1. 事件 X 的信息量(自信息)

$I(X)=-\log(P(X))$定义为概率分布的对数的相反数。一个随机产生的事件 X 所包含的自信息数量，只与事件发生的几率相关。事件发生的几率越低，在事件真的发生时，接收到的信息中，包含的自信息越大【不可能性大】。

概率大，出现机会多，不确定性小；反之就大。

$$I(X)=f(P(X))$$

如果$\mathrm{P}(X)=1$，那么 $\mathrm{I}(X)=0$。如果$\mathrm{P}(X)<1$，那么 $\mathrm{I}(X)>0$。

根据定义，自信息的量度是非负的而且是可加的。如果事件C是两个独立事件A和B的交集，那么宣告C发生的信息量就等于分别宣告事件A和事件B的信息量的和：$I(C)=I(A\cup B)=I(A)+I(B)$

因为A和B是独立事件， C的概率为$\mathrm{P}(C)=\mathrm{P}(A\cup B)=\mathrm{P}(A)\cdot\mathrm{P}(B)$

应用函数$f(\cdot)$会得到

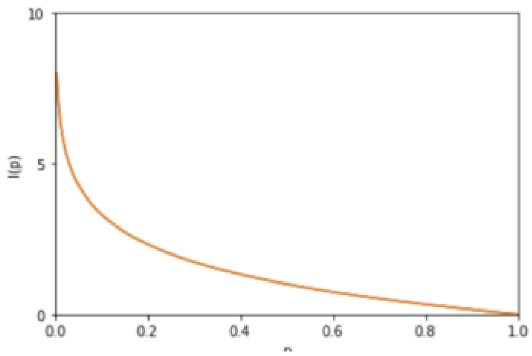
$$\mathrm{I}(C)=\mathrm{I}(A)+\mathrm{I}(B)$$

$$f(\mathrm{P}(C))=f(\mathrm{P}(A)\cdot\mathrm{P}(B))=f(\mathrm{P}(A))+f(\mathrm{P}(B))$$

所以函数$f(\cdot)$有性质$f(x\cdot y)=f(x)+f(y)$，因此可采用定义$f(x)=K\log(x)$

由于事件的概率总是在 0 和 1 之间，而信息量必须是非负的，所以 $K<0$

如果以 2 为底，单位是 bit。当使用以 e 为底的对数时，单位将是 nat。对于基底为 10 的对数，单位是 hart。



随机变量 X 的熵

随机变量信息量的均值(即期望)为该分布产生的信息量的平均值(即熵)

$$H(X)=E[I(X)]$$

$$H(X)=-\sum_{x}P(x)\log P(x)$$

描述随机变量 X 的随机性或不确定性的量度。不确定性越大，熵越大。

P=0/1，对熵的计算没有贡献)

具有均匀概率分布的信源符号集可以有效地达到最大熵$\log(n)$logb(n)：所有可能的事件是等概率的时候，不确定性最大。

- 连续性

该量度应连续，概率值小幅变化只能引起熵的微小变化。

- 对称性

重新排序后，该量度应不变。

- 极值性

当所有变量有同等机会出现的情况下，熵达到最大值(所有可能的事件同等概率时不确定性最高)。

等概率事件的熵随变量数量增加而增加。

- 可加性

熵的量与该过程如何被划分无关。

计算(X,Y)得到的熵或信息量(即同时计算 X 和 Y)等于通过进行两个连续实验得到的信息：先计算 Y 的值，然后在知道 Y 的值条件下得出 X 的值：

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

2. 条件熵$\mathbf{H}(\mathbf{Y}|\mathbf{X})$

表示在已知随机变量 X 的条件下随机变量 Y 的不确定性(利用可加性)

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

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

条件熵$\mathrm{H(Y|X)}$相当于联合熵减去单独的熵$\mathrm{H}(X)$

3. 两个分布的交叉熵

相同事件测度的两个概率分布 p 和 q 的**交叉熵**是指，用非真实分布 q(x)来表示来自真实分布 p(x)的平均编码长度，在事件集合中唯一标识一个事件所需要的平均比特数(bit)。

基于 p(x)分布的，q 的平均信息量。【基于 p(x)分布的，q 的信息量的平均值】

给定两个概率分布p和q，以p为基准的，p和q的交叉熵定义为：

$$H(p,q)=\mathrm{E}_p[-\log q]=H(p)+D_{\mathrm{KL}}(p\|q)$$

其中$H(p)$是p的熵，$D_{\mathrm{KL}}(p\|q)$是 KL 散度(也被称为以p**为基准的**，p和q的相对熵)。

对于离散分布p和q：

$$H(p,q)=\sum_x p(x)(-\log q(x))=-\sum_x p(x)\log q(x)$$

交叉熵误差(cross entropy error)也经常被用作损失函数

$$E=-\sum_k t_k\log\quad y_k$$

y_k是神经网络的输出，t_k是正确解标签。并且，t_k中只有正确解标签的索引为1，其他均为0(one-hot 表示)。正确解标签对应的输出越大，式的值越接近0；当输出为1 时，交叉熵误差为0

4. 相对熵(KL 散度)

如果用 P 来描述目标问题，而不是用 Q 来描述目标问题，得到的信息增量。

KL 散度(Kullback-Leibler Divergence，简称 KLD)，在信息系统中称为相对熵(relative entropy)，在连续时间序列中称为随机性(randomness)，在统计模型推断中称为**信息增益**(information gain)。也称信息散度(information divergence)。

