$

方差：

$var(X) = \frac{1}{\lambda^2}$

3. 正态随机变量，参数 μ， σ^2

概率密度函数：

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

期望：

$E[X] = \mu$

方差：

$var(X) = \sigma^2$

Probability Notes 4 随机变量 续

2014-04-15

4 随机变量更多话题（Further Topics on Random Variables）

4.1 分布的推导（Derived Distributions）

- X 是连续随机函数，Y 是X 的函数$Y = g(X)$ ，若要计算Y 的概率密度函数：

1. 计算Y 的累积分布函数：
   $F_Y(y) = P(g(X) \leqslant y) = \int_{\{x \mid g(x) \leqslant y\}} f_X(x)\,dx$
2. 微分获得概率密度函数：
   $f_Y(y) = \frac{dF_Y}{dy}(y)$

- 若$Y = g(X)$ 是线性函数，具有形式$Y = aX + b$ ，其中a，b为标量且$a \neq 0$，则：
  $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$

- 若$Y = g(X)$ 是连续随机函数X 的严格单调函数，则必然有反函数h满足$x = h(y)$
- 假设h是可微的，则：
  $f_Y(y) = f_X(h(y)) \left| \frac{dh}{dy}(y) \right|$

4.2 协方差（Covariance）

- 随机变量X 与Y 的协方差定义为：
  $$\begin{aligned} \operatorname{cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$
- 若$\operatorname{cov}(X, Y) = 0$ ，则X 与Y 是不相关的（uncorrelated）
- 若X 与Y 相互独立，则$\operatorname{cov}(X, Y) = 0$
- $\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X, Y)$

4.3 相关性系数（Correlation Coefficient）

- 随机变量X 与Y 的相关性系数定义为：
  $\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}$
- 相关性系数满足：
  $-1 \leqslant \rho(X, Y) \leqslant 1$

4.4 条件期望和方差的属性（Properties of the Conditional Expectation and Variance）

- $E[X \mid Y = y]$ 是一个取值依赖于y 的值
- $E[X \mid Y]$是随机变量Y 的一个函数，因此也是一个随机变量，在Y 取值为y时，其值为$E[X \mid Y = y]$
- 迭代期望定律（Law of Iterated Expectations）
  $E[E[X \mid Y]] = E[X]$

- $\mathrm{var}(X|Y)$是随机变量Y的一个函数，因此也是一个随机变量，在Y取值为y时，其值为$\mathrm{var}(X|Y=y)$
- 全方差定律（Law of Total Variance）
  $\mathrm{var}(X)=E[\mathrm{var}(X|Y)]+\mathrm{var}(E[X|Y])$
- $E[X|Y=y]$可以被视为在给定$Y=y$情况下，X取值的一个估计值（estimate）
  $E[X|Y]-X$是估计误差，该误差是一个期望为0，且与$E[X|Y]$独立的随机变量

Probability Notes 5 极限定理

2014-04-22

5 极限定理（Limit Theorem）

5.1 马尔可夫不等式（Markov Inequality）

- 设随机变量X只可取非负值，则：
  $P(X \geqslant a) \leqslant \frac{E[X]}{a}$, for all $a > 0$

5.2 契比雪夫不等式（Chebyshev Inequality）

- 设随机变量X期望为μ，方差为σ^2，则：
  $P(|X - \mu| \geqslant c) \leqslant \frac{\sigma^2}{c^2}$ for all $c > 0$

5.3 弱大数定律（Weak Law of Large Numbers, WLLN）

- 设$X_1, X_2, \cdots, X_n$是独立同步分布（independent identically distributed, i.i.d.）的随机变量，共同的期望为μ，则：
  对于任意$\epsilon > 0$有：
  $P(|M_n - \mu| \geqslant \epsilon) = P(|\frac{X_1+X_2+\cdots+X_n}{n} - \mu| \geqslant \epsilon) \to 0, \quad as \ n \to \infty$

5.4 概率收敛（Convergence in Probability）

- 设$Y_1, Y_2 \cdots Y_n$是随机变量的一个数列（sequence），且a为常数。若对于任意$\epsilon > 0$均有：
  $\lim_{n\to\infty} P(|Y_n - a| \geqslant \epsilon) = 0$
  则称数列Y_n依概率收敛于a

