

# 最大熵模型

---

3月机器学习在线班 邹博

2015年3月22日

# 本次目标

- 理解并掌握熵Entropy的定义
  - 理解“Huffman编码是所有编码中总编码长度最短的”熵含义
- 理解联合熵 $H(X,Y)$ 、相对熵 $D(X||Y)$ 、条件熵 $H(X|Y)$ 、互信息 $I(X,Y)$ 的定义和含义，并了解如下公式：
  - $H(X|Y) = H(X,Y) - H(Y) = H(X) - I(X,Y)$
  - $H(Y|X) = H(X,Y) - H(X) = H(Y) - I(X,Y)$
  - $I(X,Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y) \geq 0$
- 掌握最大熵模型Maxent
  - Maximum Entropy Models
- 了解最大熵在自然语言处理NLP中的应用
  - Natural Language Processing
- 与前序知识的联系：最大熵模型和极大似然估计MLE的关系
  - Maximum Likelihood Estimation
- 副产品：了解数据分析、函数作图的一般步骤



# 预备题目

---

□ 证明:  $-\ln x \geq 1-x, x>0$

■  $f(x)=-\ln x + x - 1, x>0,$

■ 凸函数

■ 在 $x=1$ 处取极值



# 骰子

每个面朝上的概率分别是多少

— 所有人都说是等概率，即各点的概率均为 $1/6$

为什么？

— “一无所知”的骰子

— 假定它每一个朝上概率均等是最安全的做法

新问题

—  $N$ 次投掷结果的平均值为 $\mu = 5.5$

— 六个面出现的次数各是多少？



# 优化问题

$$S(\mathbf{p}) = -\sum_i p_i \ln p_i$$

$$-\sum_i p_i = 1$$

$$-\sum_i i \cdot p_i = \mu$$

拉格朗日乘子法

$$-\zeta = -\sum_i p_i \ln p_i + \lambda_0(1 - \sum_i p_i) + \lambda_1(\mu - \sum_i i \cdot p_i)$$

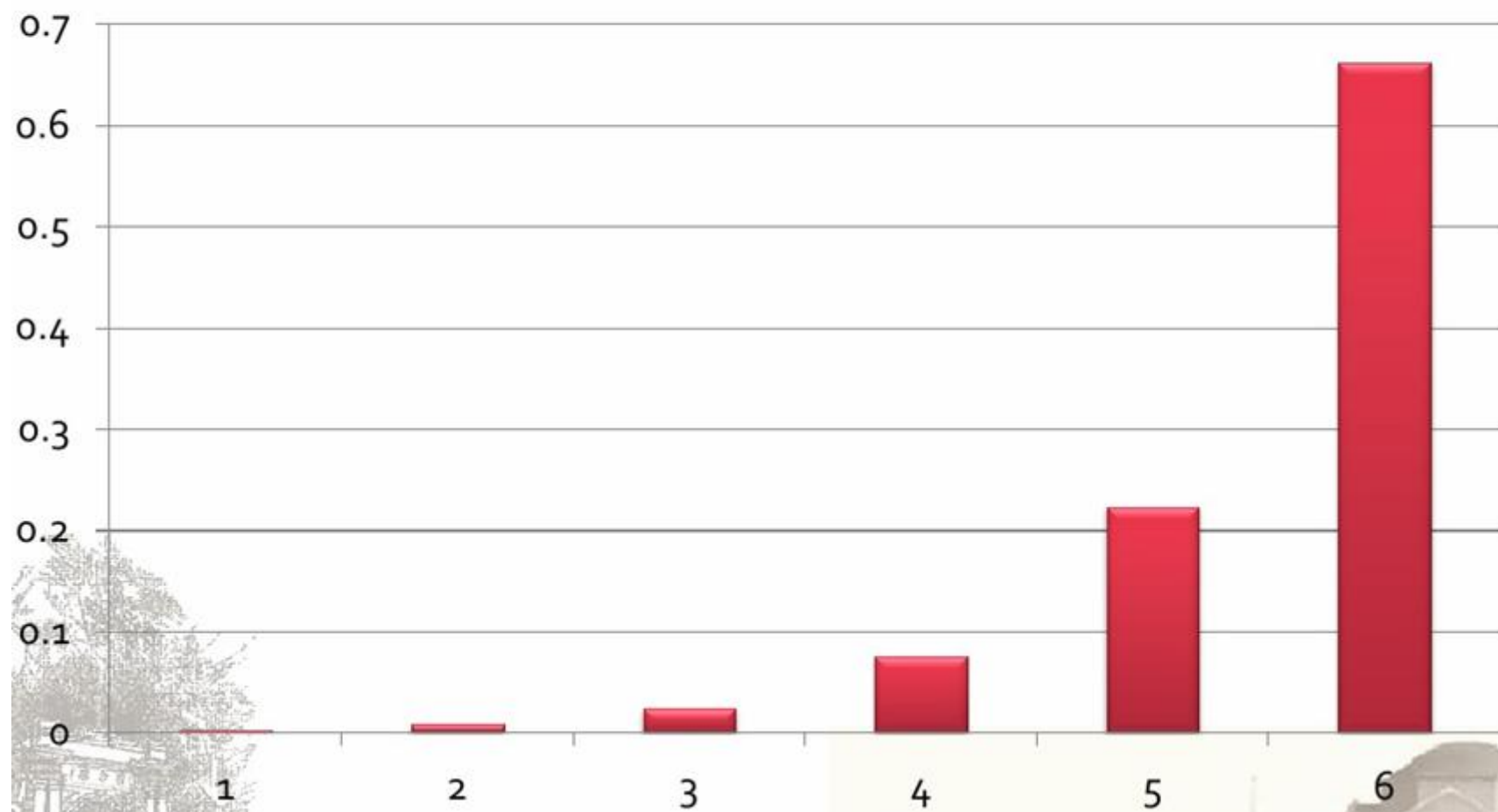
- 通过  $\frac{\partial \zeta}{\partial p_i} = 0$ , 可得

$$p_i = e^{-1-\lambda_0-i\lambda_1}$$

$$\lambda_0 = 5.932, \quad \lambda_1 = -1.087$$



# 预测结果



# 从小学数学开始

---

- 假设有5个硬币：1,2,3,4,5，其中一个是假的，比其他的硬币轻。有一个天平，天平每次能比较两堆硬币，得出的结果可能是以下三种之一：
  - 左边比右边轻
  - 右边比左边轻
  - 两边同样重
- 问：至少要使用天平多少次才能**确保**找到假硬币？

