

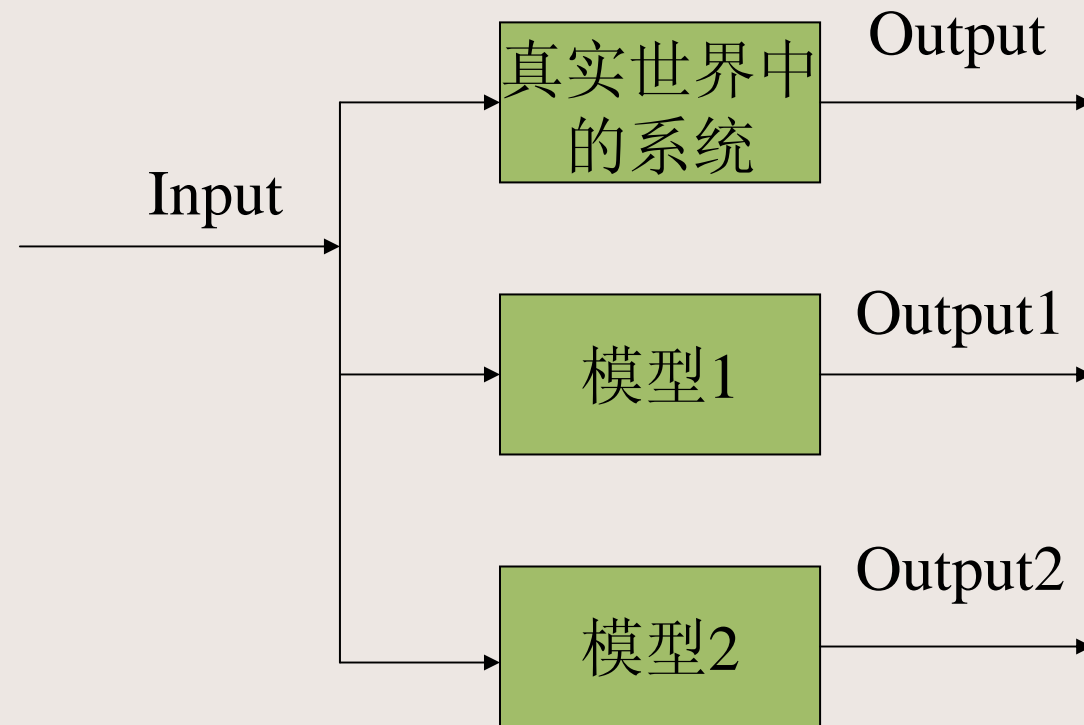
统计自然语言处理基本概念

刘挺

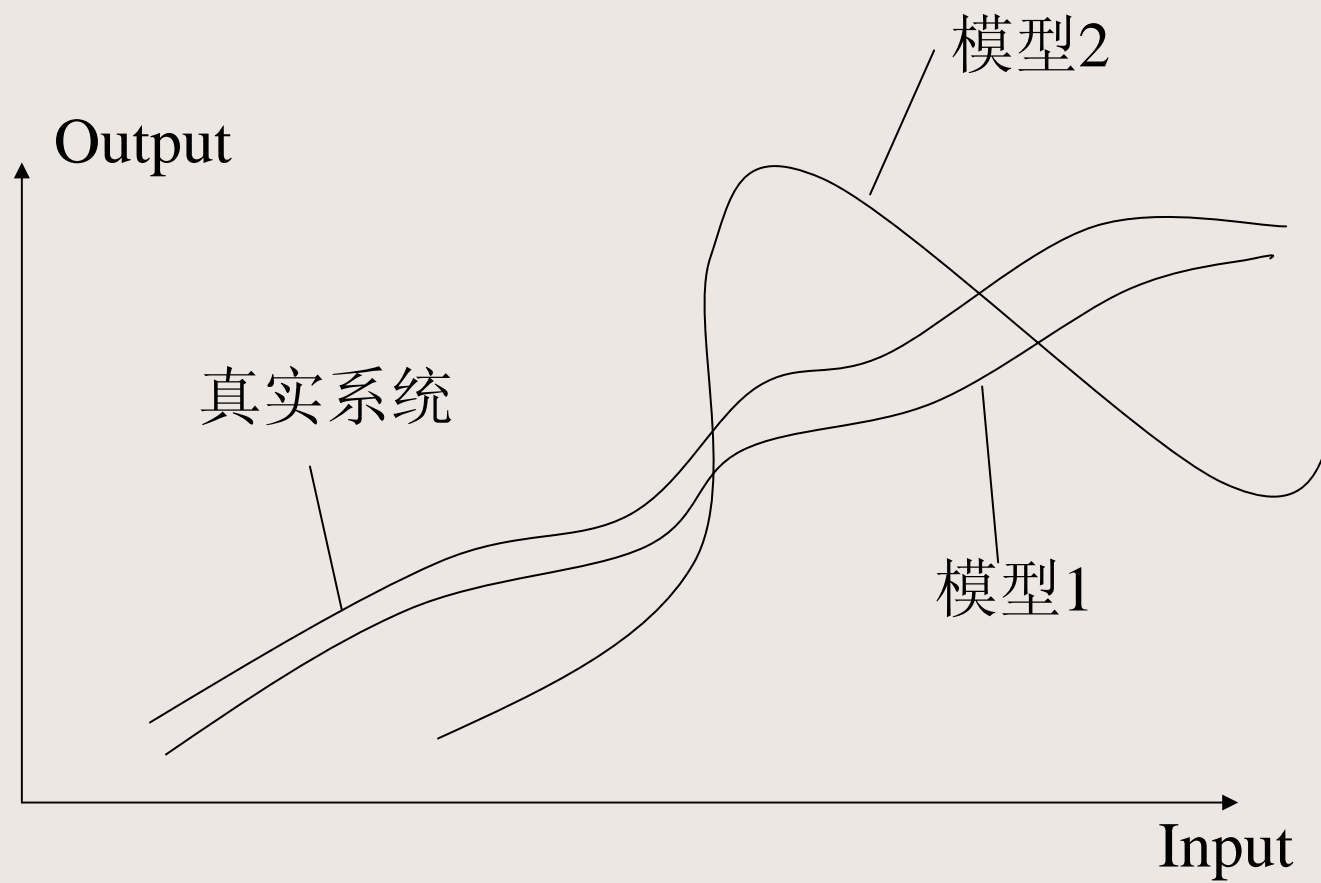
哈工大信息检索研究室

2004年春

模型



如果Output1总是和Output接近，Output2总是和Output偏离，我们就认为模型1比模型2好



- 模型由体系结构和参数两部分构成

- 举例：住宅楼

- 多层板楼
 - 高层板楼
 - 高层塔楼

- 参数

- 层数：
 - 户型：三室一厅，两室一厅，
 - 举架高度：
 - 供热方式：地热？暖气片？

目录

- 样本空间(Sample Space)
- 估计器(Estimator)和随机过程(Stochastic Process)
- 信息论(Information Theory)
- 数据集分类(Data Set Classification)
- 性能评价(Performance Measure)

The background of the slide is a spiral-bound notebook. The notebook has a brown cover and a light beige, textured fabric-like surface. A silver metal spiral binding is visible along the left edge. The text is centered on the notebook's cover.

样本空间 (Sample Space)

试验(Experiment)

- 试验

- 一个可观察结果的人工或自然的过程，其产生的结果可能不止一个，且不能事先确定会产生什么结果
- 例如
 - 连掷两次硬币

- 样本空间

- 是一个试验的全部可能出现的结果的集合
- 举例
 - 连掷两次硬币
 - $\Omega = \{HH, HT, TH, TT\}$, H:面朝上; T:面朝下

事件(Event)

- 事件
 - 一个试验的一些可能结果的集合，是样本空间的一个子集
 - 举例：连掷两次硬币
 - A: 至少一次面朝上
 - B: 第二次面朝下
 - $A=\{HT, TH, HH\}$, $B=\{HT, TT\}$

事件的概率

- 事件的概率