5.5 中央极限定理（Central Limit Theroem）

- 设$X_1, X_2, \cdots, X_n$是独立同步分布的随机变量，共同的期望为μ，方差为σ^2，定义标准值Z_n为：
  $Z_n = \frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$
- 则Z_n的累积分布函数收敛于标准正态累积分布函数：
  $\lim_{n\to\infty} P(Z_n \leqslant z) = \Phi(z)$, for every$z$
  $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx$

5.6 德莫佛-拉普拉斯二项分布近似公式（De Moivre-Laplace Approximation to the Binomial）

- 设S_n是二项随机变量，其参数为n和p，n较大且k，l为非负整数时有：
  $P(k \leqslant S_n \leqslant l) \approx \Phi(\frac{l+\frac{1}{2}-np}{\sqrt{np(1-p)}}) - \Phi(\frac{k-\frac{1}{2}-np}{\sqrt{np(1-p)}})$

5.7 强大数定律（Strong Law of Large Numbers）

- 设$X_1, X_2, \cdots, X_n$是独立同步分布的随机变量，共同的期望为μ，则：
  $P(\lim_{n\to\infty} \frac{X_1+X_2+\cdots+X_n}{n} = \mu) = 1$

5.8 依概率1收敛

- 设$Y_1, Y_2 \cdots Y_n$是随机变量的一个数列（sequence），且c为常数。若：

$P(\lim_{n\to\infty} Y_n = c) = 1$

则称数列Y_n依概率1收敛于c

Probability Notes 6 伯努利过程和泊松过程

2014-05-04

6 伯努利过程和泊松过程（Bernoulli Processes and Poisson Processes）

6.1 伯努利过程（Bernoulli Process）

- 伯努利过程是互相独立的伯努利随机变量$X_1, X_2, \cdots$的一个序列，随机变量X_i满足：
  $P(X_i = 1) = P(\text{success at the ith trail}) = p$
  $P(X_i = 0) = P(\text{failure at the ith trail}) = 1 - p$

6.2 与伯努利过程有关的随机变量

- 二项随机变量（参数n和p），描述n次试验中成功的次数S的概率：
  $$p_S(k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \cdots, n$$
  $$E[S] = np$$
  $$\text{var}(S) = np(1-p)$$
- 几何随机变量（参数p），描述直至T次试验才出现第一次成功的概率：
  $$p_T(t) = (1-p)^{t-1} p, \qquad t = 1, 2, \cdots$$
  $$E[T] = \frac{1}{p}$$
  $$\text{var}(T) = \frac{1-p}{p^2}$$

6.3 伯努利过程的独立性属性

- $X_1, X_2, \cdots$是一个伯努利过程，给定时间n，则该过程的未来（随机变量序列$X_{n+1}, X_{n+2}, \cdots$）也是一个伯努利过程，并且与过程的过去（$X_1, X_2, \cdots, X_n$）相互独立
- $X_1, X_2, \cdots$是一个伯努利过程，给定时间n，令在此后首次出现成功的时间为T，则$T - n$满足几何分布，参数为p，并且与随机变量$X_1, \cdots, X_n$独立

6.4 伯努利过程的另一种描述

- 从相互独立且共同参数为p的一个几何随机变量序列$T_1, T_2, \cdots$开始，将这些随机变量理解为两次成功试验之间的时间间隔，即：
  在$T_1, T_1 + T_2, T_1 + T_2 + T_3$ 这些时刻的试验是成功的，其余时刻均为失败

6.5 第k次成功时刻

- 第k次试验成功的时刻Y_k等于前k次成功试验之间的时间间隔之和，即：
  $$Y_k = T_1 + T_2 + \cdots + T_k$$
- 随机变量Y_k的分布成为k阶帕斯卡（Pascal）分布，其概率质量函数为：
  $$p_{Y_k}(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \qquad t = k, k+1, \cdots$$
  期望与方差分别为：
  $$E[Y_k] = E[T_1] + \cdots + E[T_k] = \frac{k}{p}$$
  $$\text{var}(Y_k) = \text{var}(T_1) + \cdots + \text{var}(T_k) = \frac{k(1-p)}{p^2}$$

