## Recall: Squared loss case

我们现在指代任意一个 $y^{(i)}$ 为 $y$，其 $\hat{y}^{(i)}$ 为 $\hat{y}$，对应的 $x^{(i)}$ 为 $x$：

$$l(y,\hat{y}) = \frac{1}{2}(y-\hat{y})^2$$

$$\begin{bmatrix}\hat{y}^{(1)}\\ \hat{y}^{(2)}\\ \vdots\\ \hat{y}^{(n)}\end{bmatrix} = \begin{bmatrix}x_1^{(1)} x_2^{(1)} \cdots x_k^{(1)}\\ x_1^{(2)} x_2^{(2)} \cdots x_k^{(2)}\\ \vdots\\ x_1^{(n)} x_2^{(n)} \cdots x_k^{(n)}\end{bmatrix} \begin{bmatrix}w_1\\ w_2\\ \vdots\\ w_k\end{bmatrix}$$

$$\hat{y}^{(i)} = w_1 x_1^{(i)} + w_2 x_2^{(i)} + \cdots + w_k x_k^{(i)} = [w_1\ w_2 \cdots w_k] \cdot \begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_k\end{bmatrix} = w^T x^{(i)}$$

## Jacobian matrix:

对于 $f: \mathbb{R}^n \to \mathbb{R}^m$

$$x = \begin{bmatrix}x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix} \mapsto \begin{bmatrix}f_1(x)\\ f_2(x)\\ \vdots\\ f_m(x)\end{bmatrix} \quad f(x)$$

$$J = \begin{bmatrix}\frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n}\\ \vdots & \vdots & \cdots & \vdots\\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n}\end{bmatrix} \begin{matrix}\to (\nabla_x f_1)^T\\ \to (\nabla_x f_2)^T\\ \\ \to (\nabla_x f_m)^T\end{matrix}$$

可记作 $(\nabla_x f)^T$

chain rule: $\nabla_x f = (\nabla_x u)^T \cdot \nabla_u f$

## By chain rule:

$$\nabla_w l(y,\hat{y}) = (\nabla_{\hat{y}} l(y,\hat{y}))^T (\nabla_w \hat{y})$$
$$= (\nabla_{\hat{y}} \tfrac{1}{2}(y-\hat{y})^2)^T (\nabla_w w^T x)$$
$$= (\hat{y}-y)^T (x) = (w^T x - y)\cdot x$$

## Now: cross-entropy loss.

($o_j = w_j \cdot x = w_{j1}x_1 + \ldots + w_{jk}x_k$)

$$l(y,\hat{y}) = -\sum_{j=1}^{q} y_j \ln \hat{y}_j$$

$\hat{y}_j = \frac{\exp(o_j)}{\sum_{k=1}^{q}\exp(o_k)} = \frac{e^{o_j}}{e^{o_1}+e^{o_2}+\ldots+e^{o_q}}$

ex: $y = \begin{bmatrix}y_1\\ y_2\\ y_3\end{bmatrix} = \begin{bmatrix}1\\ 0\\ 0\end{bmatrix}$, $\hat{y} = \begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \hat{y}_3\end{bmatrix} = \begin{bmatrix}0.7\\ 0.2\\ 0.1\end{bmatrix}$

$$\Rightarrow l(y,\hat{y}) = -y_1\ln\hat{y}_1 - y_2\ln\hat{y}_2 - y_3\ln\hat{y}_3$$
$$= -y_1\ln\hat{y}_1 = -\ln\hat{y}_1 \quad (y_2, y_3 \text{ 为 } 0)$$
$$= \ln\frac{1}{\hat{y}_1}$$

总结 (观察) $l(y,\hat{y}) = \ln\frac{1}{\hat{y}_\alpha}$, where $y_\alpha$ 为 1.

$y_\alpha$ 乘上 hard label 的 probability 的 -log 值.

于是计算 $\nabla_{w_j} l(y,\hat{y}) = (\partial_{o_j} l(y,\hat{y}))(\nabla_{w_j} o_j)$

($\partial_{o_j} l(y,\hat{y})$ → 一个数; $\nabla_{w_j} o_j = x$)

(注意这里 $W = \begin{bmatrix}w_{11} & w_{12} & \cdots & w_{1k}\\ w_{21} & w_{22} & \cdots & w_{2k}\\ \vdots & \vdots & \cdots & \vdots\\ w_{q1} & w_{q2} & \cdots & w_{qk}\end{bmatrix}$ 为一个 matrix

我们求的是对于其中任一个 $j^{th}$ row$^T$: $w_j$ 这个 vector 的 grad)

