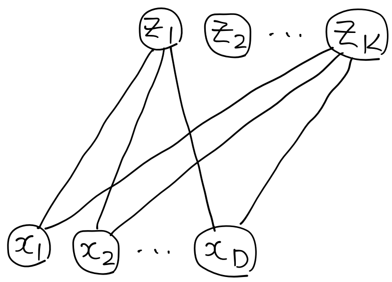
1. What weight are used in multiclass classification?



$$z_1 = \vec{w}_1^T \vec{x} + b_1 \qquad z_K = \vec{w}_K^T \vec{x} + b_K$$

$$\begin{pmatrix} w_{11} \\ w_{12} \\ \vdots \\ w_{1D} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix} \qquad \begin{pmatrix} w_{K1} \\ w_{K2} \\ \vdots \\ w_{KD} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix}$$

same K linear models

$$\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_K \end{pmatrix} \qquad \vec{z} = W\vec{x} + \vec{b} \qquad \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{pmatrix}$$

$$\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1D} \\ \vdots & & & \vdots \\ w_{K1} & & \cdots & w_{KD} \end{pmatrix}$$

how do we obtain W & $\vec{b}$? vid 22

2. provide examples of softmax

$$\vec{y} = \mathrm{softmax}(\vec{z}) = \left[\frac{e^{z_k}}{\sum_{k'=1}^{K} e^{z_{k'}}} \right]_{k \in 1 \cdots K}$$

$$\vec{z} = [3,-1,4]^T \rightarrow \vec{y} = [0.3, 0, 0.7] \quad \frac{e^4}{e^3+e^4}$$

$$\vec{z} = [200, 10, -4]^T \rightarrow \vec{y} = [1, 3\cdot 10^{-83}, 2.5\cdot 10^{-89}]$$

$$\approx [1, 0, 0]$$

3. More intuition behind connection between softmax & logistic activation?

recall $y = \sigma(z) = \dfrac{1}{1+e^{-z}}$

(binary) $\mathcal{E} = -t\log(y) - (1-t)\log(1-y)$

(multi-class) $\mathcal{E} = -\sum_{k=1}^{K} t_k \log(y_k)$

$\vec{t} = [0, 0, \cdots 0, 1, 0, \cdots 0]$

index $j = 1$ if sample is in class j

take $K=2$ (2 classes)

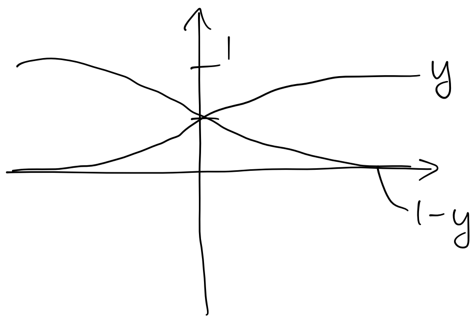
$$\mathcal{E} = -\sum_{k=1}^{2} t_k \log(y_k)$$

$$= -t_1\log(y_1) - t_2\log(y_2)$$

let $t_2 = 1 - t_1$, $y_2 = 1 - y_1$ $\Big\}$ $\mathcal{E} = -t_1\log(y_1) - (1-t_1)\log(y_2)$

what about softmax?

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



- sum always 1 for softmax

observe

$$y = \frac{1}{1+e^{-z}} = \frac{e^{z}}{e^{z}+e^{0}}$$

$$1-y = \frac{1}{1+e^{z}} = \frac{e^{0}}{e^{0}+e^{z}}$$

$$\begin{pmatrix} y \\ 1-y \end{pmatrix} = \begin{pmatrix} \frac{e^{z}}{e^{z}+e^{0}} \\ \frac{e^{0}}{e^{0}+e^{z}} \end{pmatrix} = \text{softmax}\begin{pmatrix} z \\ 0 \end{pmatrix}$$

$z \gg 0$ output 1

$z \ll 0$ output 0

4. How does cross-entropy encourage correct preds?

$$\mathcal{E} = -\sum_{k=1}^{K} t_k \log\left(\frac{e^{z_k}}{\sum_{k'=1}^{K} e^{z_{k'}}}\right)$$

$$= -\sum_{k=1}^{K} t_k \left(z_k - \log\left(\sum_{k'=1}^{K} e^{z_{k'}} \right) \right)$$

$=1$ only if $k=j$

should be huge for z_j

$$-\log\left(\sum_{k'=1}^{K} e^{z_{k'}} \right)$$

Let $m = \arg\max_k z_k$

if $m=j$ and $z_m >> z_k$ for $k \neq j$

$$\log\left(\sum_{k'=1}^{K} e^{z_{k'}} \right) \approx \log(e^{z_m}) \approx z_m = z_j$$

↳ $\varepsilon \approx -(z_j - z_j) = 0$

if $m \neq j \Rightarrow \varepsilon \simeq -(z_j - z_m)$

encourage correct prediction

penalize incorrect prediction

5. How to decide on # neurons/layers & activation function

- model zoo
  - ↳ CNN: images, videos
  - ↳ language models (attention)

- hyperparameter training
  - ↳ grid / random search
- neural architecture search