6.6 二项分布的泊松近似（Poisson Approximation to the Binomial）

- 参数为λ的泊松随机变量Z的概率质量函数为：

$p_Z(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \cdots$

期望和方差为：

$E[Z] = \lambda$

$var(Z) = \lambda$

- 给定一个非负整数k，令$p = \frac{\lambda}{n}$，则在$n \to \infty$时二项随机变量概率质量函数 $p_S(k) = \frac{n!}{(n-k)!k!}p^k(1-p)^{n-k}$ 收敛于$p_Z(k)$
- 一般地，n较大且p较小的情况下，泊松随机变量的概率质量函数是二项随机变量概率质量函数的一个很好的近似

6.7 泊松过程

- 一个到达过程（arrival process）若要是一个柏松过程，则需要满足如下属性：

1. 时间齐性（Time-homogeneity）：
   对于任意长度为τ的时间区间，发生k次到达的概率$P(k, \tau)$相同
2. 独立性：
   任何一个时间区间内的到达次数与区间外的到达情况无关
3. 小区间属性：
   在$\lim_{\tau \to 0}$时，$P(k, \tau)$满足：
   $P(0, \tau) = 1 - \lambda\tau + o(\tau)$
   $P(1, \tau) = \lambda\tau + o_1(\tau)$
   $P(k, \tau) = o_k(\tau), \quad for\ k = 2, 3, \cdots$
   其中$o(\tau), \cdots o_k(\tau)$为$\tau$的函数并且满足：
   $\lim(\tau \to 0)\frac{o(\tau)}{\tau} = 0$
   $\lim(\tau \to 0)\frac{o_k(\tau)}{\tau} = 0$

6.8 与柏松过程有关的随机变量

- 柏松随机变量，参数λ、τ描述一个到达频率为λ的柏松过程中，任意一个间隔为τ的时间区间中到达发生的次数N_τ

  概率质量函数$p_{N_\tau}(k) = P(k, \tau) = e^{-\lambda\tau}\frac{(\lambda\tau)^k}{k!}, \quad k = 0, 1, \cdots$

  期望$E[N_\tau] = \lambda\tau$

  方差$var(N_\tau) = \lambda\tau$
- 指数随机变量，参数λ，描述直至第一次到达所需要的时间T

  概率密度函数$f_T(t) = \lambda e^{-\lambda t}, \quad t \geqslant 0$

  期望$E[T] = \frac{1}{\lambda}$

  方差$var(T) = \frac{1}{\lambda^2}$

6.9 柏松过程的独立性

- 对于一个柏松过程，给定时刻$t > 0$，则t之后的过程也是一个柏松过程，并且独立于t时刻之前的过程
- 令t是一个给定时刻，T是时刻t之后第一次到达的时刻，则$T - t$满足参数为λ的指数分布，并且独立于t时刻之前的过程

6.10 柏松过程的另一种描述

- $T_1, T_2, \cdots$,是相互独立且为具有共同参数 λ的指数随机变量序列，$T_1, T_2, \cdots$表示各次到达之间的时间间隔
- 到达发生在时刻$T_1, T_1+T_2, T_1+T_2+T_3 \cdots$

6.11 第k次成功时刻

- k阶厄兰随机变量（Erlang of order k），描述第k次到达时间Y_k，Y_k等于前k次到达之间的时间间隔的和
  $Y_k = T_1 + T_2 + \cdots + T_k$
- $E[Y_k] = E[T_1] + \cdots E[T_k] = \frac{k}{\lambda}$
- $var(Y_k) = var(T_1) + \cdots + var(T_K) = \frac{k}{\lambda^2}$
- $f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geqslant 0$

