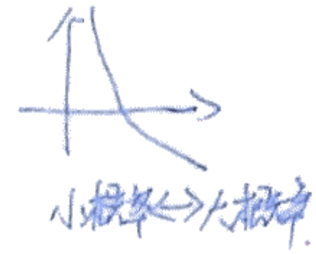
概率2：（来自Logistic & Softmax回归一节）

(29) 信息熵（见概率④） $H(P)=-\sum_{k=1}^{K}(p_k\ln p_k)$ 是n个物品装到K个箱子有多少种装法，在 $n\to+\infty$ 时的极限值。熵越高，表示混乱程度越高，分法越多，分到K个箱子的概率越趋于相同.

(30) 信息量：$h(x)=-\log_2 p(x)$ 小概率事件发生携带的信息量大



ⓐ 独立事件 X,Y 满足 $h(\underbrace{XY}_{X,Y同时发生})=h(X)+h(Y)$

$h(XY)=-\log_2 P(XY)=-\log_2(P(X)\cdot P(Y))=-\log_2 P(X)-\log_2 P(Y)=h(X)+h(Y)$

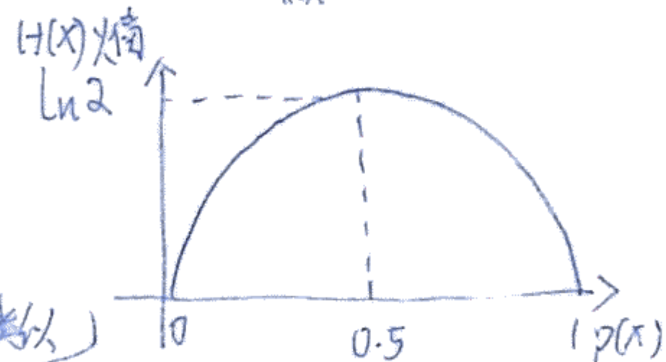
ⓑ 用信息量也可以定义信息熵：即随机事件信息量的数学期望.

$H(X)=\sum_{x\in X}p(x)h(x)=\sum_{x\in X}p(x)(-\cancel{\log_2}\ln(p(x)))=\boxed{-\sum_{x\in X}p(x)\ln(p(x))}$

底是e还是2不要紧，可以用换底公式替换.

(31) 概率分布和熵的关系

ⓐ 两点分布的熵：$H(X)=-\sum_{x\in X}p(x)\ln(p(x))=-p\ln p-(1-p)\ln(1-p)$ //根据熵的定义



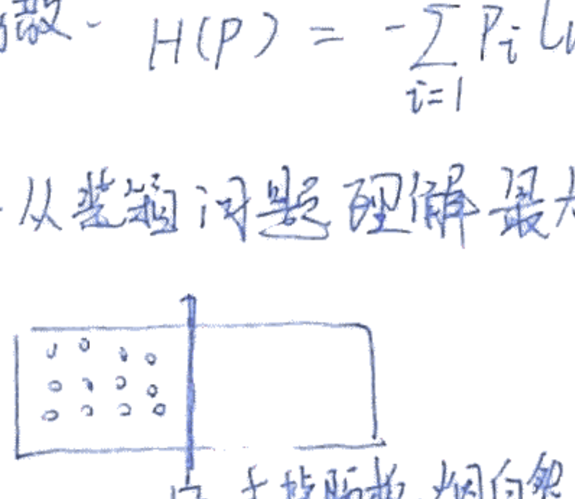
⇒ 当 $p(x=0)=p(x=1)=0.5$ 时，信息熵最高，混乱程度最高.

ⓑ 三点分布的熵（也类似）：$H(X)=-\sum_{x\in X}p(x)\ln(p(x))=-p_1\ln p_1-p_2\ln p_2-p_3\ln p_3$ 其中 $p_1+p_2+p_3=1$

ⓒ 均匀分布的熵：N表示某离散分布可取N个值，概率都是 $1/N$

离散：$H(P)=-\sum_{i=1}^{N}p_i\ln p_i=-\sum_{i=1}^{N}\frac{1}{N}\ln\frac{1}{N}=\ln N$ ⇒ 离散分布信息熵位于 $[0,\log N]$

(32) 从装箱问题理解最大熵公式



熵是随机变量不确定性的度量，不确定性越大，熵越大

熵越大 { 随机变量退化成定值，熵最小，为0.
　　　　随机变量为均匀分布，熵最大.

