Entropy: $H(Y) = \sum -P(y) \cdot \log(P(y))$ $\uparrow \Rightarrow$ uncertainty $\uparrow$

$H(Y|X=x) = \sum -P(y|x) \log_2 P(y|x)$

$H(Y|X) = \sum_{x \in X} \underbrace{p(x)} \cdot H(Y|X=x)$

$H(X,Y) = H(X|Y) + H(Y)$, $H(Y|Y) = 0$; $H(Y|X) \le H(Y)$

$IG(Y|X) = H(Y) - H(Y|X) \ge 0$

if completely uninformative about Y: $H(Y|X) = H(Y)$

$\nabla_w \, w^T a w = (a + a^T) w$

$\nabla_w \, Aw = A^T$

$\nabla_w \, w^T x = x$

$\nabla_w \, \underbrace{\|w\|^2}_{= w^T w} = 2w$

Bias - Variance Decomposition:

$$\mathbb{E}_D\left[\mathbb{E}_{p(t)}[(y-t)^2]\right] = \underbrace{\left(y_* - \mathbb{E}_D[y]\right)^2}_{\text{bias}} + \underbrace{Var(y)}_{\text{variance}} + \underbrace{Var(t)}_{\text{bayes error}} \quad , y_* = \mathbb{E}(t|x).$$

Linear regression: $y = Xw$, $X = \begin{bmatrix} 1 & [x^{(1)}]^T \\ 1 & \vdots \end{bmatrix}$, $w = \begin{bmatrix} b \\ w_1 \\ \vdots \end{bmatrix}$

$J(w) = \frac{1}{2} \|y - t\|^2 = \frac{1}{2} \| Xw - t\|^2$

$\nabla_w J = X^T X w - X^T t = 0$

Binary classification: $z = w^T x$, $y = \begin{cases} 1, & \text{if } z \ge 0 \\ 0, & \text{if } z < 0. \end{cases}$

Logistic regression: $z = w^T x$

$y = \sigma(z) = \frac{1}{1+e^{-z}}$ (sigmoid)

$\mathcal{L}_{CE}(y,t) = -t \log y - (1-t) \log(1-y)$

(cross-entropy)

$\mathcal{L}_{LCE}(z,t) = t \log(1+e^{-z}) + (1-t) \log(1+e^{z})$

$\frac{\partial \mathcal{L}_{CE}}{\partial w_j} = \frac{\partial \mathcal{L}_{CE}}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial w_j} = \left(-\frac{t}{y} + \frac{1-t}{1-y}\right) \cdot y \cdot (1-y) \cdot x_j = (y-t) \cdot x_j$

$\Rightarrow w_j \leftarrow w_j - \alpha \cdot \frac{\partial J}{\partial w_j} = w_j - \alpha \cdot \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)}) x_j^{(i)}$

backpropagation

$x, b, w \to z \to y \to L \to \mathcal{L}_{reg}$; $t \to L$; $w \to R \to \mathcal{L}_{reg}$

$\overline{\mathcal{L}_{reg}} = 1$

$\overline{L} = \overline{\mathcal{L}_{reg}} \cdot \frac{\partial \mathcal{L}_{reg}}{\partial L}$

$\overline{R} = \overline{\mathcal{L}_{reg}} \cdot \frac{\partial \mathcal{L}_{reg}}{\partial R}$

$\overline{y} = \overline{L} \cdot \frac{\partial L}{\partial y}$

$\overline{z} = \overline{y} \cdot \frac{\partial y}{\partial z}$

$\overline{w} = \overline{z} \cdot \frac{\partial z}{\partial w} + \overline{R} \cdot \frac{\partial R}{\partial w}$

where $\overline{y} := \frac{\partial L}{\partial y}$

Softmax regression (multi-class):

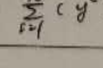
$z = Wx$

$y = \text{softmax}(z)$, $\frac{e^{z_k}}{\sum_{k'} e^{z_{k'}}} = \text{softmax}(z)_k$

$\mathcal{L}_{CE} = -t^T (\log y) = -\sum_{k=1}^{K} t_k \cdot \log y_k$

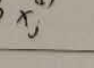
$\frac{\partial \mathcal{L}_{CE}}{\partial w_k} = (y_k - t_k) \, x$

$w \leftarrow w_k - \alpha \cdot \frac{1}{N} \sum_i^N (y_k^{(i)} - t_k^{(i)}) x^{(i)}$



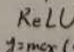
ReLU

$y = \max(0, z)$



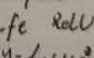
Soft ReLU

$y = \log(1+e^{z})$



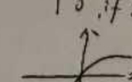
Hard Threshold

$y = \begin{cases} 1, & \text{if } z > 0 \\ 0, & \text{if } z \le 0 \end{cases}$



Logistic

$y = \frac{1}{1+e^{-z}}$



Hyperbolic Tangent

$y = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$