熵与交叉熵

Yimeng Ren

1 熵与交叉熵

在信息论中，熵用来衡量一个随机事件的不确定性。

1.1 自信息

自信息（Self Information）表示一个随机事件所包含的信息量。一个随机事件发生的概率越高，其自信息越低. 如果一个事件必然发生，其自信息为 0。

自变量 $X \in \mathbf{X}, X \sim P(x)$，自信息定义为：

$$I(x) = -\log p(x)$$

1.2 熵

熵为自信息的期望。对于分布为 $p(x)$ 的随机变量 X, 其熵（自信息的期望）的定义如下：

$$H(X) = E_X[I(x)] = E_X[-\log p(x)] = -\sum_{x \in \mathbf{X}} p(x) \log p(x),\ where\ log(0) = 0$$

熵越大，则随机变量的信息越多；熵越小，则随机变量的信息越小。如果对于一个确定的信息，那么熵为零，信息量也为零。如果一个概率分布为一个**均匀分布，则熵最大**。

对于 0-1 分布，$p = P(y = 1)$，$H(x) = -p \log p - (1-p) \log(1-p)$ 也可以看作对数似然。

1.3 交叉熵

对于分布为 $p(x)$ 的随机变量，熵 $H(p)$ 表示其最优编码长度。交叉熵是按照概率分布 q 的最优编码对真实分布 p 的信息进行编码的长度。

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

其中 p 为真实分布，q 可以看作估计分布。在给定 p 的情况下，如果 p 和 q 越接近，那么交叉熵越小；如果 p 和 q 越远，交叉熵越大。

1.4 K-L 散度（一种距离表示）

KL 散度也叫做 KL 距离或者相对熵，是用概率分布 q 来近似 p 时所造成的信息损失量。对于离散概率分布 p 和 q, 从 q 到 p 的 KL 散度定义为:

$$KL(p,q)=H(p,q)-H(p)=\sum_x p(x)\log\frac{p(x)}{q(x)}$$

KL 散度总是非负的，即 $KL(p,q)\geq 0$，可以衡量两个概率分布之间的距离。只有当 $p=q$ 时，$KL(p,q)=0$。

1.5 交叉熵损失

1.5.1 二分类

在二分的情况下，模型最后需要预测的结果只有两种情况，对于每个类别我们的预测得到的概率为 p 和 $1-p$。此时表达式为:

$$L=\frac{1}{N}\sum_i L_i=\frac{1}{N}\sum_i -[y_i\cdot\log(p_i)+(1-y_i)\cdot\log(1-p_i)]$$

注意，这里 $y_i\in\{0,1\}$ 为真实值。

1.5.2 多分类

多分类的情况实际上就是对二分类的扩展:

$$L=\frac{1}{N}\sum_i L_i=\frac{1}{N}\sum_i -\sum_{c=1}^{M} y_{ic}\log(p_{ic})$$

M 为类别的数量。

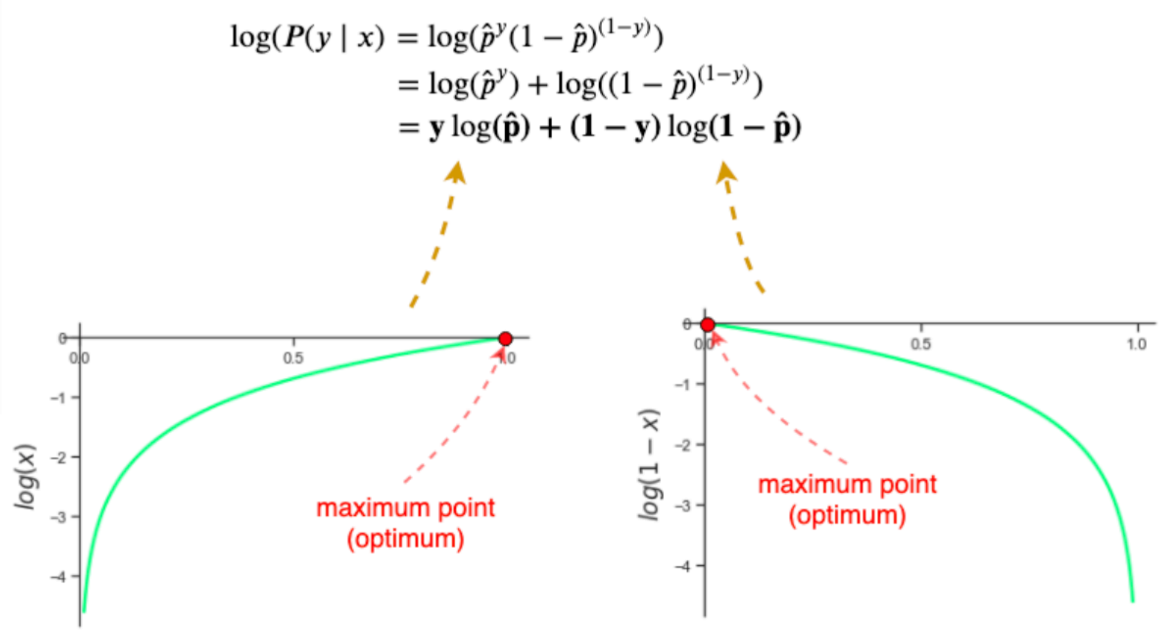
$\log(P(y \mid x) = \log(\hat{p}^y(1-\hat{p})^{(1-y)})$
$= \log(\hat{p}^y) + \log((1-\hat{p})^{(1-y)})$
$= \mathbf{y}\log(\hat{\mathbf{p}}) + (\mathbf{1}-\mathbf{y})\log(\mathbf{1}-\hat{\mathbf{p}})$
log(x)
maximum point
(optimum)
log(1 − x)
maximum point
(optimum)

图 1: 二分类交叉熵损失