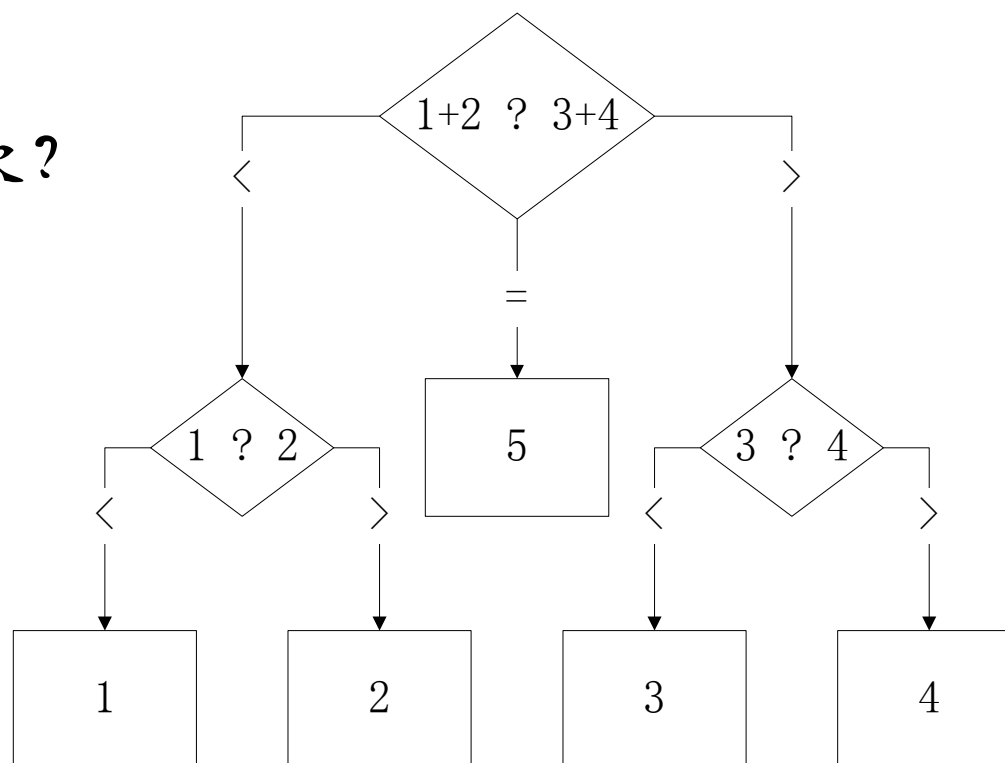


# 答案

□ 一种可能的称量方法如右图所示

□ 答案：2次

□ 追问：为什么2次？





# 分析

- 令 $x$ 表示假硬币的序号:  $x \in X = \{1, 2, 3, 4, 5\}$ ;
- 令 $y_i$ 是第 $i$ 次使用天平所得到的结果:  
 $y \in Y = \{1, 2, 3\}$ ;
  - 1表示“左轻”, 2表示“平衡”, 3表示“右轻”
- 用天平称 $n$ 次, 获得的结果是:  $y_1 y_2 \dots y_n$ ;
- $y_1 y_2 \dots y_n$ 的所有可能组合数目是 $3^n$ ;
- 根据题意, 要求通过 $y_1 y_2 \dots y_n$ 确定 $x$ 。即建立映射  
 $\text{map}(y_1 y_2 \dots y_n) = x$ ;
- 从而:  $y_1 y_2 \dots y_n$ 的变化数目大于等于 $x$ 的变化数目
  - 即  $3^n \geq 5$
  - 一般意义下:  $|Y|^n \geq |X|$



## 进一步分析

---

- 用  $y_1 y_2 \dots y_n$  表达  $x$ 。即设计编码:  $x \rightarrow y_1 y_2 \dots y_n$
- $X$  的“总不确定度”是:  $H(X) = \log |X| = \log 5$
- $Y$  的“表达能力”是:  $H(Y) = \log |Y| = \log 3$
- 至少要多少个  $Y$  才能准确表示  $X$ ?

$$\frac{H(X)}{H(Y)} = \frac{\log 5}{\log 3} = 1.46$$



# 题目的变种

---

- 假设有5个硬币：1,2,3,4,5，其中一个是真的，比其他的硬币轻。已知第一个硬币是真硬币的概率是三分之一；第二个硬币是真硬币的概率也是三分之一，其他硬币是真硬币的概率都是九分之一。
- 有一个天平，天平每次能比较两堆硬币，得出的结果可能是以下三种之一：
  - 左边比右边轻
  - 右边比左边轻
  - 两边同样重
- 假设使用天平 $n$ 次找到真硬币。问 $n$ 的期望值至少是多少？



# 解

---

□ 1/3概率的硬币有2个，1/9概率的硬币有3个：

$$\left(\frac{1}{3} + \frac{1}{3}\right) \times \frac{\log 3}{\log 3} + 3 \frac{1}{9} \times \frac{\log 9}{\log 3} = \frac{4}{3}$$

□ 定义： $-\sum p \log_a p$ 为熵



# 用熵解释Huffman编码

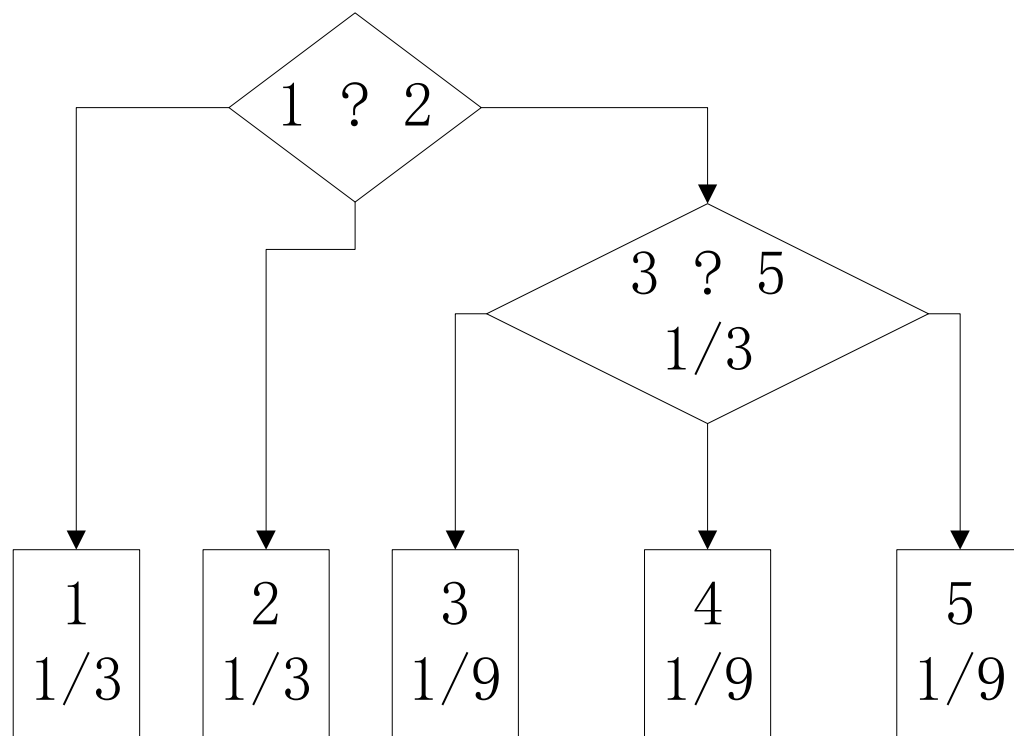
---

1	2	3	4	5
$1/3$	$1/3$	$1/9$	$1/9$	$1/9$



# 用熵解释Huffman编码

---



# Huffman编码

---

□ 本质：高概率出现的字符用更短的编码



# 广泛的结论

- 如果一个随机变量 $x$ 的可能取值为 $X=\{x_1, x_2, \dots, x_k\}$ 。要用 $n$ 位 $y: y_1 y_2 \dots y_n$ 表示(每位 $y$ 有 $c$ 种取值) $n$ 的期望值至少为:

$$\sum_{i=1}^k p(x=x_i) \frac{\log \frac{1}{p(x=x_i)}}{\log c} = \frac{\sum_{i=1}^k p(x=x_i) \log \frac{1}{p(x=x_i)}}{\log c}$$

- 一般地, 我们令 $c$ 为2(二进制表示), 于是,  $X$ 的信息量为:

$$H(X) = \sum_{i=1}^k p(x=x_i) \log \frac{1}{p(x=x_i)}$$





# 熵

---

□ 将 $P(x=x_i)$ 写成普适公式，就得到熵的定义：

$$H(X) = - \sum_{x \in X} p(x) \ln p(x)$$

- 注：经典熵的定义，底数是2，单位是bit
- 本例中，为分析方便使用底数e
- 若底数是e，单位是nat(奈特)



# 研究函数 $f(x)=x\ln x$

---

- $f(x)=x\ln x, x \in [0,1]$
- $f'(x) = \ln x + 1$
- $f''(x) = 1/x > 0$  (凸函数)
- 当  $f'(x)=0$  时,  $x=1/e$ , 取极小值;
  
- 由于  $\lim_{x \rightarrow 0} f(x) = 0$
- 定义  $f(0)=0$



# 离散采样

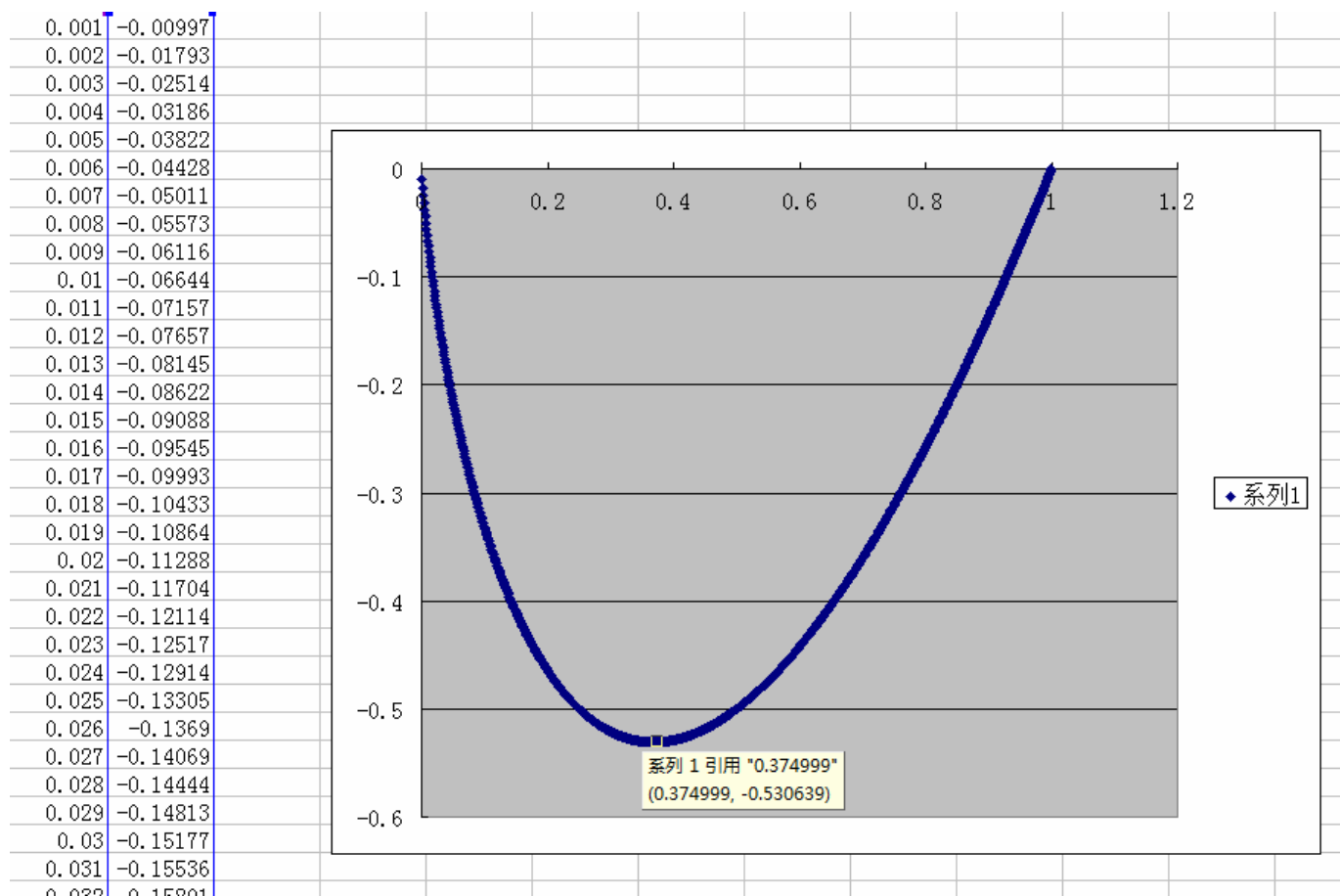
---

```
int _tmain(int argc, _TCHAR* argv[])
{
    float x = 0.001f;
    float y;
    float log2 = log(2.0f);
    ofstream oFile;
    oFile.open(_T("D:\\entropy.txt"));
    while(x < 1)
    {
        y = x * log(x) / log2;
        oFile << x << ' ' << y << '\n';

        x += 0.001f;
    }
    oFile.close();
    return 0;
}
```



# 绘图



# 对熵的理解

---

- 熵是随机变量**不确定性**的度量，不确定性越大，熵值越大；若随机变量退化成定值，熵为0
  - 均匀分布是“最不确定”的分布
- 熵其实定义了一个函数(概率分布函数)到一个值(信息熵)的映射。
  - $P(x) \rightarrow H$  (**函数**  $\rightarrow$  **数值**)
  - **泛函**
    - “变分推导”章节