6.12 随机个随机变量之和的属性

- $N, X_1, X_2, \cdots$是相互独立的随机变量，N 取值为非负整数，令$Y = X_1 + X_2 \cdots + X_N$
- X_i为参数为p的伯努利随机变量
  N 为参数为m, q 的二项随机变量
  Y 为参数为m, pq 的二项随机变量
- X_i为参数为p的伯努利随机变量
  N 为参数为 λ的柏松随机变量
  Y 为参数为 λp的柏松随机变量
- X_i为参数为p的几何随机变量
  N 为参数为q的几何随机变量
  Y 为参数为pq的几何随机变量
- X_i为参数为 λ的指数随机变量
  N 为参数为q的几何随机变量
  Y 为参数为 λq的指数随机变量

Probability Notes 7 马尔可夫链

2014-05-12

7 马尔可夫链 （Markov Chains）

7.1 马尔可夫模型 （Markov Models）

- 一个马尔可夫链模型通过如下三点定义：

1. 状态的集合（set of states）$S = \{1, 2, \cdots, m\}$
2. 可能的转变的集合（set of possible transitions），集合的元素是满足$p_{ij} > 0$ 的状态对(i, j)
3. p_{ij}的数值

- 上述马尔可夫链模型所定义的马尔可夫链是随机变量$X_0, X_1, X_2, \cdots$的序列，变量从状态集合S 中取值，并且满足，对于任何时间n、任意状态i，和任意可能的序列$i_0, \cdots, i_{n-1}$有：
  $P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \cdots, X_o = i_0) = p_{ij}$

7.2 n步转变概率：察普曼-科莫高洛夫方程式（Chapman-Kolmogorov Equation for the n-Step Transition Probabilities）

- 经过n次转变后进入某一状态的概率可以用如下递归公式进行计算：
  $r_{ij}(n) = \sum_{k=1}^{m} r_{ik}(n-1)p_{kj}, \qquad \text{for } n > 1, \text{ and all } i, j$
- $n = 1$ 时，$r_{ij}(1) = p_{ij}$

7.3 马尔可夫链的分解（Markov Chain Decomposition）

- 一个马尔可夫链可以被分解成为一个或多个循环类（recurrent classes）和可能一个或多个过渡状态（transient states）
- 任何一个循环状态都可以被其所属的循环类内所有其他状态转变而成，但是不能从其他循环类内的任意状态转变而来
- 过度状态不能从任何循环状态转变而成
- 给定一个过度状态，从其进行转变则至少能到达一个循环状态

7.4 周期性（Periodicity）

- 考虑一个循环类R ：

1. 如果类中的状态可以被划分为$d > 1$ 个互斥子集$S_1, \cdots, S_d$，使得S_k 中所有的转变都指向S_{k+1} （如果$k = d$ 则指向S_1），则称诶这个类具有周期性的
2. 当且仅当存在一个时刻n，使得$r_{ij}(n) > 0, \text{ for all } i, j \in R$，则称这个类是非周期的

7.5 稳态收敛定理（Steady-State Convergence Theorem）

- 考虑一个包含了一个具有周期性的循环类的马尔可夫链，则状态j所有的稳态概率 π_j具有如下属性：

1. 对于任意状态j，有：
   $\lim_{n\to\infty} r_{ij}(n) = \pi_j, \qquad \text{for all } i$
2. π_j是下面方程组的唯一解：
   $\pi = \sum_{k=1}^{m} \pi_k p_{kj}, \qquad j = 1, 2, \cdots, m$
   $1 = \sum_{k=1}^{m} \pi_k$

3. 对于任意过度状态j有：$\pi_j = 0$ ；对于任意循环状态j有：$\pi_j > 0$

7.6 稳态概率和期望状态频率（Steady-State Probabilities as Expected State Frequencies）

- 考虑仅包含单一一个非周期类的马尔可夫链，稳态概率 π_j满足：
  $$\pi_j = \lim_{n\to\infty} \frac{v_{ij}(n)}{n}$$
- 其中$v_{ij}(n)$是在从状态i开始的前n次转变中状态j出现次数的期望值

7.7 特定转变的期望频率（Expected Frequency of a Particular Transition）

- 考虑仅包含单一一个非周期类的马尔可夫链，从一个给定的初始状态开始，经过n次转变。令$q_{jk}(n)$是这样的从状态j到k的转变种类的期望值。无论给定的初始状态为何，均有：
  $$\lim_{n\to\infty} \frac{q_{jk}(n)}{n} = \pi_j p_{jk}$$

