Softmax

定义

熵

系统的不确定性程度，或系统的混乱程度。

信息熵

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{17}$$

X：随机变量
$p(x)$：X的概率函数

实例

X	Probability
1	0.6
2	0.1
3	0.1
4	0.1
5	0.1

$$H(X) = -0.6 \log 0.6 + 4 \times -0.1 \log 0.1 \approx 0.53308 \tag{18}$$

相对熵（KL散度）

两个概率分布之间的非对称性度量

$$D_{KL}(p||q) = \sum_{i=1}^{n} p(x_i) \log \left(\frac{p(x_i)}{q(x_i)} \right) \tag{19}$$

$$D_{KL}(p||q) = H(P,Q) - H(P) \tag{20}$$

KL散度=交叉熵-信息熵

交叉熵

主要应用：度量随机变量X的预测分布Q与真实分布P之间的差距

$$H(P,Q) = -\sum_{i=1}^{n} p(x_i) \log q(x_i) \tag{21}$$

实例

俺给大家打个比方：大家当分类1 2 3为猫狗鼠，识别图片给出预测。

分类	预测值$Q(x)$	真实标签$P(x)$
1	0.7	1
2	0.1	0
3	0.2	0

$$H(P,Q) = -1\log 0.7 = 0.1549 \tag{22}$$

相对准确的预测。

分类	预测值$Q(x)$	真实标签$P(x)$
1	0.3	1
2	0.6	0
3	0.1	0

$$H(P,Q) = 0.5229 \tag{23}$$

不准确的预测。

分类	预测值$Q(x)$	真实标签$P(x)$
1	0.1	0
2	0.1	0
3	0.8	1

$$H(P,Q) = 0.0969 \tag{24}$$

几乎准确的预测。

交叉熵：总结

1. 预测越准确，交叉熵越小。
2. 交叉熵只和真实标签的预测概率有关。(真实标签为one hot时)

最简公式

$$\text{Corss\_Entropy}(p,q) = -\log q(c_i) \tag{25}$$

Sigmoid函数

$$S(x) = \frac{1}{1+e^{-x}} \tag{26}$$

一些损失函数

L2 Loss

$$l(y, y') = \frac{1}{2}(y - y')^2 \tag{27}$$

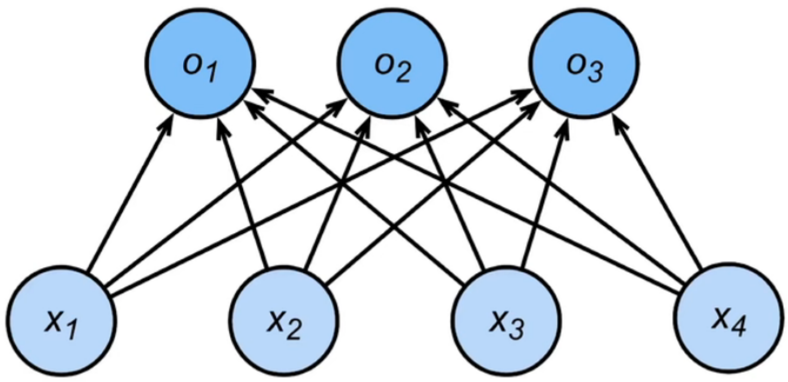
L1 Loss

$$l(y, y') = |y - y'| \tag{28}$$

Huber's Robust Loss

$$l(y, y') = \begin{cases} |y - y'| - \frac{1}{2} & if\,|y - y'| > 1 \\ \frac{1}{2}(y - y')^2 & otherwise \end{cases} \tag{29}$$

Softmax回归



softmax在于将输出匹配概率（非负、和为1）

处理输出

$$\begin{aligned} \hat{\boldsymbol{y}} &= softmax(\boldsymbol{o}) \\ \hat{y}_i &= \frac{\exp(o_i)}{\sum_k \exp(o_k)} \end{aligned} \tag{30}$$

损失函数

交叉熵（衡量两个概率的区别）

$$l(y, \hat{y}) = -\sum_i y_i \log \hat{y}_i = -\log \hat{y}_y \tag{31}$$

因为是one hot编码，只有y_y是1，其余都是0，因此只有符合预期的预测概率$\hat{y}_y$有效。

只关心对正确类的预测值。

现在对o_i求损失函数的梯度

$$\begin{aligned} l(y, \hat{y}) &= -\sum_k y_k \log \hat{y}_k = \log \sum_k \exp(o_k) - \sum_k y_k o_k \\ \partial_{o_k} l(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \frac{\exp(o_k)}{\sum_k \exp(o_k)} - y_k = \text{softmax}(o_k) - y_k \\ &\text{特别的，只有}y_y = 1, \text{其余}y_k = 0, k \neq y \end{aligned} \tag{32}$$

梯度是预测概率与真实概率的差异。

动手做：图片分类