《信息论与编码理论基础》
第三章
博弈与数据压缩

马猛
北京大学

I. 序言

- 赌博及投资与信息论有很强的关系
  - 它们都依赖于随机事件的概率；
  - 它们都需要依据概率做出判断和选择
- 投资的增长率与熵有关
- 在本章的最后，我们会发现一个优秀的投资（赌博）者也是一个优秀的数据压缩器；我们甚至可以应用赌博的方法获得英文熵的估计。

II. 博弈模型

- 博弈模型
  - 假设某赌民将其资金分散购买所有参赛的马匹的马票，b_i表示其下注在第i匹马的资金占总资金的比例。如果第i匹马获胜，获得的回报是下注在第i匹马的资金的o_i倍。考虑到赌民反复下注，则在n场赛马结束时的资产为

$$S_n = \prod_{i=1}^{n} S(X_i)$$

其中$S(X) = b(X)o(X)$是当第X匹马获胜时，赌民购买该只马票所得收益的乘积因子。

  - 双倍率（**doubling rate**）定义：$W(\mathbf{b}, \mathbf{p}) = \mathrm{E}[\log S(X)] = \sum_{k=1}^{m} p_k \log b_k o_k$，其中$p_k$表示第$k$个投资对象成功的概率。

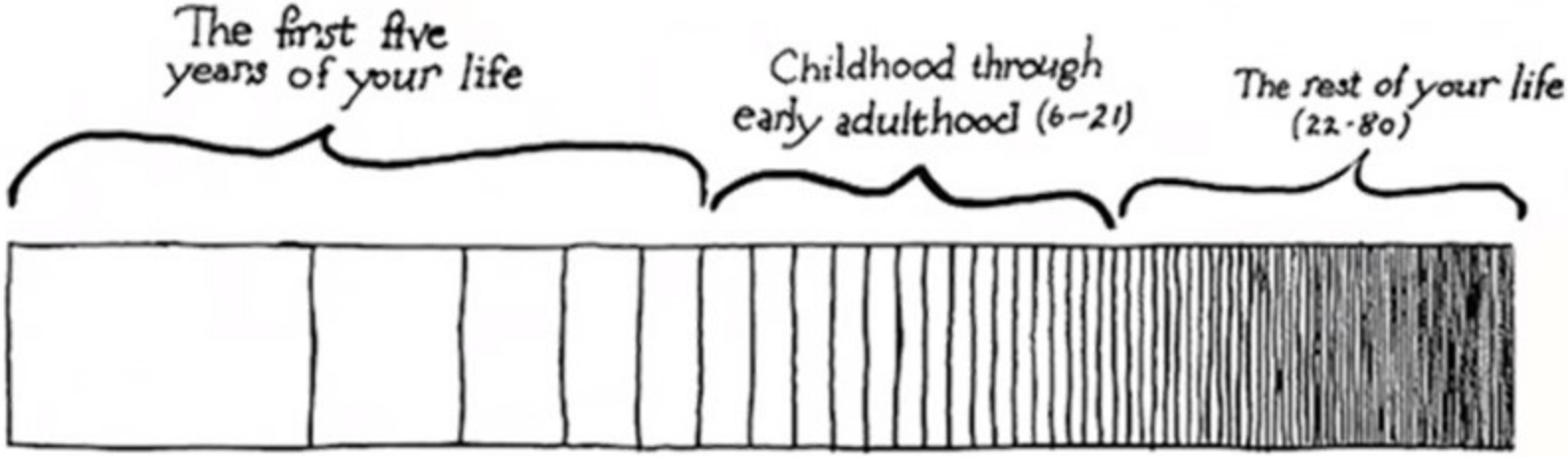
II. 博弈模型

- 双倍率的定义源于我们关心的是“比例”而非“绝对值”
- 类似地，许多人类对世界的感知都是对数的（Logarithmic）而非线性累加的（additive）
  - 听觉：人类的耳朵对声音强度的反应是成对数形式的。通常音量需要增加7~10分贝，我们才觉得音量有1倍的增加，但功率就有了5~10倍的增加。

II. 博弈模型

- 时间：2~3岁经历了一年，和我已有生命的总长度一样；到了80岁，再经历一年，只是我经历岁月的1/80

- 如果一件事对于我们来说是”激动人心“的，新奇的，未知的，这样的记忆在我们脑海中”感觉“时间更长。



- 要抵消我们的“对数感知”，我们需要让我们的“生命精彩程度”**指数级**增长---我们要不停的创造”激动人心“的回忆

III. 按比例下注

- 定理6.1.2（按比例下注是对数最优化的）最优化双倍率的公式计算如下：

$$W^*(\mathbf{b},\mathbf{p}) = \sum_{k=1}^{m} p_k \log o_k - H(\mathbf{p})$$

证明：$W(\mathbf{b},\mathbf{p}) = \sum p_k \log b_k o_k$

$$= \sum p_k \log \frac{b_k}{p_k} p_k o_k = \sum p_k \log o_k - H(\mathbf{p}) - D(\mathbf{p}\|\mathbf{b})$$