7.8 吸收概率方程（Absorption Probability Equations）

- 考虑一个马尔可夫链，其中的状态不是过渡状态就是吸收状态，给定一个吸收状态s，则从状态i开始，最后终于到达状态s的概率a_i是如下方程组的唯一解：
  $$a_s = 1$$
  $$s_i = 0, \quad \text{for all absorbing } i \neq s$$
  $$a_i = \sum_{j=1}^{m} p_{ij} a_j, \quad \text{for all transient } i$$

7.9 吸收时间期望方程（Equations for the Expected Time to Absorption）

- 到达吸收的时间的期望值 $\mu_1, \cdots, \mu_m$ 是下面方程组的唯一解：
  $$\mu_i = 0, \quad \text{for all recurrent states } i$$
  $$\mu_i = 1 + \sum_{j=1}^{m} p_{ij}\mu_j \quad \text{for all transient states } i$$

7.10 首次通过和再次出现时间平均值的方程（Equations for Mean First Passage and Recurrence Times）

- 考虑一个只包含一个循环类的马尔可夫链，令s是一个特定的循环状态：

1. 从状态i开始，首次转变至状态s所需时间t_i的期望是下面方程组的唯一解：
   $$t_s = 0$$
   $$t_i = 1 + \sum_{j=1}^{m} p_{ij} t_j, \quad \text{for all } i \neq s$$
2. 状态s的再次出现时间t_s^*的期望值为：
   $$t_s^* = 1 + \sum_{j=1}^{m} p_{sj} t_j$$

Probability Notes 8 贝叶斯统计推断（Bayesian Statistical Inference）

2014-04-17

8 贝叶斯统计推断（Bayesian Statistical Inference）

8.1 重要概念

- 贝叶斯统计（Bayesian Statistics）将一个未知的参数当作一个已知先验分布（prior distribution）的随机变量来处理。
- 在参数估计（parameter estimation）中，希望生成与参数真值接近的估计值
- 在假说检验（hypothesis testing）中，未知的参数可以取有限种可能值，每个取值对应于一种假说，希望选择一个假说使得误差的概率最小
- 主要贝叶斯推断方法（Principal Bayesian Inference Methods）：

1. 后验概率最大法则（Maximum a Posteriori Probability Rule, MAP Rule）：
   在所有可能的参数值的估计值或假说中针对手头的数据选择一个可使条件/后验概率最大的参数
2. 最小均方估计（Least Mean Squares Estimation, LMS Estimation）：
   针对手头的数据选择一个使得参数值和参数估计值之间误差的均方最小的估计量或函数
3. 线性最小均方估计（Linear Least Mean Squares Estimation, LLMS Estimation）：
   针对手头的数据选择一个使得参数值和参数估计值之间误差的均方最小的线性函数

8.2 贝叶斯推断（Bayesian Inference）

- 从对未知随机变量 Θ 的先验分布 p_θ 或 f_θ 开始
- 对观测 X （向量）建模 $p_{X|\Theta}$ 或 $f_{X|\Theta}$
- 在获得具体的观测值 x 后，通过合适的贝叶斯法则构建 Θ 的后验分布

8.3 贝叶斯法则

1. Θ 离散，X 离散：
   $$p_{\Theta|X} = \frac{p_\Theta(\theta)p_{X|\Theta}(x|\theta)}{\sum_{\theta'} p_\Theta(\theta)p_{X|\Theta}(x|\theta)}$$
2. Θ 离散，X 连续：
   $$p_{\Theta|X} = \frac{p_\Theta(\theta)f_{X|\Theta}(x|\theta)}{\sum_{\theta'} p_\Theta(\theta)f_{X|\Theta}(x|\theta)}$$
3. Θ 连续，X 离散：
   $$f_{\Theta|X} = \frac{f_\Theta(\theta)p_{X|\Theta}(x|\theta)}{\int f_\Theta(\theta)p_{X|\Theta}(x|\theta)d\theta}$$
4. Θ 连续，X 连续：
   $$f_{\Theta|X} = \frac{f_\Theta(\theta)f_{X|\Theta}(x|\theta)}{\int f_\Theta(\theta)f_{X|\Theta}(x|\theta)d\theta}$$