以上为无条件的最大熵分布

若有条件呢？给定期望和方差前提下，最大熵的分布形式呢

(33) 如果 $\ln(p(x))$ 是 $\alpha x^2+\beta x+\gamma$ 形式，则 $p(x)$ 服从正态分布

证明：利用 $\ln(p(x))$ 计算过程的可逆性

$$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ // 正态分布概率密度

$$\ln p(x)=\ln\frac{1}{\sqrt{2\pi}}-\ln\sigma-\frac{\boxed{(x-\mu)^2}}{2\sigma^2}=\alpha x^2+\beta x+\gamma$$ // 对数正态分布

该对数正态分布是关于随机变量 x 的二次函数，且计算过程可逆

(34) 如果 $\ln(p(x))$ 是 $A\cdot x+B\ln(x)+C$ 的形式，则 $p(x)$ 服从Gamma分布

证明：

总结：Beta分布，Gamma函数 ⇒ 概率15

Gamma函数：$\Gamma(\alpha)=\int_0^{\infty}t^{\alpha-1}e^{-t}dt$

$\Gamma(n)=(n-1)!$ 是阶乘在实数域的推广

Gamma分布 $$f(x;\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\cdot x^{\alpha-1}e^{-\beta\cdot x}\qquad x\geq 0\ (\text{常参数}\ \alpha,\beta>0)$$

数学期望 $E(x)=\frac{\alpha}{\beta}$

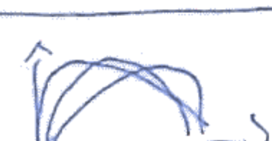
△

对数Gamma分布：

$$\ln f(x;\alpha,\beta)=\alpha\ln\beta+(\alpha-1)\ln x-\beta x-\ln\Gamma(\alpha)$$
$$=A\cdot x+B\cdot\ln x+C$$

该计算过程可逆

对比Beta分布



抛物线下面积

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

数学期望是 $\frac{\alpha}{\alpha+\beta}$

△

㉟给定方差的最大熵分布：带约束条件求极值问题，用拉格朗日乘子法

目标函数：$\underset{p(x)}{\arg\max}\ H(x) = -\sum_x p(x)\ln p(x)$　s.t. $\begin{cases} E(x)=\mu & ⓐ \\ Var(x)=\sigma^2 \Leftrightarrow E(x^2)=(E(x))^2+Var(& ⓑ \end{cases}$

$= \mu^2+\sigma^2$

根据约束条件ⓐ,ⓑ 用拉格朗日乘子法建模

根据方差公式 $Var(x)=E(x^2)-(E(x)$

$$\mathcal{L}(p) = -\sum_x p(x)\cdot\ln p(x) + \lambda_1(E(x)-\mu) + \lambda_2(E(x^2)-\mu^2-\sigma^2)$$

$$= -\sum_x p(x)\ln p(x) + \lambda_1\left(\sum_x x\cdot p(x)-\mu\right) + \lambda_2\left(\sum_x x^2 p(x)-\mu^2-\sigma^2\right)$$

要求p(x)即p为何值时，$H(x)=-\sum_x p(x)\ln p(x)$ 即 $-\sum_x p\ln p$ 值最大

对p求偏导，另偏导为0，得到驻点

$$\frac{\partial L}{\partial p} = \frac{\partial\left(-\sum_x p\ln p + \lambda_1\left(\sum_x x\cdot p-\mu\right)+\lambda_2\left(\sum_x x^2\cdot p-\mu^2-\sigma^2\right)\right)}{\partial p}$$

分部求导公式

$$= -\ln p - 1 + \lambda_1 x + \lambda_2 x^2$$

令偏导为0，得到

$$\ln p(x) = \ln p = \lambda_2 x^2 + \lambda_1 x + 1$$

是关于随机变量x的二次形，根据引理（概率㉝），

p服从正态分布

$\Rightarrow$ $\boxed{\text{最大熵模型是建立在p服从正态分布的假设之上}}$

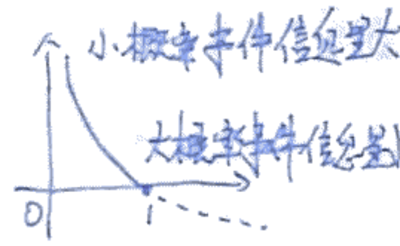
㊱ 联合熵 $H(X,Y) = -\sum_{x,y} p(x,y)\ln(p(x,y))$

随机变量 X, Y 联合分布的信息熵

对比信息熵定义（概率④㉙㊴）

$H(z) = -\sum_{z} p(z)\ln p(z)$

其中 $-\ln p(z)$ 是信息量



$p(z)$ 是概率密度，

$H(z)$ 也可理解为信息量的数学期望。

㊲ 条件熵 $H(Y|X) = H(X,Y) - H(X)$ ⓐ

*表示 X 发生的前提下，X,Y 同时发生"新"带来的熵（混乱度）

即事件X前提下Y发生

$H(Y|X) = -\sum_{x,y} p(x,y)\ln p(y|x)$ ⓑ 信息量

$= -\sum_{x,y} \boxed{p(y|x)\cdot p(x)} \underbrace{\ln p(y|x)}_{}$

?