 - 重复 m 试验，如果事件 A 出现的次数为 n ，则事件 A 的概率为 $P(A)=n/m$ ，这称为概率的频率解释，或称统计解释
 - 频率的稳定性又称为经验大数定理
 - 举例：连掷两次硬币
 - A : 至少一次面朝上
 - B : 第二次面朝下
 - $P(A)=3/4$, $P(B)=1/2$
 - 当试验不能重复时，概率失去其频率解释的含义，此时概率还有其他解释：贝叶斯学派和信念学派
 - 一个人出生时的体重，一个人只能出生一次

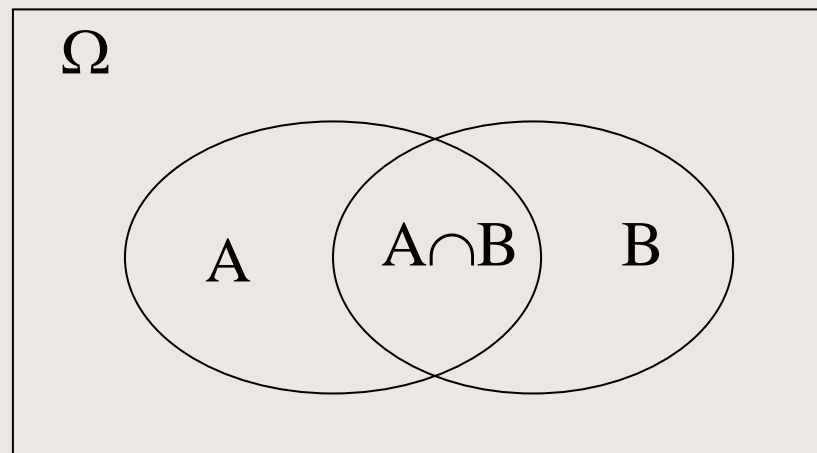
举例

- 举例：连续三次掷硬币
 - 样本空间
 - $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
 - 事件A：恰好两次面朝下
 - $A = \{HTT, THT, TTH\}$
 - 做1000次试验，计数得386次为两次面朝下
 - 估计： $P(A) = 386/1000 = 0.386$
 - 继续做7组试验，得：373，399，382，355，372，406，359，共8组试验
 - 计算平均值： $P(A) = (0.386 + 0.373 + \dots)/8 = 0.379$ ，或
累计： $P(A) = (386 + 373 + \dots)/8000 = 3032/8000 = 0.379$
 - 统一的分布假设为： $3/8 = 0.375$

概率空间

- 概率空间的三个公理
 - $P(A) \geq 0$
 - $P(\Omega) = 1$
 - $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \Phi$
 - 这三条公理也是概率的原始定义
- 推论：
 - $P(\Phi) = 0$; $A \subseteq B \Rightarrow P(A) < P(B)$; $P(\bar{A}) = 1 - P(A)$
- 不是所有0和1之间的值都是概率
 - 例如: $|\cos(x)|$ 就不是概率

概率空间图示



联合事件

- A和B两个事件的联合概率就是A和B两个事件同时出现的概率
 - A和B的联合概率表示为： $P(A, B)$ 或 $P(A \cap B)$
 - 举例：连掷两次硬币
 - 事件A：第一次面朝上， $A=\{HH, HT\}$
 - 事件B：第二次面朝下， $B=\{HT, TT\}$
 - 联合事件 $A \cap B=\{HT\}$

