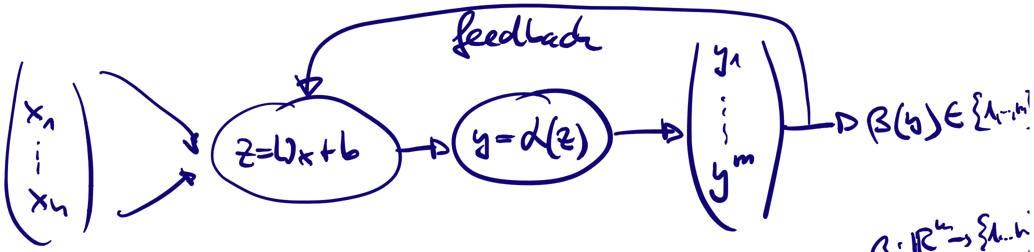
Multiclass Classification

$m \in \mathbb{N}$ classes, $n \in \mathbb{N}$ inputs



$$W \in \mathbb{R}^{m \times n} \qquad \alpha : \mathbb{R}^m \to \mathbb{R}^m$$

$$b \in \mathbb{R}^m \qquad z \mapsto \begin{pmatrix} \alpha(z) \\ \vdots \\ \alpha(z_n) \end{pmatrix}$$

Convention: for all functions $f : \mathbb{R} \to \mathbb{R}$ we def. $f : \mathbb{R}^n \to \mathbb{R}^n$

$$x \mapsto \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

Loss function

$$L(W, b) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(y^{(i)}, \alpha(Wx^{(i)} + b)\right)$$

Update:

$$W \mapsto W^{new} := W - \eta \frac{\partial L}{\partial W}$$

$$b \mapsto b^{new} := b - \eta \frac{\partial L}{\partial b}$$

Example: (1) Argmax Logistic regression model

Training data $(x^{(i)}, y^{(i)})_{1 \le i \le N}$ $\quad y^{(i)} \in \{1 \dots m\}$

$$\alpha(z) = \frac{1}{1 + e^{-z}} \in (0, 1), \quad p(x^{(i)}) = \alpha(W \cdot x^{(i)} + b)$$

$$y = \alpha(Wx + b)$$

$$P\left(y_j^{(i)} = p_j(x^{(i)})\right) = p_j(x^{(i)})^{y_j^{(i)}} \left[1 - p_j(x^{(i)})\right]^{1 - y_j^{(i)}}$$

$$P\left(y^{(i)} = p(x^{(i)})\right) = \prod_{j=1}^{m} p_j(x^{(i)})^{y_j^{(i)}} \left[1 - p_j(x^{(i)})\right]^{1 - y_j^{(i)}}$$

$$P\left(y^{(i)} = p(x^{(i)}) \text{ for } i = 1 \dots N\right)$$
$$= \prod_{i=1}^{N} \prod_{j=1}^{m} p_j(x^{(i)})^{y_j^{(i)}} \left[1 - p_j(x^{(i)})\right]^{1 - y_j^{(i)}}$$

$$-\log P\left(y^{(i)} = p(x^{(i)}) \text{ for } i = 1 \dots N\right)$$

$$= -\sum_{i=1}^{N} \sum_{j=1}^{m} \left[y_j^{(i)} \log p_j(x^{(i)}) + (1 - y_j^{(i)}) \log\left(1 - p_j(x^{(i)})\right) \right]$$

$$L(w,b) = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{m}\left[y_j^{(i)}\log p(x_j^{(i)}) + (1-y_j^{(i)})\log\left(1-p(x_j^{(i)})\right)\right]$$

$\beta(y) :=$ choose element from $\operatorname{argmax}_{i=1\ldots m} y_i$

$\operatorname{argmax}_{i=1\ldots m} y_i := \{ j = 1\ldots m \mid y_j \geq y_i \text{ for all } i = 1\ldots m \}$

(2) Argmax Softmax model

$$\alpha(z) = \left(\frac{e^{z_i}}{\sum_{j=1}^{m} e^{z_j}}\right)_{i=1\ldots m} \in (0,1)$$

models the argmax by putting large z_i on a different scale than small ones.

everything btw above

$$\frac{d\alpha(z)}{dz_i} = \frac{e^{z_i}}{\sum_{j=1}^{m} e^{z_j}} - \frac{e^{2z_i}}{\left(\sum_{j=1} e^{z_j}\right)^2}$$

$$= \alpha(z)\,(1-\alpha(z))$$

So that Cross-entropy is a good loss function.