概率密度

*也可以理解为条件概率信息量的数学期望（

增

*公式ⓐ等价于公式ⓑ

$H(Y|X) = H(X,Y) - H(X)$ //条件熵定义式

$= -\sum_{x,y} p(x,y)\ln p(x,y) + \sum_{x} p(x)\ln p(x)$ //信息熵，联合熵定义式

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

$= H(Y) - \underline{I(X,Y)}$

↓

互信息：见

概率㊴㊵

$= -\sum_{x,y} p(x,y)\ln p(x,y) + \sum_{x}\left(\sum_{y} p(x,y)\right)\ln p(x)$

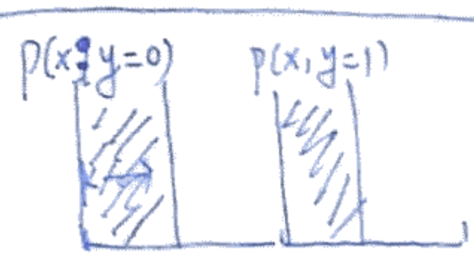
$= -\sum_{x,y} p(x,y)\ln p(x,y) + \sum_{x,y} p(x,y)\ln p(x)$

$= -\sum_{x,y} p(x,y)\ln \frac{p(x,y)}{p(x)}$

$= -\sum_{x,y} p(x,y)\ln p(y|x)$ //公式ⓑ

没用 $p(y|x)$

因为要表示信息增量



$p(x,y)$ 的分母是整个概率空间

$p(x|Y)$ 的分母是 $p(y)$，不能表示增量

$= -\sum_{x} p(x)\cdot \underline{p(y|x)\ln p(y|x)}$

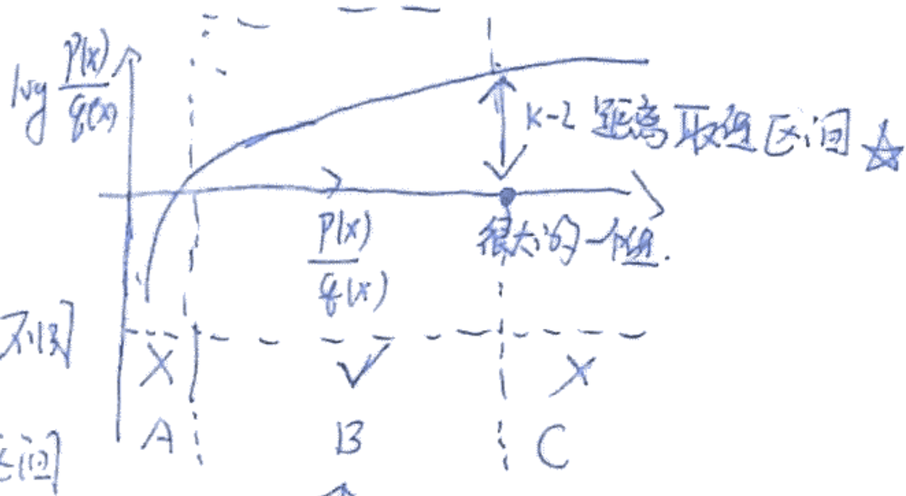
↓

$= \sum_{x} p(x)\,\underline{H(Y|X=x)}$ ⓒ

*等价于公式ⓒ：$\boxed{\text{给定 } X=x \text{ 时，} Y \text{ 的熵的数学期望}}$

㊳ 相对熵（也叫互熵，交叉熵，鉴别信息，Kullback熵，Kullback-leible散度，简称K-L距离），可以理解为对 $\log \frac{p(x)}{q(x)}$ 求数学期望.