条件概率

- 在事件B发生的条件下事件A发生的概率
 - $P(A|B)=P(A,B)/P(B)$
 - $P(A|B)=(c(A,B)/T)/(c(B)/T)=c(A,B)/c(B)$
 - $c(A)$ 代表事件A出现的次数, $c(B)$ 同理
 - T 是试验总次数
 - 举例: 两次掷硬币问题
 - 事件A: 第一次面朝上, $A=\{HH,HT\}$
 - 事件B: 第二次面朝下, $B=\{HT,TT\}$
 - $A \cap B=\{HT\}$
 - $P(A|B)=1/2$
 - 条件概率可以被视为从另外一个样本空间产生

概率的乘法原理

- $P(A,B)=P(A|B)\times P(B)=P(B|A)\times P(A)$
- Chain Rule
 - $P(A_1,A_2,\dots,A_n)=P(A_1)\times P(A_2|A_1)\times P(A_3|A_1,A_2)$
 $\times \dots \times P(A_n|A_1,A_2,\dots,A_{n-1})$
- 举例1：词性标注
 - $P(\text{det},\text{adj},n)=P(\text{det})\times P(\text{adj}|\text{det})\times P(n|\text{det},\text{adj})$
- 举例2：计算一个句子的概率
 - $p(w_1,w_2,\dots,w_n)=p(w_1)p(w_2|w_1)\dots p(w_n|w_1\dots w_{n-1})$

独立和条件独立

- 独立

- 定义: $P(A,B)=P(A)\times P(B)\Rightarrow P(A|B)=P(A), P(B|A)=P(B)$

- 条件独立

- 定义: $\underline{P(A,B|C)}=P(A|B,C)\times P(B|C)=P(A|C)\times P(B|C)\Rightarrow P(A|B,C)=P(A|C), P(B|A,C)=P(B|C)$
- Naïve Bayesian: 假定各特征之间条件独立
 - $P(A_1,A_2,\dots,A_n|B)=\prod_{i=1,\dots,n}P(A_i|B)$
- 避免一个错误: $\underline{P(A|B,C)}=P(A|B)\times P(A|C)$

独立和条件独立

- 独立不意味着条件独立
 - 举例：色盲和血缘关系
 - A: 甲是色盲
 - B: 乙是色盲
 - C: 甲和乙有血缘关系
 - $P(A,B)=P(A)\times P(B)$
 - $P(A,B|C) \neq P(A|C)\times P(B|C)$
- 条件独立不意味着独立
 - $P(\text{肺癌}, \text{买雪茄} | \text{吸烟}) = P(\text{肺癌} | \text{吸烟}) \times P(\text{买雪茄} | \text{吸烟})$
 - $P(\text{肺癌}, \text{买雪茄}) \neq P(\text{肺癌}) \times P(\text{买雪茄})$

Bayes' Rule

- 根据乘法原理：
 - $P(A,B)=P(A)\times P(B|A)=P(B)\times P(A|B)$
 - 得到贝叶斯原理： $P(A|B)=P(A)\times P(B|A)/P(B)$
- 应用1
 - $\operatorname{argmax}_A P(A|B)=\operatorname{argmax}_A P(A)P(B|A)/P(B)$
 $=\operatorname{argmax}_A P(A)P(B|A)$
- 应用2
 - A_1, A_2, \dots, A_n 是特征， B 是结论
 - $P(B|A_1, A_2, \dots, A_n)=P(A_1, A_2, \dots, A_n|B)P(B)/P(A_1, A_2, \dots, A_n)$
 - 其中： $P(A_1, A_2, \dots, A_n|B)=\prod_{i=1,n} P(A_i|B)$

Bayes举例

- 应用3

- 英汉统计机器翻译

- $P(CW_1, \dots, CW_m | EW_1, \dots, EW_n) =$
 $P(EW_1, \dots, EW_n | CW_1, \dots, CW_m) \times P(CW_1, \dots, CW_m) / P(EW_1, \dots, EW_n)$

- 汉语句子的 CW_1, \dots, CW_m

- 英语句子 EW_1, \dots, EW_m

- 翻译模型: $P(EW_1, \dots, EW_n | CW_1, \dots, CW_m)$

- 目标语语言模型: $P(CW_1, \dots, CW_m)$

随机变量(Random Variable)

- 随机变量是一个函数 $X:\Omega\rightarrow\mathbf{R}$ 。 Ω 是样本空间， \mathbf{R} 是实数集合
 - 人们常常关心和样本点有关的数量指标
 - 数值也比事件更易于处理，举例打靶的环数
- 举例：
 - $[X=0]=\{TT\}$ ； $[X=1]=\{TH,HT\}$ ； $[X=2]=\{HH\}$
 - X 是两次掷硬币面朝上的次数
- 数值可以是连续值，也可以是离散值
- $P_X(x)=P(X=x)=_{\text{df}}P(A_x)$, $A_x=\{a\in\Omega:X(a)=x\}$ ，通常简写作 $P(x)$

期望Expectation

- 期望是随机变量的均值
 - $E(X) = \sum_{x \in X(\Omega)} x \cdot P_X(x)$ (对于离散值)
 - $E(X) = \int_{\mathbb{R}} x P(x) dx$ (对于连续值)
 - 举例:
 - 六面骰子问题: $E(X) = 3.5$
 - $1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 6 \cdot 1/6 = 3.5$
 - 两次六面骰子得到的点数和: $E(X) = 7$
 - $2 \cdot 1/36 + 3 \cdot 2/36 + 4 \cdot 3/36 + \dots = 7$
- 方差(Variance)
 - $E((X-E(X))^2) = \sum_{x \in X(\Omega)} (x-E(X))^2 \cdot P_X(x)$ (对于离散值)
 - $E((X-E(X))^2) = \int_{\mathbb{R}} (x-E(X))^2 P(x) dx$ (对于连续值)
 - 王励勤和王皓的期望接近, 王励勤的方差大