上式中$\sum p_k \log o_k$可以看作完全准确的投资所获得的收益，$H(\mathbf{p})$是比赛结果不确定性所带来的投资损失，$D(\mathbf{p}\|\mathbf{b})$是错误的下注方式带来的损失。由相对熵$D(\mathbf{p}\|\mathbf{b}) \geq 0$可得

$$W(\mathbf{b},\mathbf{p}) \leq \sum p_k \log o_k - H(\mathbf{p})$$

等号成立的充要条件是$\mathbf{b} = \mathbf{p}$，即应该按照每匹马获胜的概率按比例分散地购买马票。按此策略下注称为**Kelly博弈**。

- 回报o_k影响双倍率，即赌民的收益，但不应影响投资策略。

III. 按比例下注

- 例题6.1.1. 考虑两匹马比赛，第一匹马赢的概率为p_1，第二匹马赢的概率为p_2。最优的投资策略是$b_1 = p_1, b_2 = p_2$, 赔率2：1，则双倍率为$W^*(\mathbf{b}, \mathbf{p}) = \sum_{k=1}^{m} p_k \log o_k - H(\mathbf{p}) = 1 - H(\mathbf{p})$，财富依如下速率趋紧于无穷：

$$S_n = 2^{n\left(1-H(\mathbf{p})\right)}$$

III. 按比例下注

- 从统计平均的角度，即当投注次数$n \to \infty$时，实际赛马结果为典型集中的概率趋近于1，即仅考虑典型集中事件带来的误差趋近于0。
- 而典型集中事件具有相同的经验分布，按比例下注对典型集中任何事件的收益是相同的。
- 一个优秀的马民也是一个优秀的数据压缩器
  - 马民的每个下注策略可以认为是对数据的概率分布给出的估计。
  - 一个优秀的马民必然得到该概率分布的优秀估计

III. 按比例下注

- 假设赔率有$\sum \frac{1}{o_i} = 1$,我们定义$r_i = \frac{1}{o_i}$，称为Bookie's estimate。则双倍率为：

$$W(\mathbf{b}, \mathbf{p}) = \sum p_k \log b_k o_k$$

$$= \sum p_k \log \frac{b_k}{p_k} \frac{p_k}{r_k} = D(\mathbf{p}\|\mathbf{r}) - D(\mathbf{p}\|\mathbf{b})$$

- 双倍率是bookie's estimate与真实分布的距离减去赌博者的估计与真实分布的距离。
- 如果赌博者的估计比bookie's estimate更好则可以赚钱。
- 如果所有马的赔率相同，为$o_k = m$,则双倍率为

$$W^*(\mathbf{p}) = \log m - H(\mathbf{p})$$

上式第一项相当于带压缩字母集的对数，第二项为压缩后的编码长度。

III. 按比例下注

- Superfair odds: $\sum \frac{1}{o_i} < 1$
  - 此时将所有钱进行投资且按比例下注是最优的
  - 也可以选择$b_i = c\frac{1}{o_i}$，其中$c = {}^{1}/_{\sum \frac{1}{o_i}}$，尽管该策略不是最优的，但财富仍然增长，比例为：$S(X) = {}^{1}/_{\sum \frac{1}{o_i}} > 1$
- Subfair odds: $\sum \frac{1}{o_i} > 1$
  - 这种情况在现实生活中较为常见
  - 此时应将一部分钱进行赌博，而不是全部，按比例投注也不是最优的