# 联合熵和条件熵

---

- 两个随机变量 $X$ ,  $Y$ 的联合分布, 可以形成联合熵Joint Entropy, 用 $H(X,Y)$ 表示
- $H(X,Y) - H(Y)$ 
  - $(X,Y)$ 发生所包含的熵, 减去 $Y$ 单独发生包含的熵: 在 $Y$ 发生的前提下,  $X$ 发生“新”带来的熵
  - 该式子定义为 $Y$ 发生前提下,  $X$ 的熵:
    - 条件熵 $H(X|Y)$



# 推导条件熵的定义式

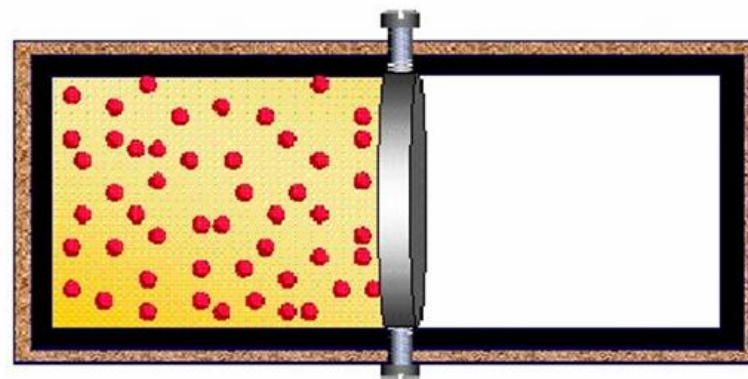
---

$$\begin{aligned} & H(X, Y) - H(Y) \\ &= -\sum_{x,y} p(x, y) \log p(x, y) + \sum_y p(y) \log p(y) \\ &= -\sum_{x,y} p(x, y) \log p(x, y) + \sum_y \left( \sum_x p(x, y) \right) \log p(y) \\ &= -\sum_{x,y} p(x, y) \log p(x, y) + \sum_{x,y} p(x, y) \log p(y) \\ &= -\sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(y)} \\ &= -\sum_{x,y} p(x, y) \log p(x | y) \end{aligned}$$



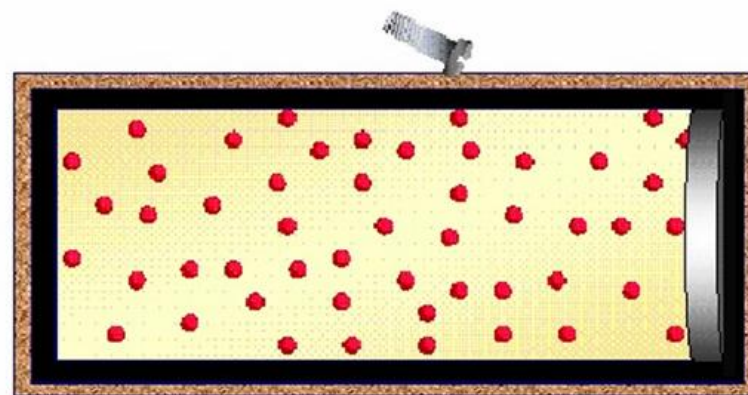
# 自封闭系统的运动总是倒向均匀分布

- 密封箱子中间放一隔板
- 隔板左边空间注入烟，  
右边真空



去掉隔板会怎样？

- 左边的烟就会自然（自发）地向右边扩散，最后均匀地占满整个箱体





# 相对熵

- 相对熵，又称互熵，交叉熵，鉴别信息，Kullback熵，Kullback-Leibler散度等
- 设 $p(x)$ 、 $q(x)$ 是 $X$ 中取值的两个概率分布，则 $p$ 对 $q$ 的相对熵是

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_{p(x)} \log \frac{p(x)}{q(x)}$$

- 两点说明  
■ 在一定程度上，相对熵可以度量两个随机变量的“距离”  
■ 一般的， $D(p \parallel q) \neq D(q \parallel p)$



# 相对熵

- 相对熵，又称互熵，交叉熵，鉴别信息，Kullback熵，Kullback-Leibler散度等
- 设 $p(x)$ 、 $q(x)$ 是 $X$ 中取值的两个概率分布，则 $p$ 对 $q$ 的相对熵是

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_{p(x)} \log \frac{p(x)}{q(x)}$$

- 说明：

- 相对熵可以度量两个随机变量的“距离”
  - 在“贝叶斯网络”、“变分推导”等章节会再次遇到
- 一般的， $D(p \parallel q) \neq D(q \parallel p)$
- $D(p \parallel q) \geq 0$ 、 $D(q \parallel p) \geq 0$  提示：凸函数中的Jensen不等式



# 思考

---

- 假定已知随机变量 $P$ ，求相对简单的随机变量 $Q$ ，使得 $Q$ 尽量接近 $P$ 
  - 方法：使用 $P$ 和 $Q$ 的K-L距离。
  - 难点：K-L距离是非对称的，两个随机变量应该谁在前谁在后呢？
- 假定使用 $KL(Q||P)$ ，为了让距离最小，则要求在 $P$ 为0的地方， $Q$ 尽量为0。会得到比较“窄”的分布曲线；
- 假定使用 $KL(P||Q)$ ，为了让距离最小，则要求在 $P$ 不为0的地方， $Q$ 也尽量不为0。会得到比较“宽”的分布曲线；



# 互信息

---

- 两个随机变量 $X$ ,  $Y$ 的互信息, 定义为 $X$ ,  $Y$ 的联合分布和独立分布乘积的相对熵。
- $I(X, Y) = D(P(X, Y) \parallel P(X)P(Y))$

$$I(X, Y) = \sum_{x, y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$



# 计算 $H(X)-I(X,Y)$

---

$$\begin{aligned} & H(X) - I(X, Y) \\ &= -\sum_x p(x) \log p(x) - \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= -\sum_x \left( \sum_y p(x, y) \right) \log p(x) - \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= -\sum_{x,y} p(x, y) \log p(x) - \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(y)} \\ &= \sum_{x,y} p(x, y) \log p(x | y) \\ &= H(X | Y) \end{aligned}$$