概率分布

- 多项式分布(Multinomial Distribution)
 - $P(n_1, \dots, n_m) = n! / (n_1! \dots n_m!) \times p_1^{n_1} \dots p_m^{n_m}$
 - $\sum_i n_i = n$, 做n次试验
 - 输出第i种结果的次数是 n_i , 第i种结果出现的概率是 p_i
- 二项式分布(Binomial Distribution)
 - 输出: 0或1
 - 做n次试验
 - 关心的是试验成功的次数的概率
 - $P_b(r|n) = C_n^r p^r (1-p)^{n-r}$
 - C_n^r 是从n个元素中任意取出r个元素的组合数
 - p是成功的概率
 - 如果是等概率分布, 则 $p=1/2$, $P_b(r|n) = C_n^r / 2^n$

协方差和相关系数

- 协方差(Covariance)
 - $C_{xy} = E[(X - E(X))(Y - E(Y))]$
- 相关系数(Correlation Coefficient)
 - $\rho_{xy} = C_{xy} / (\sigma_x \sigma_y)$
 - σ_x 是随机变量X的方差
 - σ_y 是随机变量Y的方差
 - $-1 \leq \rho \leq 1$, $\rho > 0$ 正相关, $\rho < 0$ 负相关, $\rho = 0$ 不相关

The image shows the front cover of a spiral-bound notebook. The cover has a light beige, textured fabric-like surface. A silver-colored metal spiral binding is visible along the left edge. The title is centered on the cover in a black serif font.

参数估计 Parameter Estimation

参数估计

- 研究对象的全体所构成的集合成为总体 (population)
- 数理统计的任务：已经知道总体的一部分个体的指标变量值，以此为出发点来推断总体分布的性质
- 简单样本 (simple sample) 是指这样的样本 (X_1, X_2, \dots, X_n) ，它的分量 X_i ， $i=1, \dots, n$ 是独立同分布的随机变量（向量）

估计器

- 设 (X_1, X_2, \dots, X_n) 为一个样本，它的一个与总体分布无关的函数（或向量函数） $f(X_1, X_2, \dots, X_n)$ 称为一个统计量(statistics)
 - 举例：掷硬币问题
 - X ：面朝上/面朝下
 - $T(X_1, X_2, \dots, X_n)$ ：面朝上的次数
- 估计器(Estimator)
 - 根据样本计算参数
 - 一个估计器是随机变量的函数，同时其自身也可以视为一个随机变量
 - 估计器的准确率依赖于采样数据的大小

参数估计

- 所有参数都是从一个有限的样本集合中估计出来的
 - 一个好的估计器的标准：
 - 无偏(unbias): 期望等于真实值
 - 有效(efficient): 方差小
 - 一致(consistent): 估计的准确性随样板数量的增加而提高
- 一些常用的估计方法
 - 极大似然估计
 - 最小二成估计
 - 贝叶斯估计

极大似然估计

- 极大似然估计
 - Maximum Likelihood Estimation(MLE)
- 选择一组参数 θ ，使似然函数 $L(\theta)$ 达到最大
 - $L(\theta)=f(x_1, x_2, \dots, x_n|\theta)=\prod_{i=1, n} f(x_i|\theta)$
- 举例：
 - 罐里有黑球和白球，比例3:1，今连续抽取两球全为黑球，问罐里黑球多还是白球多？
 - 设黑球概率为 p ，抽取 n 次拿到 x 次黑球的概率符合二项分布： $f_n(x, p)=C_n^x p^x (1-p)^{n-x}$
 - 今抽取两次全是黑球 $f_2(2, p)=C_2^2 p^2 (1-p)^0=p^2$
 - 若 $p=1/4$ ，则 $f_2(2, p)=1/16$ ；若 $p=3/4$ ，则 $f_2(2, p)=9/16$
 - 选择概率大的： $p=3/4$ ，黑球多

随即过程

- 随即过程(Stochastic Process)
 - $X(t), t \in T$
 - X 是一组随机变量
 - T 是过程的索引集合，例如时间或位置
 - 如果 T 是可数集，则 $X(t)$ 是离散时间过程
- 举例：词性标注
 - $C(t)$, C 是词性, t 是位置
 - $C(1)=\text{noun}, C(2)=\text{verb}, \dots, C(n)=\text{pron}$

马尔可夫过程

- 马尔可夫过程，也称马尔可夫链
 - Markov Chain
 - 离散时间，离散状态
 - 无后效性：已知现在状态，则未来和过去无关
 - $P(X_n=x_n|X_1=x_1, X_2=x_2, \dots, X_{n-1}=x_{n-1}) = P(X_n=x_n|X_{n-1}=x_{n-1})$
 - 举例：拼音输入法
 - 一本[书]（输，淑，叔，舒，……）
 - $P(\text{书}|\text{一}, \text{本}) = P(\text{书}|\text{本})$

信息论

信息

- 信息是对物质存在和运动形式的一般描述
- 信息是物质和能量在空间和时间中分布的不均匀程度
- 信息存在于客体间的差别，而非客体本身
 - 题帕三绝
- 新消息的信息量大
 - 布什是美国总统（熟知，信息量小）
 - 马其顿总统遇难（新知，信息量大）

信息论

- 1948年美国Shannon香农“通信的数学理论”，用概率测度和数理统计的方法，系统地讨论了通信的基本问题，奠定了信息论的基础
- 信息的度量有三个基本方向：结构的、统计的和语义的
- 香农所说的信息是狭义的信息，是统计信息，依据是概率的不确定性度量