8.4 后验概率最大法则

- 给定观测值 x，后验概率最大法则在 θ 中选择一个可使后验分布 $p_{\Theta|X}(\theta|x)$ 或 $f_{\Theta|X}(\theta|x)$ 最大的 θ
- 等价于：选择 θ 使得如下量最大

1. Θ 离散，X 离散：

$p_\Theta(\theta)p_{X|\Theta}(x|\theta)$

2. Θ离散,X 连续：
   $p_\Theta(\theta)f_{X|\Theta}(x|\theta)$
3. Θ连续,X 离散：
   $f_\Theta(\theta)p_{X|\Theta}(x|\theta)$
4. Θ连续,X 连续：
   $f_\Theta(\theta)f_{X|\Theta}(x|\theta)$

- 如果 Θ只可在有限个值中选择，则后验概率最大法则使得选择错误假说的概率最低

8.5 点估计值（Point Estimates）

- 估计量（Estimator）是一个随机变量，其形式为 $\hat{\Theta}=g(X)$ ，是X 的函数
  选择不同的函数g，即是在选择不同的估计量
- 估计值（Esitmate）是估计量的一个具体值，是根据获得的X 的观测值x而确定的值
- 给定X 的一个具体观测值x，后验概率最大估计量将选择估计值 $\hat{\theta}$使得后验概率分布最大
- 给定X 的一个具体观测值x，条件期望估计量将选择估计值 $\hat{\theta}$为$E[\Theta|X=x]$

8.6 假说检验的后验概率最大法则（The MAP Rule for Hypothesis Testing）

- 给定X 的一个具体观测值x，后验概率最大法则选择一个假说H_i使得后验概率$P(\Theta=\theta_i|X=x)$ 最大
- 等价于，选择假说H_i使得$p_\Theta(\theta_i)p_{X|\Theta}(x|\theta_i)$或$p_\Theta(\theta_i)f_{X|\Theta}(x|\theta_i)$最大
- 后验概率最大法则使得在给定观测x的情况下，选择错误假说的概率最低

8.7 最小均方估计

- 在没有任何观测的情况下，选择 $\hat{\theta}=E[\Theta]$可使的$E[(\Theta-\hat{\theta})^2]$最小：
  $E[(\Theta-E[\Theta])^2]\leqslant E[(\Theta-\hat{\theta})^2],\quad \text{for all } \hat{\theta}$
- 给定一个观测值x，选择 $\hat{\theta}=E[\Theta|X=x]$ 可使的$E[(\Theta-\hat{\theta})^2|X=x]$最小：
  $E[(\Theta-E[\Theta]|X=x)^2|X=x]\leqslant E[(\Theta-\hat{\theta})^2|X=x],\quad \text{for all } \hat{\theta}$
- 在所有对于Θ的估计量$g(X)$ 中，选择$g(X)=E[\Theta|X]$可使均方估计误差$E[(\Theta-g(X))^2]$最小：
  $E[(\Theta-E[\Theta|X])^2]\leqslant E[(\Theta-g(X))^2],\quad \text{for all estimators } g(X)$

8.8 估计误差的属性

- 估计误差 $\tilde{\Theta}$是无偏倚的（unbiased），即：
  $\tilde{\Theta}=0,\quad E[\tilde{\Theta}|X=x]=0,\quad \text{for all } x$
- 估计误差 $\tilde{\Theta}$与估计值 $\hat{\Theta}$是不相关的（uncorrelated）：
  $\text{cov}(\tilde{\Theta},\hat{\Theta})=0$
- Θ的方差可以被分解为：
  $\text{var}(\Theta)=\text{var}(\hat{\Theta})+\text{var}(\tilde{\Theta})$