# 整理得到的等式

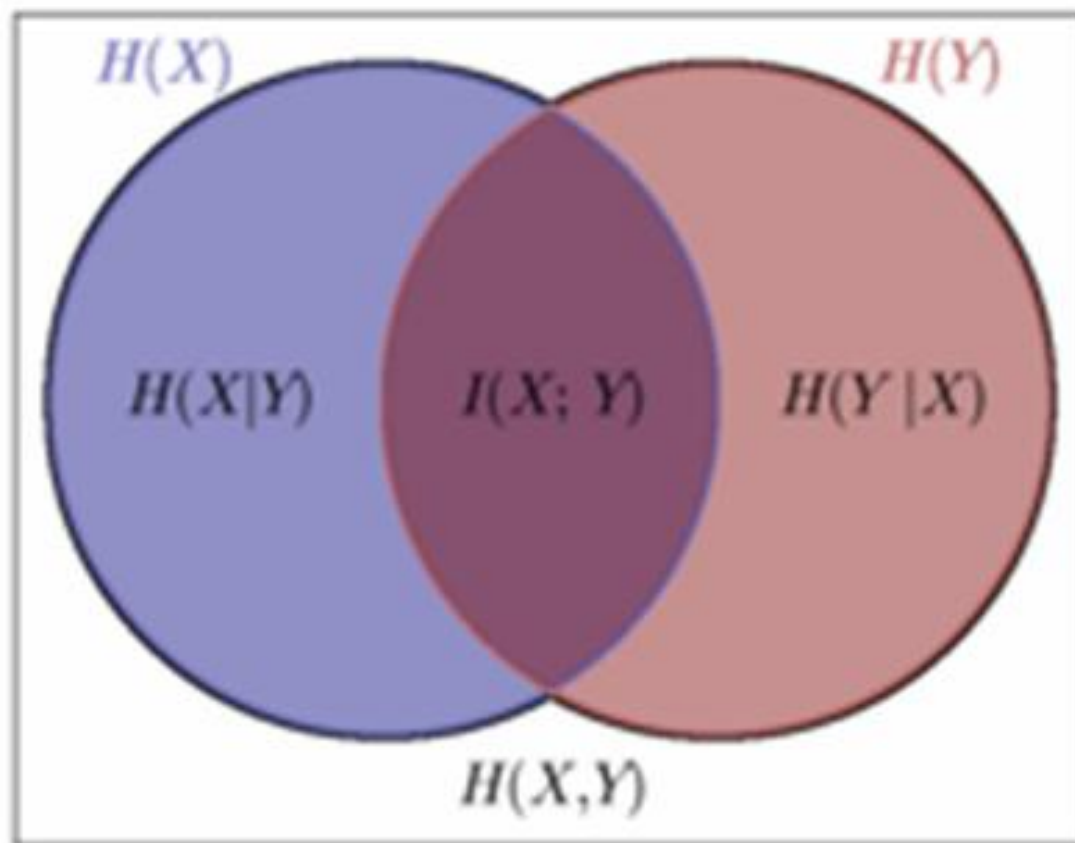
---

- $H(X|Y) = H(X,Y) - H(Y)$ 
  - 条件熵定义
- $H(X|Y) = H(X) - I(X,Y)$ 
  - 根据互信息定义展开得到
  - 有些文献将  $I(X,Y) = H(Y) - H(Y|X)$  作为互信息的定义式
- 对偶式
  - $H(Y|X) = H(X,Y) - H(X)$
  - $H(Y|X) = H(Y) - I(X,Y)$
- $I(X,Y) = H(X) + H(Y) - H(X,Y)$ 
  - 有些文献将该式作为互信息的定义式
- 试证明:  $H(X|Y) \leq H(X)$ ,  $H(Y|X) \leq H(Y)$



# 强大的Venn图：帮助记忆

---



# 最大熵模型的原则

---

- ☐ 承认已知事物(知识)
- ☐ 对未知事物不做任何假设，没有任何偏见





# 两点分布的最大熵

---

□  $H(X) = -p\ln p - (1-p)\ln(1-p)$

■ 思考：如何求最大值？

■ 提示：  $f(p) = -p\ln p - (1-p)\ln(1-p)$

□ 一阶导、二阶导



# X满足均匀分布时，熵最大

---

- 当 $p=0.5$ 时，取 $H(X)$ 取最大值；
- 思考：若“多点”分布呢？
  - $X$ 是随机变量，可以取从1到 $K$ 的 $K$ 个数。问： $X$ 满足什么分布时， $X$ 的熵最大？
  - $p(X)=1/K$ ：均匀分布

$$0 \leq H(X) \leq \log |X|$$



# 例如

□ 已知：

■ “学习”可能是动词，也可能是名词。

■ “学习”可以被标为主语、谓语、宾语、定语……

□ 令 $x_1$ 表示“学习”被标为名词， $x_2$ 表示“学习”被标为动词。

□ 令 $y_1$ 表示“学习”被标为主语， $y_2$ 表示被标为谓语， $y_3$ 表示宾语， $y_4$ 表示定语。得到下面的表示：

$$p(x_1) + p(x_2) = 1 \quad \sum_{i=1}^4 p(y_i) = 1$$

□ 根据无偏原则  $p(x_1) = p(x_2) = 0.5$

$$p(y_1) = p(y_2) = p(y_3) = p(y_4) = 0.25$$



# 引入新知识

---

□ 若已知：“学习”被标为定语的可能性很小，只有0.05  $p(y_4) = 0.05$

□ 仍然坚持无偏原则：

$$p(x_1) = p(x_2) = 0.5$$

$$p(y_1) = p(y_2) = p(y_3) = \frac{0.95}{3}$$



# 再次引入新知识

---

- 当“学习”被标作动词的时候，它被标作谓语的  
概率为0.95

$$p(y_2 | x_1) = 0.95$$

- 除此之外，仍然坚持无偏见原则，尽量使概  
率分布平均。
- 问：怎么样能尽量无偏见的分布？



# 最大熵模型Maximum Entropy

---

- 概率平均分布 等价于 熵最大
- 问题转化为：计算X和Y的分布，使 $H(Y|X)$ 达到最大值，并且满足条件

$$p(x_1) + p(x_2) = 1$$

$$\sum_{i=1}^4 p(y_i) = 1$$

$$p(y_4) = 0.05$$

$$p(y_2 | x_1) = 0.95$$



# 最大熵模型Maxent

---

$$\max H(Y | X) = - \sum_{\substack{x \in \{x_1, x_2\} \\ y \in \{y_1, y_2, y_3, y_4\}}} p(x, y) \log p(y | x)$$

$$p(x_1) + p(x_2) = 1$$

$$p(y_1) + p(y_2) + p(y_3) + p(y_4) = 1$$

$$p(y_4) = 0.05$$

$$p(y_2 | x_1) = 0.95$$



# Maxent的一般式

---

□ 一般模型：

$$\max_{p \in P} H(Y | X) = - \sum_{(x,y)} p(x,y) \log p(y | x)$$

□  $P = \{p \mid p \text{ 是 } X \text{ 上满足条件的概率分布}\}$

■ 注意区分这里的p和P。





# 特征(Feature)和样本(Sample)

---

## □ 特征: $(x, y)$

- $y$ : 这个特征中需要确定的信息
- $x$ : 这个特征中的上下文信息

## □ 样本: 关于某个特征 $(x, y)$ 的样本, 特征所描述的语法现象在标准集合里的分布:

- $(x_i, y_i)$ 对
- $y_i$ 是 $y$ 的一个实例
- $x_i$ 是 $y_i$ 的上下文
- $(x_1, y_1) (x_2, y_2) (x_3, y_3) \dots$