自信息量

- 自信息量(Self-information)
 - $I(X) = -\log P(X)$
 - 小概率事件包含的信息量大，大概率事件包含的信息量小

互信息

Mutual Information

- $I(x,y)=\log_2 p(x,y)/(p(x)p(y))$
- 比如计算两个词的搭配
 - $I(\text{伟大}, \text{祖国})=\log_2 p(\text{伟大}, \text{祖国})/(p(\text{伟大})p(\text{祖国}))$
 - 此值较高，说明“伟大”和“祖国”是一个比较强的搭配
 - $I(\text{的}, \text{祖国})=\log_2 p(\text{的}, \text{祖国})/(p(\text{的})p(\text{祖国}))$
 - 此值较低，因为 $p(\text{的})$ 太高，“的”和“祖国”不是一个稳定的搭配
- $I(x,y) \gg 0$: x 和 y 关联强度大
- $I(x,y)=0$: x 和 y 无关
- $I(x,y) \ll 0$: x 和 y 具有互补的分布

熵(Entropy)

- 熵(Entropy)
 - Chaos（混沌），无序
 - 物理学：除非施加能量，否则熵不会降低
 - 举例：把房间弄乱很容易，整理干净不容易
 - 是不确定性(Uncertainty)的衡量
 - 不确定性越高，熵越高，我们从一次实验中得到的信息量越大

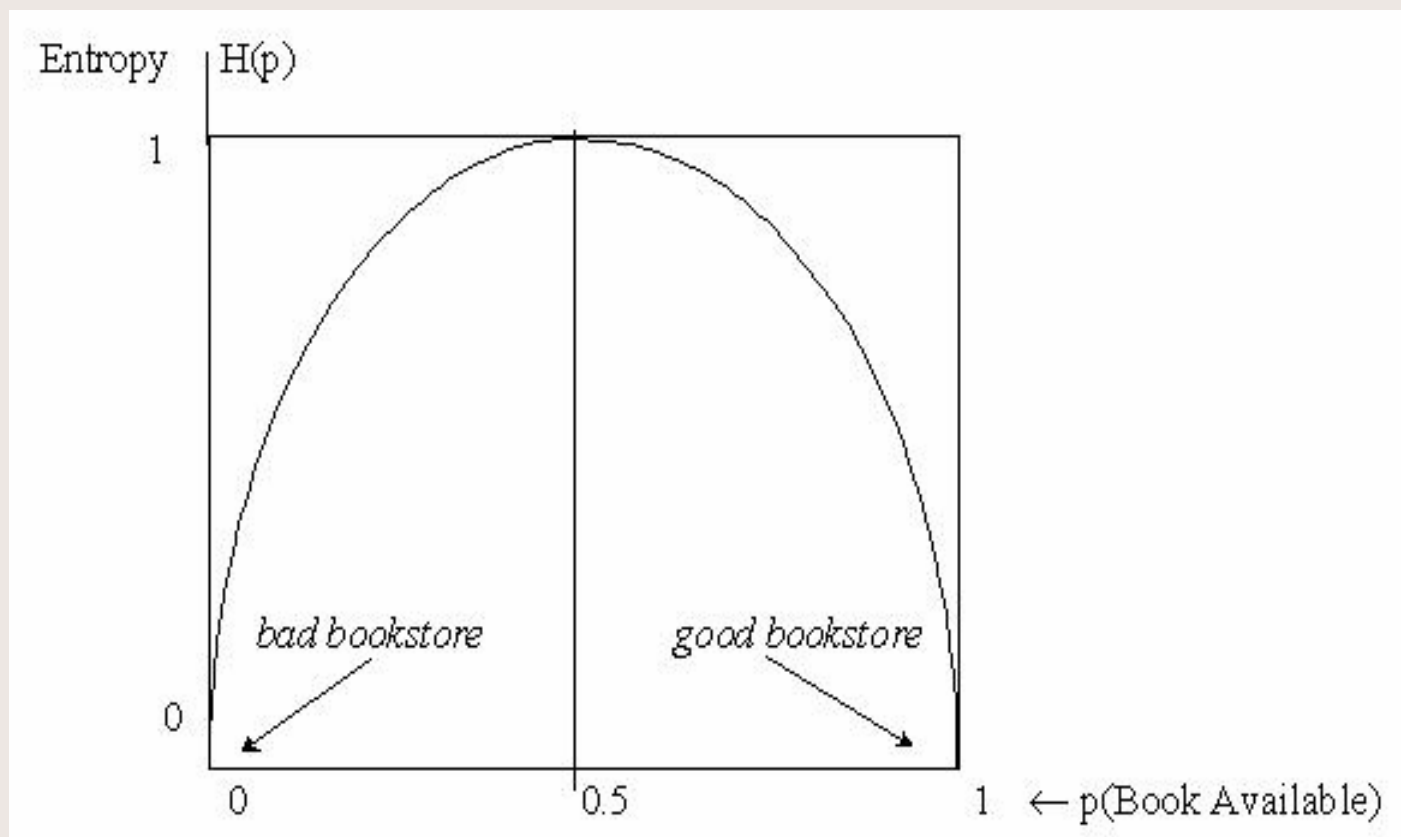
熵的公式

- 熵 $H(X) = -\sum_{x \in \Omega} p(x) \log_x p(x)$
 - 假设 $P_X(x)$ 是随机变量 X 的分布
 - 基本输出字母表是 Ω
 - 单位: bits
- 熵是 X 的平均信息量, 是自信息量的期望
 - $E(X) = \sum_{x \in \Omega} p(x) x$
 - $I(X) = -\log p(x)$, 取2为底, $I(X) = -\log_2 p(x)$
 - $E(I(X)) = E(-\log_2 p(x)) = \sum_{x \in \Omega} p(x) (-\log_2 p(x)) = H(X)$
- $H(X) = H(p) = H_p(X) = H_X(p) = H(p_X)$