8.9 线性最小均方估计公式

- 根据观测X，对未知参数Θ的线性最小均方估计量$\hat{\Theta}$为：
  $$\hat{\Theta} = E[\Theta] + \frac{cov(\Theta, X)}{var(X)}(X - E) = E + \rho \frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$
- 其中：
  $$\rho = \frac{cov(\Theta, X)}{\sigma_\Theta \sigma_X}$$
  为相关性系数（correlation coefficient）
- 线性最小均方估计误差为：
  $$(1 - \rho^2)\sigma_\Theta^2$$

Probability Notes 9 经典统计

2014-04-25

9 经典统计（Classical Statistics）

9.1 重要概念

- 经典统计将未知的参数当作常数处理，对于参数值的每一个估计值都对应于一个模型
- 在参数估计（parameter estimation）中，我们希望生成在未知参数任何取值情况下均尽可能正确的估计
- 在假说检验（hypothesis testing）中，未知参数只可取有限种可能值，对应于相应数量个假说，我们希望选择一个假说，使得错误的概率较小
- 在显著性检验（significance testing）中，对于一个特定的假说，我们希望决定是接受还是否定该假说，希望决定错误的概率较小
- 经典的推断方法有：

1. 最大似然估计（Maximum likelihood estimation, ML）：选择能使得获得手头所有数据的可能性最大的参数值
2. 线性回归（Linear Regression）：发现适合手头数据的线性关系，使得模型和数据间差异的平方和最小
3. 似然比值检验（Likelihood ratio test）：给定两个假说，根据两个假说的似然比例确定选择哪一个
4. 显著性检验（Significance testing）：给定一个假说，当且仅当观察到的数据在特定的否定区域内时否定该假说

9.2 有关估计量（estimator）的一些术语

- 令 Θ_n 是未知参数 θ 的一个估计量（estimator），即 Θ_n 是 n 个观测值 $X_1, X_2, \cdots, X_n$ 的一个函数，且 Θ_n 的分布取决于 θ
- Θ_n 的估计误差（estimation error）Θ_n 定义为：$\Theta_n = \Theta_n - \theta$
- 估计量的偏倚（bias）记为 $b_\theta(\Theta_n)$ 是估计误差的期望值：$b_\theta(\Theta_n) = E_\theta[\Theta_n] - \theta$
- 估计量 Θ_n 的期望、方差和偏倚都取决于 θ，而估计误差 Θ_n 还额外取决于观测值 $X_1, X_2, \cdots, X_n$
- 若 $\forall \theta$，$E_\theta[\Theta_n]$ 则称 Θ_n 是无偏倚（unbiased）的估计量
- 若 $\forall \theta$，$\lim_{n\to\infty} E_\theta[\Theta_n]$ 则称 Θ_n 是渐近无偏倚（asymptotically unbiased）的估计量
- 若 $\forall \theta$，Θ_n 依概率收敛于参数值 θ，则称 Θ_n 是一致的（consistent）

9.3 最大似然估计（Maximum likelihood estimation）

- 我们手头已有数据是依据概率密度函数 $f_X(x;\theta)$ 或概率质量函数 $p_X(x;\theta)$ 分布的随机向量 $X = (X_1, \cdots, X_n)$ 的一个实现 $x = (x_1, \cdots, x_n)$
- 最大似然估计是选择一个 θ 值，使得似然函数 $p_X(x;\theta)$ 或 $f_X(x;\theta)$ 最大
- 若 h 是 θ 的一个一一映射的函数，则对 h 的最大似然估计是 $h(\hat{\theta}_n)$，其中 $\hat{\theta}_n$ 是 θ 的最大似然估计

9.4 随机变量期望与方差的估计

- 设 $X_1, X_2, \cdots, X_n$ 是独立同步分布（i.i.d.）的随机变量，共同的期望为 θ，方差为 v，两者均未知
- 样本期望为：$M_n = \frac{X_1+X_2+\cdots+X_n}{n}$ 样本期望是 θ 的一个无偏倚的估计量，其均方差（mean

squared error, MSE）是：$\frac{v}{n}$

- 方差的估计量为：$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2$　S_n^2 是偏倚但渐近无偏倚的
  $\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$　$\hat{S}_n^2$ 是无偏倚的