P原始分布，q近似的分布

$$KL(p||\mathrm{q})=\int_{-\infty}^{+\infty}\mathrm{p(x)}\log\frac{\mathrm{p(x)}}{\mathrm{q(x)}}\mathrm{dx}=\int_{-\infty}^{+\infty}\mathrm{p(x)}(\log\mathrm{p(x)}-\log\mathrm{q(x)})\mathrm{dx}=E[\log\mathrm{p(x)}-\log\mathrm{q(x)}]$$

表示原始分布 p 和近似分布 q 之间的对数差值的期望(以 p 为基准，评价 q 和 p 的差异，所以权重是p(x))KL 散度非对称的,不能作为距离度量。

Q 的分布越接近 P(Q 分布越拟合 P)，散度值越小，即损失值越小。

对数函数是凸函数，所以 KL 散度的值为非负数

(利用 Jensen 不等式,EM 算法推导也可以类似 https://en.wikipedia.org/wiki/Jensen%27s_inequality)。

当且仅当$\mathrm{p==q}$时，$\mathrm{KL(p||q)=0}$。

$$\mathrm{H}(p,q)=\mathrm{E}_p[-\log q]=\mathrm{H}(p)+D_{\mathrm{KL}}(p\|q)$$

KL 散度 = 交叉熵 - 熵。也就是说，熵固定，KL 散度和交叉熵在优化时等价。(深度学习时采用)

$$D_{\mathrm{KL}}(p\|q)=\mathrm{H}(p,q)-H(p)$$

5. JS 散度

$$\mathrm{JS}(p,q)=\frac{1}{2}\mathrm{KL}(p||\frac{1}{2}(p+q))+\frac{1}{2}\mathrm{KL}(q||\frac{1}{2}(p+q))$$

(用 2 为底的对数计算，则 K-L 散度值表示信息损失的二进制位数)

JS 散度是对称的，其取值是 0 到 1 之间。

6. 互信息(mutual Information，简称 MI)

在概率论和信息论中，两个随机变量的互信息(mutual Information，简称 MI)或转移信息(transinformation)是变量间相互依赖性的量度。决定着联合分布$p(x,y)$和分解的边缘分布的乘积 $p(x)p(y)$的相似程度。互信息是点间互信息(PMI)的期望值。互信息最常用的单位是 bit。

相关性大，互信息越大。

一般地，两个离散随机变量 X 和 Y 的互信息可以定义为：

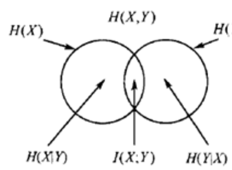
$$I(X;Y)=\sum_{y\in Y}\sum_{x\in X}p\,(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

在连续随机变量的情形下：

$$I(X;Y)=\int_Y\int_X p\,(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy$$

其中$p(x,y)$当前是X和Y的联合概率密度函数，而$p(x)$和 $p(y)$ 分别是X和Y的边缘概率密度函数。

如果对数以 2 为基底，互信息的单位是 bit。



直观上，互信息度量 X 和 Y 共享的信息：它度量知道这两个变量其中一个，对另一个不确定度减少的程度。例如，如果 X 和 Y 相互独立，则知道 X 不对 Y 提供任何信息，反之亦然，所以它们的互信息为零。在另一个极端，如果 X 是 Y 的一个确定性函数，且 Y 也是 X 的一个确定性函数，那么传递的所有信息被 X 和 Y 共享：知道 X 决定 Y 的值，反之亦然。因此，在此情形互信息与 Y(或 X)单独包含的不确定度相同，称作 Y(或 X)的熵。而且，这个互信息与 X 的熵和 Y 的熵相同。(这种情形的一个非常特殊的情况是当 X 和 Y 为相同随机变量时。)

互信息是 X 和 Y 的联合分布相对于假定 X 和 Y 独立情况下的联合分布之间的内在依赖性。 于是互信息以下面方式度量依赖性：I(X;Y) = 0 当且仅当 X 和 Y 为独立随机变量。从一个方向很容易看出：当 X 和 Y 独立时，p(x,y) = p(x) p(y),互信息为 0。

互信息非负，对称。

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X,Y) \\
&= H(X,Y) - H(X|Y) - H(Y|X)
\end{aligned}
$$

互信息的直观意义为: "因 X 而有 Y 事件"的熵(基于已知随机变量的不确定性) 在"Y 事件"的熵之中具有多少影响地位("Y 事件所具有的不确定性" 其中包含了多少 "Y|X 事件所具有的不确性")，意即"Y 具有的不确定性"有多少程度是起因于 X 事件;

互信息越小，两个来自不同事件空间的随机变量彼此之间的关系性越低; 互信息越高，关系性则越高 。$H(X) = I(X;X)$

$$I(X;Y) = D_{\mathrm{KL}}(p(x,y) \| p(x)p(y))$$

互信息和 KL 散度

$$I(X;Y) = D_{\mathrm{KL}}(P(X,Y) \| P(X)P(Y)) = \mathbb{E}_X\{D_{\mathrm{KL}}(P(Y|X) \| P(Y))\} = \mathbb{E}_Y\{D_{\mathrm{KL}}(P(X|Y) \| P(X))\}$$