熵的例子

- 掷均匀硬币, $\Omega=\{H,T\}$
 - $p(H)=.5, p(T)=.5$
 - $H(p)=-0.5\log_2 0.5+(-0.5\log_2 0.5)=1$
- 32面的均匀骰子, 掷骰子
 - $H(p)=-32((1/32)\log_2(1/32))=5$
- 事实上, $2^1=2, 2^5=32(\text{perplexity})$
- 掷不均匀硬币
 - $p(H)=0.2, p(T)=0.8, H(p)=0.722$
 - $p(H)=0.01, p(T)=0.99, H(p)=0.081$

好书店，差书店



- 什么时候 $H(p)=0$?
 - 试验结果事先已经知道
 - 即: $\exists x \in \Omega, p(x)=1; \forall y \in \Omega, p(y)=0 \text{ if } y \neq x$
- 熵有没有上限?
 - 没有一般的上限
 - 对于 $|\Omega|=n$, $H(p) \leq -\log_2 n$
 - 均衡分布的熵是最大的

- 等概率分布

- 2个输出的等概率分布, $H(p)=1\text{bit}$
- 32个输出的等概率分布, $H(p)=5\text{bits}$
- 43亿输出的等概率分布, $H(p)=32\text{bits}$

- 非等概率分布

- 32个输出, 2个0.5, 其余为0, $H(p)=1\text{bit}$
- 怎样比较具有不同数量输出的“熵”

混乱度Perplexity

- 混乱度
 - $G(p)=2^{H(p)}$
 - 平均每次试验有多少种可能的结果
- 在NLP中，如果词表中的词具有统一的分布概率，则最难预测，熵最大，混乱度最高
- 反之，分布越不均衡，熵越小，混乱度越小

联合熵和条件熵

- 两个随机变量： X (空间是 Ω), $Y(\Psi)$
- 联合熵(Joint Entropy)
 - (X,Y) 被视为一个事件
 - $H(X,Y)=-\sum_{x \in \Omega} \sum_{y \in \Psi} p(x,y) \log_2 p(x,y)$
- 条件熵(Conditional Entropy)
 - $H(Y|X)=-\sum_{x \in \Omega} \sum_{y \in \Psi} p(x,y) \log_2 p(y|x)$
 - $p(x,y)$ 是加权，权值是没有条件的

条件熵

- $$\begin{aligned} H(Y|X) &= \sum_{x \in \Omega} p(x) H(Y|X=x) \\ &= \sum_{x \in \Omega} p(x) \left(- \sum_{y \in \Psi} p(y|x) \log_2 p(y|x) \right) \\ &= - \sum_{x \in \Omega} \sum_{y \in \Psi} p(y|x) p(x) \log_2 p(y|x) \\ &= - \sum_{x \in \Omega} \sum_{y \in \Psi} p(x, y) \log_2 p(y|x) \end{aligned}$$

熵的性质

- 熵的非负的
 - $H(X) \geq 0$
- Chain Rule
 - $H(X, Y) = H(Y|X) + H(X)$
 - $H(X, Y) = H(X|Y) + H(Y)$
- $H(X, Y) \leq H(X) + H(Y)$, X 和 Y 独立时相等
- $H(Y|X) \leq H(Y)$, 条件熵比熵小

熵的编码意义

- 如果一个符号序列是满足概率分布 p 的随机过程产生的，那么对这个序列进行编码至少需要的bit数是 $H(p)$
- 压缩问题
 - 如果数据中有很多重复的模式，则易于压缩，因为熵小
 - 否则，熵大，不容易压缩

编码实例

- 怎样给ISO Latin 1编码？
 - 通常用8位
- 经验表明：有的字符经常出现，有的字符很少出现
 - 我们可以给经常出现的字用较少的bit来表示，给很少出现的字符用较多的bit来表示
 - 假设： $p('a')=0.3$, $p('b')=0.3$, $p('c')=0.3$, 其余 $p(x)=0.0004$
 - 编码： a:00, b:01, c:10, 其余： $11b_1b_2\dots b_8$
 - 对于符号串： acbbécbaac，编码为：
 - a c b b é c b a a c
 - 0010010111000011111001000010
 - 如果每个符号用8位编码，需要80位，现在需要28位

语言的熵

- $p(c_{n+1}|c_1 \dots c_n)$
 - c_i 是语言中的一个字符
 - $c_1 \dots c_n$ 是历史 h
 - 举例：汉语， $n=3$
 - $p(\text{赵}|\text{围魏救})$ ：高
 - $p(\text{去}|\text{我曾经})$ ：低
- 计算语言的条件熵
 - $-\sum_{h \in H} \sum_{c \in \Omega} p(c, h) \log_2 p(c|h)$

各种语言的熵

- 按字母计算的零阶熵
 - 法文：3.98 bits 意大利文：4.00 bits
 - 西班牙文：4.01 bits 英文：4.03 bits
 - 德文：4.10 bits 俄文：4.35 bits
 - 中文（按汉字计算）：9.65 bits
 - 中文（按笔画计算）：3.43 bits
- 按词汇计算的零阶熵
 - 英语：10.0 bits 汉语：11.46 bits
 - 说明汉语的词汇丰富
- 语言的冗余度
 - 英语：73%； 俄语：70%； 汉语：63%； 古文更低

Kullback-Leibler距离

- 假设通过一组试验估计得到的概率分布为 p ，样本空间 Ω ，随机变量 X
- 真实的分布为 q ，相同的 Ω 和 X
- 现在的问题是： p 和 q 相比，误差多大？
- Kullback-Leibler距离给出的答案是：
 - $D(q||p)=\sum_{x \in \Omega} q(x) \log_2 q(x)/p(x)$
 $=E_p \log(q(x)/p(x))$