因为 $l(y,\hat{y}) = -\sum_{j=1}^{q} y_j \ln\left(\frac{\exp(o_j)}{\sum_{k=1}^{q}\exp(o_k)}\right)$

$$= \sum_{j=1}^{q} y_j \left(\ln\left(e^{o_1}+e^{o_2}+\ldots+e^{o_q}\right) - \ln(e^{o_j})\right)$$
$$= \sum_{j=1}^{q} y_j \left(\ln\left(e^{o_1}+\ldots+e^{o_q}\right)\right) - \sum_{j=1}^{q} y_j o_j$$
$$= y_\alpha\left(\ln\left(e^{o_1}+\ldots+e^{o_q}\right)\right) - o_\alpha$$
$$= \ln\left(e^{o_1}+\ldots+e^{o_q}\right) - o_\alpha$$

因而对其中任一个 row $j$,

$\partial_{o_j} l(y,\hat{y}) = \frac{\exp(o_j)}{\sum_{k=1}^{q}\exp(o_k)} - y_j = \text{softmax}(o)_j - y_j = \hat{y}_j - y_j$

($y_j = 0$, if $j \neq \alpha$; $y_j = 1$, if $j = \alpha$)

($o = \begin{bmatrix}o_1\\ o_2\\ \vdots\\ o_q\end{bmatrix}$, $\text{softmax}(o) = \begin{bmatrix}\text{softmax}(o)_1\\ \text{softmax}(o)_2\\ \vdots\\ \text{softmax}(o)_q\end{bmatrix}$,

$\text{softmax}(o)_j = \frac{e^{o_j}}{e^{o_1}+e^{o_2}+\ldots+e^{o_q}}$)

因而 $\nabla_{w_j} l(y,\hat{y}) = (\text{softmax}(o)_j - y_j)\cdot x$

$= (\hat{y}_j - y_j)\cdot x$ ✡

因而我们可以用 $w_j := w_j - \eta \nabla_{w_j} l(y,\hat{y})$

来更新每个 $w_j$. 用一个循环来更新 $w_1 \sim w_q$

也可以用 matrix form: $W \in \mathbb{R}^{q\times k}$

因而 $\nabla_W l(y,\hat{y}) \in \mathbb{R}^{q\times k}$. ($\nabla_W$ 一定和 W 形状一样)

$$\nabla_W l = \begin{matrix}\uparrow\\ q\\ \downarrow\end{matrix}\begin{bmatrix}(\nabla_{w_1} l(y,\hat{y}))^T\\ (\nabla_{w_2} l(y_1,\hat{y}))^T\\ \vdots\\ (\nabla_{w_q} l(y_1,\hat{y}))^T\end{bmatrix}$$
$$\leftarrow k \rightarrow$$

$$W := W - \eta \nabla_W l$$

## Entropy

(like a,b...,z)

— Suppose: encode an "alphebat" where symbol $j$ occurs with prob $P(j)$,

— The encoding need to be binary (0/1)

— A good encoding 会 assign shorter codes to frequent symbols.

比如 a 出现很高频，那么 a 的 encode 应该很短；q 只出现很少次，那么 q 的 encode 应该很长.

— optimal 的 encoding：

$j^{th}$ letter 的 binary code 的长度约为 $\simeq \log_2 \frac{1}{P(j)}$. (即 $-\log_2 P(j)$)

— 所有 bits on average: $-\sum_j P(j) \log_2 P(j)$

定义 $H[P]$ 为分布 $P$ 的 entropy，代表其 uncertainty 的大小.

$$H[P] = -\sum_j P(j) \log P(j) \quad \left(= \sum_j P(j) \log \frac{1}{P(j)}\right)$$

其以 nats 为单位，1 nat $\approx$ 1.44 bits.

可以想到 $P$ 为 uniform distribution 时，$H[P]$ 最大，因为这个时候什么都不知道. 如果 $P(A) = 60\%$ 那么我们知道 $A$ 最可能发生. 而 uniform distribution 下一步每个事件发生可能性相等，uncertainty 最大.

## Cross Entropy

(对于 discrete variables)

$$H(P, Q) = -\sum_j P(j) \ln Q(j)$$

($P$: 真实分布，$Q$: 预测分布)

其意义是：在 true distribution $P$ 的条件下，用预测分布 $Q$ 来编码事件所需的信息量. 把 $-\ln Q(j)$ 这个长度 encode 给 $j$ 这个 symbol. $H(P, Q)$ 越低，表示 $Q$ 越接近 $P$.

## KL divergence 或称 relative entropy

如果我们知道两个概率分布有差异，并且我们期望 $Q$ 比 $P$ 的信息量稍大一些.

我们可以使用

KL divergen 表示：如果我们需要用 $P$ 来 encode $Q$，我们所需的额外信息量.

$$KL(P \| Q) = H(P, Q) - H(P)$$

$$= \sum_j P(j) \ln \frac{P(j)}{Q(j)}$$