# 特征函数

---

- 特征函数：对于一个特征 $(x_0, y_0)$ ，定义特征函数：

$$f(x, y) = \begin{cases} 1 & x = x_0 \text{ 且 } y = y_0 \\ 0 & \text{otherwise} \end{cases}$$

- 对于一个特征 $(x_0, y_0)$ ，在样本中的期望值是： $\bar{p}(f) = \sum_{(x_i, y_i)} \bar{p}(x, y) f(x, y)$
- $\bar{p}(x, y)$  是 $(x, y)$ 在样本中出现的概率



# 条件Constraints

---

- 对每一个特征 $(x,y)$ , 模型所建立的条件概率分布要与训练样本表现出来的分布相同。
- 假设样本的分布是(已知):

$$\bar{p}(x) = x \text{ 出现的概率}$$

$$\bar{p}(x, y) = xy \text{ 出现的概率}$$

$$\bar{p}(f) = \text{特征} f \text{ 在样本中的期望值}$$



# 条件Constraints

---

□ 特征 $f$ 在模型中的期望值：

$$\begin{aligned} p(f) &= \sum_{(x_i, y_i)} p(x_i, y_i) f(x_i, y_i) \\ &= \sum_{(x_i, y_i)} p(y_i | x_i) p(x_i) f(x_i, y_i) \\ &= \sum_{(x_i, y_i)} p(y_i | x_i) \bar{p}(x_i) f(x_i, y_i) \end{aligned}$$

$$p(f) = \bar{p}(f)$$



# 最大熵模型在NLP中的完整提法

---

□ NLP 模型：

$$p^* = \arg \max_{p \in P} H(Y | X) = - \sum_{(x,y)} p(x,y) \log p(y | x)$$

□  $P = \{p \mid p \text{ 是 } y|x \text{ 的概率分布并且已知条件}\}$

□ 对训练样本，对任意给定的特征  $f_i$ ：

$$p(f) = \overline{p(f)}$$



# 最大熵模型在NLP中的完整提法

---

$$p^* = \arg \max_{p \in P} H(Y | X) = - \sum_{(x,y)} p(x,y) \log p(y | x) = - \sum_{(x,y)} p(y | x) \bar{p}(x) \log p(y | x)$$

$$P = \left\{ p(y | x) \left| \forall f_i : \sum_{(x,y)} p(y | x) \bar{p}(x) f_i(x, y) = \sum_{(x,y)} \bar{p}(x, y) f_i(x, y), \quad \forall x : \sum_y p(y | x) = 1 \right. \right\}$$



# 最大熵模型总结

定义条件熵  $H(y|x) = - \sum_{(x,y) \in Z} p(y,x) \log p(y|x)$

模型目的  $p^*(y|x) = \arg \max_{p(y|x) \in P} H(y|x)$

定义特征函数  $f_i(x, y) \in \{0, 1\} \quad i = 1, 2, \dots, m$

约束条件  $\sum_{y \in Y} p(y|x) = 1 \quad (1)$

$$E(f_i) = \tilde{E}(f_i) \quad i = 1, 2, \dots, m \quad (2)$$

$$\tilde{E}(f_i) = \sum_{(x,y) \in Z} \tilde{p}(x,y) f_i(x,y) = \frac{1}{N} \sum_{(x,y) \in T} f_i(x,y) \quad N = |T|$$

$$E(f_i) = \sum_{(x,y) \in Z} p(x,y) f_i(x,y) = \sum_{(x,y) \in Z} p(x) p(y|x) f_i(x,y)$$



# 求解Maxent模型

- 该条件约束优化问题的Lagrange函数

$$\Lambda(p, \vec{\lambda}) = H(y|x) + \sum_{i=1}^m \lambda_i (E(f_i) - \tilde{E}(f_i)) + \lambda_{m+1} \left( \sum_{y \in Y} p(y|x) - 1 \right)$$

- 分析：

- 已知若干条件，要求若干变量的值使到目标函数(熵)最大

- 数学本质：

- 最优化问题(Optimization Problem)

- 条件：线性、等式

- 目标函数：非线性

- 非线性规划(线性约束)(non-linear programming with linear constraints)





# 拉格朗日函数L

---

$$\begin{aligned} L = & \sum_{(x,y)} p(y|x) \bar{p}(x) \log \frac{1}{p(y|x)} \\ & + \sum_i \lambda_i \sum_{(x,y)} f_i(x,y) [p(y|x) \bar{p}(x) - \bar{p}(x,y)] \\ & + \lambda_0 \left[ \sum_y p(y|x) - 1 \right] \end{aligned}$$



## 最优解形式Exponential: 求偏导, 等于0

$$\frac{\partial L}{\partial p(y|x)} = \bar{p}(x) \left( \log \frac{1}{p(y|x)} - 1 \right) + \sum_i \lambda_i \bar{p}(x) f_i(x, y) + \lambda_0$$

$$p^*(y|x) = e^{\sum_i \lambda_i f_i(x, y) - \frac{\lambda_0}{p(x)} - 1}$$

$$p^*(y|x) = \frac{1}{Z_\lambda(x)} e^{\sum_i \lambda_i f_i(x, y)}$$

$$\sum_y \frac{1}{Z_\lambda(x)} e^{\sum_i \lambda_i f_i(x, y)} = 1 \quad Z_\lambda(x) = \sum_y e^{\sum_i \lambda_i f_i(x, y)}$$



# $\lambda$ 未知

---

- 由  $Z_{\lambda}(x) = \sum_y e^{\sum_i \lambda_i f_i(x,y)}$ ，Maxent模型是对数线性模型
- 因为包含指数函数，几乎不可能有解析解
  - 退一步说：有了解析解，仍然需要数值解
- 能不能找到另一种逼近？构造函数 $f(\lambda)$ ，求其最大/最小值？



# 当前任务

---

## □ 理论问题

- 解释通过最大熵模型建立的目标函数和最大似然估计的关系

## □ 实践问题

- 找到有效的求解 $\lambda$ 的算法

## □ 先解决理论问题

- 发现Maxent和MLE的关系后，有利于 $\lambda$ 的求解



# 最大似然估计Maximum likelihood estimate

---

□ 找出与样本的分布最接近的概率分布模型。

□ 简单的例子

■ 10次抛硬币的结果是：正正反正正正反反正正

□ 假设 $p$ 是每次抛硬币结果为正的概率。则：

□ 得到这样的实验结果的概率是：

$$\begin{aligned} P &= pp(1-p)ppp(1-p)(1-p)pp \\ &= p^7(1-p)^3 \end{aligned}$$



# 极大似然估计MLE

□ 目标函数:  $\max P = \max_{0 \leq p \leq 1} p^7 (1-p)^3$

□ 最优解是:  $p=0.7$

■ 思考: 如何求解?