KL距离（相对熵）

- 习惯上
 - $0\log 0=0$
 - $p\log(p/0)=\infty$
- Distance or Divergence（分歧）
 - 不对称 $D(q\|p)\neq D(p\|q)$
 - 也不满足三角不等式
 - 事实上， $D(q\|p)$ 不是距离，而是分歧
- $H(q)+D(q\|p)$ ：根据 q 分布，对 p 进行编码需要的bit数（交叉熵）

平均互信息

- 随机变量: $X; Y; p_{X \cap Y}(X, Y); p_X(x); p_Y(y)$
- 两个离散集之间的平均互信息
 - $I(X, Y) = D(p(x, y) \| p(x)p(y))$
$$= \sum_{x \in \Omega} \sum_{y \in \Psi} p(x, y) \log_2(p(x, y) / p(x)p(y))$$
 - 这里说的是两个离散集的平均互信息
 - 互信息衡量已知Y的分布时，对X的预测有多大的帮助，或者说Y的知识降低了H(X)
 - 或者说 $p(x, y)$ 和 $p(x)p(y)$ 之间的距离

$$I(X,Y) = \sum_{x \in \Omega} \sum_{y \in \Psi} p(x,y) \log_2 (p(x,y)/p(y)p(x)) =$$

...use $p(x,y)/p(y) = p(x|y)$

$$= \sum_{x \in \Omega} \sum_{y \in \Psi} p(x,y) \log_2 (p(x|y)/p(x)) =$$

...use $\log(a/b) = \log a - \log b$ ($a \sim p(x|y)$, $b \sim p(x)$), distribute sums

$$= \sum_{x \in \Omega} \sum_{y \in \Psi} p(x,y) \log_2 p(x|y) - \sum_{x \in \Omega} \sum_{y \in \Psi} p(x,y) \log_2 p(x) =$$

...use def. of $H(X|Y)$ (left term), and $\sum_{y \in \Psi} p(x,y) = p(x)$ (right term)

$$= -H(X|Y) + (-\sum_{x \in \Omega} p(x) \log_2 p(x)) =$$

...use def. of $H(X)$ (right term), swap terms

$$= H(X) - H(X|Y)$$

...by symmetry, $= H(Y) - H(Y|X)$

互信息的性质

- $I(X, Y) = H(X) - H(X|Y)$
- $= H(Y) - H(Y|X)$
- $I(X, Y) = H(X) + H(Y) - H(X, Y)$
 - 因为: $H(X, Y) = H(X|Y) + H(Y)$
- $I(X, X) = H(X)$ (因为 $H(X, X) = 0$)
- $I(X, Y) = I(Y, X)$
- $I(X, Y) \geq 0$

交叉熵

Cross-Entropy

- 典型情况：
 - 我们得到一个观察序列
 - $T = \{t_1, t_2, \dots, t_n\}, t_i \in \Omega$
 - 估计：
 - $\forall y \in \Omega: p(y) = c(y)/|T|$, 定义: $c(y) = |\{t \in T, t=y\}|$
 - 但是, 真实的 q 不知道, 再大的数据也不够
 - 问题: 用 p 对 q 进行估计是否准确?
 - 方法: 用一个不同的观察序列 T' 估计实际的 q

交叉熵

- $H_{p'}(p) = H(p') + D(p' \| p)$
- $H_{p'}(p) = -\sum_{x \in \Omega} p'(x) \log_2 p(x)$
- p' 当然也不是真实的分布，但是我们视为真实世界的分布，以便测试 p
- 交叉混乱度： $G_{p'}(p) = 2^{H_{p'}(p)}$

条件交叉熵

- 实践中计算的往往是条件交叉熵
- 两个样本空间
 - 样本空间: ψ , 随机变量 Y , $y \in Y$
 - 上下文样本空间: Ω , 随机变量 X , $x \in X$
- 实验得到的分布 $p(y|x)$, “真实”分布 $p'(y|x)$
- $H_{p'}(p) = -\sum_{y \in \psi, x \in \Omega} p'(y, x) \log_2 p(y|x)$
 - 条件熵中的权值是 $p'(y, x)$, 不是 $p'(y|x)$

- 在实际应用中，在全部两个样本空间上做累加通常不是很方便，因此常常简化
- 使用如下公式：

$$- H_{p'}(p) = - \sum_{y \in \Psi, x \in \Omega} p'(y, x) \log_2 p(y|x)$$

$$= -1/|T'| \sum_{i=1 \dots |T'|} \log_2 p(y_i|x_i)$$

- 事实上，就是在 T' 上进行累加，然后归一化

$$= -1/|T'| \log_2 \prod_{i=1 \dots |T'|} p(y_i|x_i)$$

举例

- $\Omega=\{a,b,\dots,z\}$, 概率分布 (估计值)
 - $p(a)=0.25, p(b)=0.5, p(\alpha)=1/64, \alpha \in \{c,\dots,r\}, p(\beta)=0, \beta \in \{s,\dots,z\}$
- 测试数据为: barb , $p(a)=p(r)=0.25, p(b)=0.5$
- 在 Ω 上做累加
 - α

	a	b	c	d	...	q	r	s	...	z
– $-p'(\alpha)\log_2 p(\alpha)$	0.5	0.5	0	0		0	1.5	0		$0=2.5$
- 也可以在测试数据上进行累加, 然后归一化
 - s_i

	b	a	r	b
– $-\log_2 p(s_i)$	1	2	6	1
 - $= 10 \quad (1/4) \times 10 = 2.5$

- $H(p)$ 和 $H_{p'}(p)$ 之间可能有各种关系
 - 包括‘<’, ‘=’, ‘>’
 - 举例（参照上例）
 - $H(P)=2.9$
 - 测试数据: barb
 - $H_{p'}(p) = 1/4(1+2+6+1)=2.5$
 - 测试数据: probable
 - $H_{p'}(p) = 1/8(6+6+6+1+2+1+6+6)=4.25$
 - 测试数据: abba
 - $H_{p'}(p) = 1/4(2+1+1+2)=1.5$