9.5 置信区间（Confidence Intervals）

- 未知标量参数 θ 的置信区间是一个端点为 $\hat{\Theta}_n^-$ 和 $\hat{\Theta}_n^+$ 并且有高概率包含 θ 的区间
- $\hat{\Theta}_n^-$ 和 $\hat{\Theta}_n^+$ 也是随机变量，并且取决于手头的数据 $X_1, X_2, \cdots, X_n$
- 一个 $1-\alpha$ 置信区间满足：$\forall\ \theta,\quad P_\theta(\hat{\Theta}_n^- \leqslant \theta \leqslant \hat{\Theta}_n^+) \geqslant 1-\alpha$

9.6 线性回归

- 给定 n 个数据 (x_i, y_i)，可以使残差平方和（sum of squared residuals, SSR）最小的估计是：
  $\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sum_{i=1}^{n}(x_i - barx)^2}$　$\hat{\theta}_0 = x - \hat{\theta}_1 x$　其中：$x = \frac{1}{n}\sum_{i=1}^{n}x_i$，$y = \frac{1}{n}\sum_{i=1}^{n}y_i$

9.7 贝叶斯线性回归

- 模型：

1. 假设线性关系 $Y_i = \Theta_0 + \Theta_1 x_i + W_i$
2. x_i 是已知的常数（手头的数据）
3. 随机变量 Θ_0，Θ_1，W_1，W_2，$\cdots W_n$ 是正态随机变量，并且相互独立
4. 随机变量 Θ_0，Θ_1 期望为 0 方差分别为 σ_0^2 和 σ_1^2
5. 随机变量 W_i 期望为 0，方差为 σ^2

- 估计量公式：

1. 给定数据 (x_i, y_i)，Θ_0 和 Θ_1 的后验概率最大（MAP）估计为：
   $\hat{\theta}_1 = \frac{\sigma_1^2}{\sigma^2 + \sigma_1^2\sum_{i=1}^{n}(x_i - x)^2}\cdot\sum_{i=1}^{n}(x_i - x)(y_i - y)$　$\hat{\theta}_0 = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}(y - \hat{\theta}_1 x)$
2. 其中：$x = \frac{1}{n}\sum_{i=1}^{n}x_i$　$y = \frac{1}{n}\sum_{i=1}^{n}y_i$

9.8 似然比例检验（Likelihood Ratio Test）

- 从目标错误拒绝概率（false rejection probability）α 开始
- 选择一个 ξ 值，使得错误拒绝概率等于 α　$P(L(X) > \xi ; H_0) = \alpha$
- 一旦获得了 X 的观测数据，若 $L(x) > \xi$ 则拒绝零假说 H_0

9.9 显著性检验（Significance Testing）

- 根据观测值 $X_1, \cdots X_n$，对假说 H_0：$\theta = \theta$ 进行统计检验
- 在获得观测数据之前进行如下步骤：

1. 选择统计量（statistic）S，即一个能够总结手头数据的标量随机变量，通常涉及一个函数 h：$R^n \to R$，统计量 $S = h(X_1 \cdots X_n)$
2. 判断拒绝区域的形状，即将能够拒绝零假说 H_0 的 S 的值写成未知变量 ξ 的一个函数

3. 选择显著性级别，即想要的错误拒绝概率 α
4. 选择关键值 c 使得错误拒绝概率等于 α

- 一旦获得 $X_1, X_2, \cdots, X_n$ 的观测值 $x_1, x_2, \cdots, x_n$：

1. 计算统计量 S 的值 $s = h(x_1, x_2, \cdots, x_n)$
2. 如果 s 属于拒绝区域，则拒绝零假说 H_0

9.10 卡方检验（The Chi-Square Test）

- 选择统计量 $S = \sum_{k=1}^{m} N_k \log\left(\frac{N_k}{n q_k}\right)$ 或相应的 T 统计量
- 拒绝区域为 $2S > \gamma$ 或 $T > \gamma$
- 关键值 γ 从自由度 $m-1$ 的卡方分布 x^2 的累积分布函数值查表获得，使得：
  $P(2S > \gamma ; H_0) = \alpha$，其中 α 是给定的显著性级别