□ 一般形式:

$$L_{\bar{p}} = \prod_x p(x)^{\bar{p}(x)}$$

$p(x)$ 模型是估计的概率分布  
 $\bar{p}(x)$ 是实验结果的分布

$$\log L(\theta_1, \theta_2, \dots, \theta_k) = \sum_{i=1}^n \log f(x_i; \theta_1, \theta_2, \dots, \theta_k)$$



# 取对数

□ 对数极大似然估计：

$$L_{\bar{p}} = \log \left( \prod_x p(x)^{\bar{p}(x)} \right) = \sum_x \bar{p}(x) \log p(x)$$

$$\begin{aligned} L_{\bar{p}}(p) &= \sum_{x,y} \bar{p}(x,y) \log p(x,y) \\ &= \sum_{x,y} \bar{p}(x,y) \log [\bar{p}(x) p(y|x)] \\ &= \sum_{x,y} \bar{p}(x,y) \log p(y|x) + \sum_{x,y} \bar{p}(x,y) \bar{p}(x) \end{aligned}$$

□ 第二项是常数，可忽略



# MLE与条件熵

---

□ 此目标式，与条件熵具有相同的形式。

$$L_{\bar{p}}(p) = \sum_{x,y} \bar{p}(x,y) \log p(y|x)$$

□ 既然函数式相同，极有可能二者殊途同归，  
目标函数是相同的。

■ 演示推导

$$L = \sum_{(x,y)} p(y|x) \bar{p}(x) \log \frac{1}{p(y|x)} + \sum_{i=1}^k \lambda_i \sum_{(x,y)} f_i(x,y) [p(y|x) \bar{p}(x) - \bar{p}(x,y)] + \lambda_0 \left[ \sum_y p(y|x) - 1 \right]$$





# 求L的对偶函数

□ 最优解  $p_\lambda(y|x) = \frac{1}{Z_\lambda(x)} e^{\sum_i \lambda_i f_i(x,y)}$  代入L得到关于  $\lambda$  的函数  $L(\lambda)$

$$\begin{aligned} &= -\sum_{x,y} p(y|x) \bar{p}(x) \log p(y|x) + \sum_{i=1}^k \lambda_i \sum_{x,y} f_i(x,y) [p(y|x) \bar{p}(x) - \bar{p}(x,y)] + \lambda_0 \left[ \sum_y p(y|x) - 1 \right] \\ &= -\sum_{x,y} p_\lambda(y|x) \bar{p}(x) \log p_\lambda(y|x) + \sum_{i=1}^k \lambda_i \sum_{x,y} f_i(x,y) [p_\lambda(y|x) \bar{p}(x) - \bar{p}(x,y)] \\ &= -\sum_{x,y} \bar{p}(x) p_\lambda(y|x) \log p_\lambda(y|x) + \sum_{i=1}^k \bar{p}(x) p_\lambda(y|x) \lambda_i \sum_{x,y} f_i(x,y) - \sum_{i=1}^k \bar{p}(x,y) \lambda_i \sum_{x,y} f_i(x,y) \\ &= -\sum_{x,y} \bar{p}(x) p_\lambda(y|x) \log p_\lambda(y|x) + \sum_{x,y} \bar{p}(x) p_\lambda(y|x) \sum_{i=1}^k \lambda_i f_i(x,y) - \sum_{i=1}^k \bar{p}(x,y) \lambda_i \sum_{x,y} f_i(x,y) \\ &= \sum_{x,y} \bar{p}(x) p_\lambda(y|x) \log Z_\lambda(x) - \sum_{i=1}^k \bar{p}(x,y) \sum_{x,y} \lambda_i f_i(x,y) \end{aligned}$$



将最大熵最优解  $p_{\lambda}(y|x) = \frac{1}{Z_{\lambda}(x)} e^{\sum_i \lambda_i f_i(x,y)}$  带入MLE

---

$$\begin{aligned} L_{\bar{p}}(p) &= \sum_{x,y} \bar{p}(x,y) \log p(y|x) \\ &= \sum_{x,y} \bar{p}(x,y) \left( \sum_{i=1}^n \lambda_i f_i(x,y) - \log Z_{\lambda}(x) \right) \\ &= \sum_{x,y} \bar{p}(x,y) \sum_{i=1}^n \lambda_i f_i(x,y) - \sum_{x,y} \bar{p}(x,y) \log Z_{\lambda}(x) \\ &= \sum_{x,y} \bar{p}(x,y) \sum_{i=1}^n \lambda_i f_i(x,y) - \sum_x \bar{p}(x) \log Z_{\lambda}(x) \end{aligned}$$



# 结论

---

- 可以看到，二者的右端具有完全相同的目标函数。
- 根据MLE的正确性，可以断定：最大熵的解(无偏的对待不确定性)同时是最符合样本数据分布的解，进一步证明了最大熵模型的合理性。
- 做点思考：
  - 熵：不确定度
  - 似然：与知识的吻合程度
  - 最大熵模型：对不确定度的无偏分配
  - 最大似然估计：对知识的无偏理解

知识 = 不确定度的补集



# $\lambda$ 的求解

---

- 因为没有显式的解析式，使用IIS计算最大熵模型的数值解
  - IIS是目前最大熵模型的最优化算法，优于梯度下降算法
  - IIS, Improved Iterative Scaling, 改进的迭代尺度算法



# 改进的迭代尺度法IIS

---

$$p^*(y|x) = \frac{1}{Z_\lambda(x)} e^{\sum_i \lambda_i f_i(x,y)}$$

$$Z_\lambda(x) = \sum_y e^{\sum_i \lambda_i f_i(x,y)}$$

$$L_{\bar{p}}(p) = \sum_{x,y} \bar{p}(x,y) \sum_{i=1}^n \lambda_i f_i(x,y) - \sum_x \bar{p}(x) \log Z_\lambda(x)$$



# IIS的思想

---

- 假设最大熵模型当前的参数向量是 $\lambda$ ，希望找到新的参数向量 $\lambda + \delta$ ，使得模型的对数似然函数值 $L$ 增加。重复这一过程，直至找到对数似然函数的最大值。



$$L_{\bar{p}}(p) = \sum_{x,y} \bar{p}(x,y) \sum_{i=1}^n \lambda_i f_i(x,y) - \sum_x \bar{p}(x) \log Z_{\lambda}(x)$$

$$L(\lambda + \delta) - L(\lambda)$$

$$= \sum_{x,y} \bar{p}(x,y) \sum_{i=1}^n \delta_i f_i(x,y) - \sum_x \bar{p}(x) \log \frac{Z_{\lambda+\delta}(x)}{Z_{\lambda}(x)}$$

$$\geq \sum_{x,y} \bar{p}(x,y) \sum_{i=1}^n \delta_i f_i(x,y) + 1 - \sum_x \bar{p}(x) \frac{Z_{\lambda+\delta}(x)}{Z_{\lambda}(x)}$$