交叉熵的使用

- 不是比较数据，而是比较分布
- 如果我们有两个分布p和q，哪一个更好呢？
 - 面对“真实数据”S，p和q谁的交叉熵低，谁就更好
 - $H_T(p) = -1/|S| \log_2 \prod_{i=1 \dots |S'|} p(y_i | x_i)$
 - $H_T(q) = -1/|S| \log_2 \prod_{i=1 \dots |S'|} q(y_i | x_i)$

Test data S: probable

- $p(\cdot)$ from prev. example:

$$H_S(p) = 4.25$$

$p(a) = .25, p(b) = .5, p(\alpha) = 1/64$ for $\alpha \in \{c..r\}, = 0$ for the rest: s, t, u, v, w, x, y, z

- $q(\cdot|\cdot)$ (conditional; defined by a table):

$q(\cdot \cdot) \rightarrow$ ↓	a	b	e	l	o	p	r	other
a	0	.5	0	0	0	.125	0	0
b	1	0	0	0	1	.125	0	0
e	0	0	0	1	0	.125	0	0
l	0	.5	0	0	0	.125	0	0
o	0	0	0	0	0	.125	1	0
p	0	0	0	0	0	.125	0	1
r	0	0	0	0	0	.125	0	0
other	0	0	1	0	0	.125	0	0

ex.: $q(o|r) = 1$


$q(r|p) = .125$

$$(1/8) (\log(p|oth.) + \log(r|p) + \log(o|r) + \log(b|o) + \log(a|b) + \log(b|a) + \log(l|b) + \log(e|l))$$

$$(1/8) (0 + 3 + 0 + 0 + 1 + 0 + 1 + 0)$$

$$H_S(q) = .625$$

数据集分类

- 
- 训练集Training Set
 - 用来获得模型参数
 - 测试集Testing Set
 - 从训练集以外独立采样
 - 反映系统面对真实世界的处理能力
 - 测试集经常被无意识地“做了手脚”
 - 交叉确认集Cross-Validation Set
 - 从训练集和测试集以外独立采样
 - 主要用来帮助做设计决策

测试集

- 测试集
 - 从训练集去评价系统的性能，结果往往过于乐观
 - 如果模型的参数比需要的多很多时，获得100%的准确率也是可能的
 - 过拟合(Over-fitting)常常出现在训练数据的数量不足以支持模型的复杂程度之时
 - 为此，我们需要另一个数据集来模拟用户的真实需要

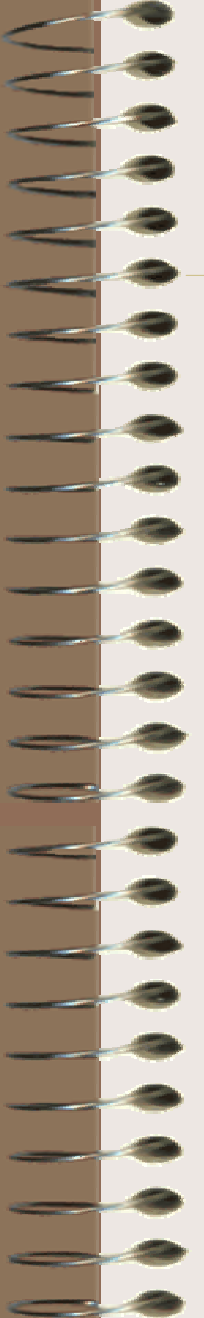
- 在设计阶段，不允许偷看测试数据的细节，以保证测试数据不被污染
 - 你不能参照测试数据来决定模型的复杂度，特征空间的维数，以及什么时候决定停止训练过程等
 - 设计决策可以参照交叉确认数据进行
- 每一个阶段采用一个不同测试集
 - 当你试图选择一个最好的方法是测试效果达到最佳时，实际上已经在无意识地使你的系统偏向测试集
 - 问题的关键在于测试集并不是真实数据本身，如果面向测试集调整参数，可能造成系统对于从未见过的真实数据效果下降

A spiral-bound notebook with a brown cover and a light beige page. The spiral binding is on the left side. The page has a horizontal line near the top.

• 交叉确认集

- 如果在训练集合上获得了比较差的结果，我们必须重新设计
- 如果在训练集合上获得了比较好的结果，那可能是因为：
 - 模型确实好（在测试数据上性能一样会好）
 - 模型过拟合（在测试数据上性能会下降）
- 由于不允许使用测试集来改进系统设计，因此需要另一个数据集

性能评价

- 
- 使用有限的样本进行性能测试
 - 有估计误差
 - 性能评价的结果和测试数据的大小有关
 - 不同数据集的测试结果往往不同
 - 性能上限Performance Upper Bound
 - 人与人取得一致的指标就是系统性能的上限

- 联立表(Contingency table)

第一类： + 第二类： -		系统给出的标记	
		+	-
正确标记	+	N11	N21
	-	N12	N22

- 准确率P(Precision)
 - $N_{11}/(N_{11}+N_{21})$
- 召回率R(Recall)
 - $N_{11}/(N_{11}+N_{12})$
- 错误率E(Error Rate)
 - $(N_{12}+N_{21})/(N_{11}+N_{12}+N_{21}+N_{22})$
- F-measure
 - $2PR/(P+R)$

A spiral-bound notebook with a textured, light brown cover and a dark brown border. The spiral binding is on the left side. The text "谢谢!" is centered on the cover.

谢谢！