IV. 边信息

- 边信息
  - 如果马民拥有某些参赛马匹的历史记录，那么这些信息到底有多少价值呢？
- 定理：由于获得某场赛马X中边信息Y而引起的双倍率的增量ΔW满足

$$\Delta W = I(X;Y)$$

证明：在具有边信息的条件下，按照条件比例买马票，即$b^*(x|y) = p(x|y)$，则

$$W^*(X|Y) = \sum p(x,y) \log o(x)\, p(x|y)$$

$$= \sum p(x,y) \log o(x) - H(x|y)$$

当无边信息时$W^*(X) = \sum p(x) \log o(x) - H(x)$

则$\Delta W = W^*(X|Y) - W^*(X) = H(x) - H(x|y) = I(X;Y)$

- 独立的边信息不会提高双倍率

V. 股票市场投资

- 股票市场的建模：
  - 一个股票市场是由各只股票为分量组成的列向量$\mathbf{X} = (X_1, X_2, \cdots, X_m)^t, X_i \geq 0$称为相对价格，为第$i$只股票当天的收盘价与开盘价之比。
  - 设$F(\mathbf{x})$是相对价格向量的联合分布。
  - 投资组合是列向量$\mathbf{b} = (b_1, b_2, \cdots, b_m)^t, b_i \geq 0, \sum b_i = 1$, b_i是某人投资第i只股票占其总投资的比例
  - 假设每天都进行再投资，到了第n天收盘时，相对收益是这n天中每天的相对收益之乘积。增长率由期望值的对数决定。

V. 股票市场投资

- 股票市场中的增长率为：

$$W(\mathbf{b}, F) = E[\log \mathbf{b}^t \mathbf{X}]$$

如果对数的基底是2，增长率也称为双倍率。

- 定理16.2.1 一个股票市场$\mathbf{X} \sim F$的对数最优投资组合$\mathbf{b}^*$（即使的增长率$W(\mathbf{b}, F)$达到最大值的投资组合）满足下面的充要条件：

$$E\left[\frac{X_i}{\mathbf{b}^t \mathbf{X}}\right] \begin{cases} = 1 \text{ 当} b_i^* > 0 \\ \leq 1 \text{ 当} b_i^* = 0 \end{cases}$$

证明：略（见教材349）

V. 股票市场投资

- 假如资金的初始分配为$\mathbf{b}^*$，那么当天收盘后，第i只股票的相对收益与整个投资组合的相对收益的比例为$\frac{b_i^* X_i}{\mathbf{b}^t \mathbf{X}}$，其期望为：

$$E\left[\frac{b_i^* X_i}{\mathbf{b}^t \mathbf{X}}\right] = b_i^* E\left[\frac{X_i}{\mathbf{b}^t \mathbf{X}}\right] = b_i^*$$

因此，第i只股票当天收盘后的相对收益占整个投资组合的相对收益的比例的数学期望与当天开盘时投资该股的资金比例相同。这是Kelly按比例博弈的翻版。

VI. 语言的熵

- 熵率
  - 给定一个长度为n的随机变量序列，该序列的熵关于n的增长率称为**熵率**
  - 当如下极限存在时，随机过程$\{x_i\}$的熵率定义为：

$$H(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n)$$

  - 对于i.i.d.随机序列$X_1, X_2, \cdots, X_n$有

$$H(X) = H(X_1)$$

  - 对于平稳随机过程有

$$H(X) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \cdots, X_1)$$

  - 英文文本不是i.i.d.，也不是平稳的

VI. 语言的熵

- 英语的熵率
  - 英文中有26个字母和空格，共27个字符
  - 通过收集一些文本样本，根据文本中字符的经验分布建立英文模型
  - 字母出现的概率远不是均匀的
    - 字母E出现的频率最高达13%，频率最低的字母Q和Z大约为0.1%
    - 双字母出现的概率也远不是均匀的。TH出现的概率为3.7%
  - 可以利用马尔可夫过程对英文建模
- 英文的分布对于加密的英文文本译码十分有用
  - 例如，在简单的替代加密（即任何一个字母都用另外一个字母替换）的密文中，通过搜索频率最高的字母来取定该字母替换了E

VI. 语言的熵

- 香侬猜字游戏
  - 在此游戏中，给出一篇英文文章的样本，要求猜出下一个字母是什么。
  - 一个优秀的嘉宾应该首先估计下一个可能出现的字母的概率，然后以概率大小从大到小依次猜测。实验者记录猜中下一个字母所需要的次数
  - 嘉宾猜测次数构成的序列实际上可以作为英文的一种编码方式，将嘉宾模拟为计算机，则计算机根据每个字母的猜测次数可以确认该字母
  - 根据猜测次数序列的熵可以获得英文的熵为1.3比特/字符

VI. 语言的熵

- 西方几种主要语言的熵率值
  - $\log_2 K$表示按照字母表大小为$K = 26$计算的最大值
  - $H_1(U)$是根据各字母的实际概率计算所得的熵
  - $H(W)$是按照单词计算所得的平均每个字母的熵值

西方几种主要语言的熵值(单位:bit)

熵	英语	法语	德语	西班牙
$\log_2 K$	4 .70	4 .70	4 .70	4 .70
$H_1(U)$	4 .124	3 984	4 095	4 .015
$H(W)$	1 .65	3 .02	1 .08	1 97

习题

- 阅读教材6.1，6.2，6.4，6.6
- 习题6.1，6.3