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



难点：D(p||q) 与 D(q||p) 选用哪个？

因为K-L距离是非对称的，上面2个值不同

选择依据：尽量让 $\frac{p(x)}{q(x)}$ 或 $\frac{q(x)}{p(x)}$ 落在B区间

这样K-L距离取值区间不会太夸张

两个K-L散度的区别：PPT

概率 58

㊴ 互信息：随机变量X, Y的 ⓐ$\boxed{\text{联合分布}}$ 与 $\boxed{\text{ⓑ独立分布}}$的乘积 的 $\boxed{\text{相对熵}}$

$$I(X,Y) = D(p(x,y)\,\|\,p(x)p(y))$$

$$= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)q(y)}$$

表示没有X时Y的不确定性与给定X时Y的不确定性的差值

㊵ 条件熵与互信息关系

$$H(Y|X) = H(X,Y) - H(X)$$ // 定义式：X发生前提下，X,Y同时发生新带来的混乱度

$$H(Y|X) = \underbrace{H(Y)}_{\text{不考虑X时Y不确定性}} - \underbrace{I(X,Y)}_{\text{差值}}$$ // 表示~~给定~~考虑X时，Y的不确定性. 即

推导：$H(Y) - I(X,Y)$

$$= -\sum_y p(y)\log p(y) - \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

$$= -\sum_{x,y} p(x,y)\log(p(y)) - \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

$$= -\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)}$$

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

$$= H(Y|X)$$

(41) 互信息与信息熵，联合熵之间的关系

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

证明： $H(X) + H(Y) - H(X,Y)$

$$= -\sum_x p(x)\log p(x) - \sum_y p(y)\log p(y) + \sum_{x,y} p(x,y)\log p(x,y)$$

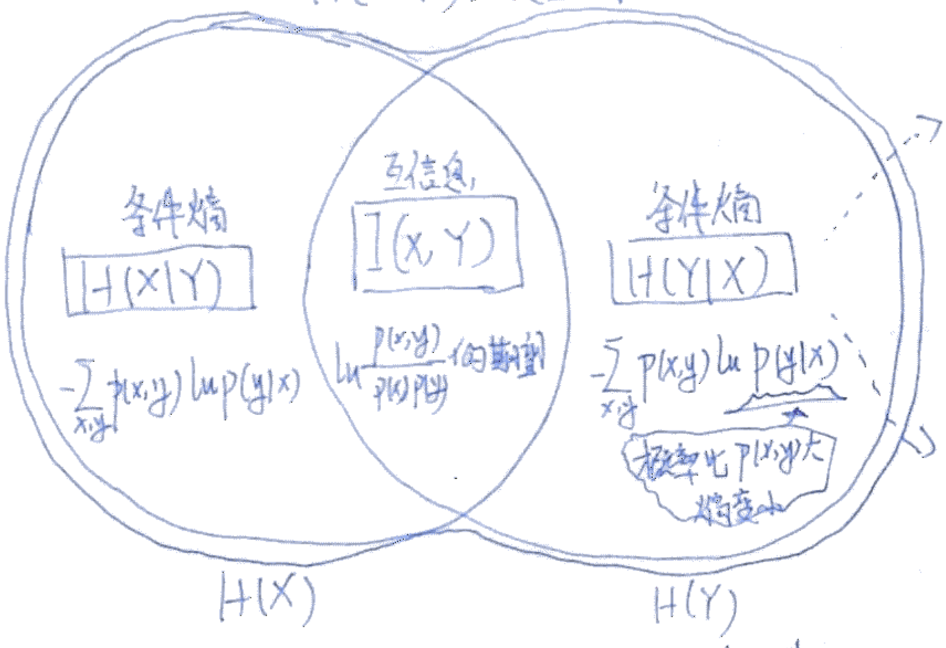
$$= -\sum_{x,y} p(x,y)\log p(x) - \sum_{x,y} \log(x,y)\log p(y) + \sum_{x,y} p(x,y)\log(x,y)$$

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

(42) Venn图：信息熵，联合熵，条件熵，互信息之间的关系.

$-\sum_{x,y} p(x,y)\ln p(x,y)$ 概率越小，混乱程度越高，熵越高

$H(X,Y)$：联合熵，X,Y联合分布的信息熵



X的信息熵 $-\sum_x p(x)\ln p(x)$

Y的信息熵 $-\sum_Y p(y)\ln p(y)$

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

X发生前提下

X,Y同时发生(概率小熵大)

新带来的不确定性

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

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

不考虑X取值，Y自身的不确定性

差值

X发生前提下 Y的不确定性

指 $\boxed{P(y|x)}$

(不是 $p(x,y)$)

概率大熵小