注:  $-\ln x \geq 1-x, x > 0$

$$= \sum_{x,y} \bar{p}(x,y) \sum_{i=1}^n \delta_i f_i(x,y) + 1 - \sum_x \bar{p}(x) \frac{Z_{\lambda+\delta}(x)}{Z_{\lambda}(x)}$$

$$Z_{\lambda}(x) = \sum_y e^{\sum_i \lambda_i f_i(x,y)}$$

$$= \sum_{x,y} \bar{p}(x,y) \sum_{i=1}^n \delta_i f_i(x,y) + 1 - \sum_x \bar{p}(x) \exp\left(\sum_{i=1}^n \delta_i f_i(x,y)\right)$$

$$= \sum_{x,y} \bar{p}(x,y) \sum_{i=1}^n \delta_i f_i(x,y) + 1 - \sum_x \bar{p}(x) \sum_y p_{\lambda}(y|x) \exp\left(\sum_{i=1}^n \delta_i f_i(x,y)\right)$$



# 针对凸函数 $f(x)=e^x$ 使用Jensen不等式

$$f^\#(x, y) = \sum_i f_i(x, y)$$

$$\begin{aligned} A(\delta | \lambda) &= \sum_{x,y} \bar{p}(x, y) \sum_{i=1}^n \delta_i f_i(x, y) + 1 - \sum_x \bar{p}(x) \sum_y p_\lambda(y | x) \exp\left(\sum_{i=1}^n \delta_i f_i(x, y)\right) \\ &= \sum_{x,y} \bar{p}(x, y) \sum_{i=1}^n \delta_i f_i(x, y) + 1 - \sum_x \bar{p}(x) \sum_y p_\lambda(y | x) \exp\left(f^\#(x, y) \sum_{i=1}^n \frac{\delta_i f_i(x, y)}{f^\#(x, y)}\right) \\ &\geq \sum_{x,y} \bar{p}(x, y) \sum_{i=1}^n \delta_i f_i(x, y) + 1 - \sum_x \bar{p}(x) \sum_y p_\lambda(y | x) \sum_{i=1}^n \frac{f_i(x, y)}{f^\#(x, y)} \exp(\delta_i f^\#(x, y)) \end{aligned}$$





# 对该下界求偏导，令为0，求出 $\delta$

$$B(\delta | \lambda) = \sum_{x,y} \bar{p}(x,y) \sum_{i=1}^n \delta_i f_i(x,y) + 1 - \sum_x \bar{p}(x) \sum_y p_\lambda(y|x) \sum_{i=1}^n \frac{f_i(x,y)}{f^\#(x,y)} \exp(\delta_i f^\#(x,y))$$

$$\begin{aligned} \frac{\partial B(\delta | \lambda)}{\partial \delta_i} &= \sum_{x,y} \bar{p}(x,y) f_i(x,y) - \sum_x \bar{p}(x) \sum_y p_\lambda(y|x) f_i(x,y) \exp(\delta_i f^\#(x,y)) \\ &= \sum_{x,y} \bar{p}(x,y) f_i(x,y) - \sum_{x,y} \bar{p}(x) p_\lambda(y|x) f_i(x,y) \exp(\delta_i f^\#(x,y)) \\ &= E_{\bar{p}}(f_i) - \sum_{x,y} \bar{p}(x) p_\lambda(y|x) f_i(x,y) \exp(\delta_i f^\#(x,y)) \end{aligned}$$

令梯度为0，得到：

$$\sum_{x,y} \bar{p}(x) p_\lambda(y|x) f_i(x,y) \exp(\delta_i f^\#(x,y)) - E_{\bar{p}}(f_i) = 0$$



$\delta$  的求法：若  $f^\#(x,y)=M$  为常数

---

$$\delta_i = \frac{1}{M} \log \frac{E_{\bar{p}}(f_i)}{E_p(f_i)}$$



## $\delta$ 的求法：若 $f^\#(x,y)$ 不是常数

□ 令  $g(\delta_i) = \sum_{x,y} \bar{p}(x)p_\lambda(y|x)f_i(x,y)\exp(\delta_i f^\#(x,y)) - E_{\bar{p}}(f_i)$

■ 转换为求  $g(\delta) = 0$  的根。

□ 牛顿法：

$$\delta_i^{(k+1)} = \delta_i^{(k)} - \frac{g(\delta_i^{(k)})}{g'(\delta_i^{(k)})}$$

□ 说明：

- 因为需要计算  $g(\delta) = 0$  的根而不是求  $g(\delta)$  的极小值，上式是函数值除以一阶导，而不是一阶导除以二阶导；
- 实践中，可采用拟牛顿BFGS或者L-BFGS的方法。



# 最终解

---

□ 上述求解过程中得到的权值  $\lambda$ ，回代到下式中，即得到最大熵模型的最优估计。

$$p^*(y|x) = \frac{1}{Z_\lambda(x)} e^{\sum_i \lambda_i f_i(x,y)}$$

$$Z_\lambda(x) = \sum_y e^{\sum_i \lambda_i f_i(x,y)}$$



# 再次强调

---

- 熵是描述不确定度的
- 知识是不确定度的补集
  - 不确定度越小，模型越准确。
- 直观的过程：
  - 什么特征都不限定：熵最大
  - 加一个特征：熵少一点
    - Condition Reduces Entropy (C.R.E.)
  - 加的特征越多，熵越少



# 生态学

## 生态群落结构中各物种的相对多度问题

- Shipley, B. and Vile, D. and Garnier, E.
- “From plant traits to plant communities: a statistical mechanistic approach to biodiversity”  
Science, 2006
- 在法国南部选取12块样地
- 研究30种植物的叶、茎、叶子等相关的8个形状与各物种相对多度之间的关系
  - 用群落水平上各性状的平均值作为约束条件
  - 用最大熵原理对各物种在不同群落中的相对多度进行预测
  - 结果与实测数据有很高的吻合度 (94%)



# 总结

---

- MaxEnt已经是比较成功的一个NLP模型，并获得广泛应用
- 从信息论获得启发(1948-)：自然语言处理也是信息处理的一种。
  - 词性标注也可以看作一种编码的过程？
  - 思考：身边的哪些问题，可以看做或类别编码过程？
- 求极值的技术手段：Lagrange对偶问题
- 最大熵模型，涉及了很多前序的数学知识
  - 事实上，机器学习本身就是多种手段的综合应用。



# Last words on MaxEnt...

---

- ☐ All Models are wrong. Some are useful.





# 参考文献

---

- ☐ Elements of Information Theory (Cover & Thomas)
- ☐ Linear and Nonlinear Programming (Nash & Sofer)
- ☐ A maximum entropy approach to natural language processing (Adam Berger)
- ☐ A Brief MaxEnt Tutorial (Adam Berger)
- ☐ Learning to parse natural language with maximum entropy models (Adwait Ratnaparkhi)
- ☐ A simple Introduction to Maximum Entropy Models for Natural Language Processing (Adwait Ratnaparkhi)
- ☐ 统计学习方法，李航著，清华大学出版社，2012年



---

谢谢大家！

恳请大家批评指正！

