Entropy: $H(Y) = \sum -p(i) \cdot \log(p(i))$ ↑ ⇒ uncertainty ↑

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

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

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

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

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

$\nabla_w \; w^T A w = (A + A^T) w$

$\nabla_w \; A w = A^T$

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

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

Bias - Variance Decomposition:

$$\mathbb{E}_D\left[\mathbb{E}_{p(t)}[(y-t)^2]\right] = \left(y_x - \mathbb{E}_D[y]\right)^2 + Var(y) + Var(t)$$

, $y_x = \mathbb{E}(t|x)$.

bias² variance bayes error

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

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

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

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

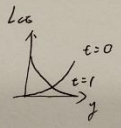
Logistic regression:

$z = w^T x$

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

$L_{CE}(y,t) = -t \log y - (1-t) \log(1-y)$

(cross-entropy)

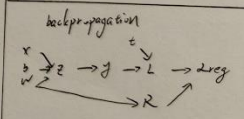


$L_{LCE}(w, z, t) = t \log(1+e^{-z}) + (1-t) \log(1+e^{z})$

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

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

backpropagation



$\overline{L_{reg}} = 1$

$\bar{L} = \overline{L_{reg}} \cdot \frac{\partial L_{reg}}{\partial L}$

$\bar{R} = \overline{L_{reg}} \cdot \frac{\partial L_{reg}}{\partial R}$

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

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

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

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

Softmax regression: (multi-class):

$z = Wx$

$y = \text{softmax}(z)_\theta$ $\frac{e^{z_k}}{\sum_i e^{z_i}} = \text{softmax}(z)_k$

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

$\frac{\partial L_{CE}}{\partial w_k} = (y_k - t_k) x$

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