Lecture -7 .

08/02/25

Gumbel Distribution

PDF $\quad \text{Gumbel}(\mu,\beta) = \dfrac{e^{-(z+e^{-z})}}{\beta} \qquad z = \dfrac{x-\mu}{\beta}.$

$\downarrow$

CDF $\quad P(X \le x) = e^{-e^{-\frac{(x-\mu)}{\beta}}}$

$\mu \to$ location param.

$\beta > 0 \to$ scale param.

For Gumbel(0,1) $\Rightarrow \boxed{F_X^{-1}(y) = -\log_e(-\log_e(y))}$.

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\beta} e^{-(z+e^{-z})} \, dx$$

$\boxed{z = \dfrac{x-\mu}{\beta}}$

$dz = \dfrac{1}{\beta} dx.$

$$= \int_{-\infty}^{\frac{x-\mu}{\beta}} \frac{1}{\cancel{\beta}} e^{-(z+e^{-z})} \cancel{\beta} \, dz$$

when
$x \to -\infty$
$z \to -\infty$

when
$x \to x$
$z \to \dfrac{x-\mu}{\beta}$

$$= \int_{-\infty}^{\frac{x-\mu}{\beta}} e^{-(z+e^{-z})} dz = \int_{-\infty}^{\frac{x-\mu}{\beta}} \underline{e^{-z}} \, e^{-e^{-z}} \, \underline{dz}$$

$e^{-z} = u.$

$\underline{du} = \underline{-e^{-z} dz}$

$$= \int_{\infty}^{e^{-\frac{(x-\mu)}{\beta}}} e^{-u} \cdot -du$$

when, $z \to -\infty$, $u \to \infty$

$z = \dfrac{x-\mu}{\beta}$, $u \to e^{-\frac{(x-\mu)}{\beta}}$

$$= -\int_{\infty}^{e^{-\frac{(x-\mu)}{\beta}}} e^{-u}\,du = +\left[\frac{e^{-u}}{+1}\right]_{\infty}^{e^{-\frac{(x-\mu)}{\beta}}}$$

$$= \left[e^{-e^{-\frac{(x-\mu)}{\beta}}} - e^{-\infty}\right]$$

$$\frac{1}{e^{\infty}} = \frac{1}{\infty} = 0.$$

$$\boxed{F_X(x) = e^{-e^{-\frac{(x-\mu)}{\beta}}}}$$

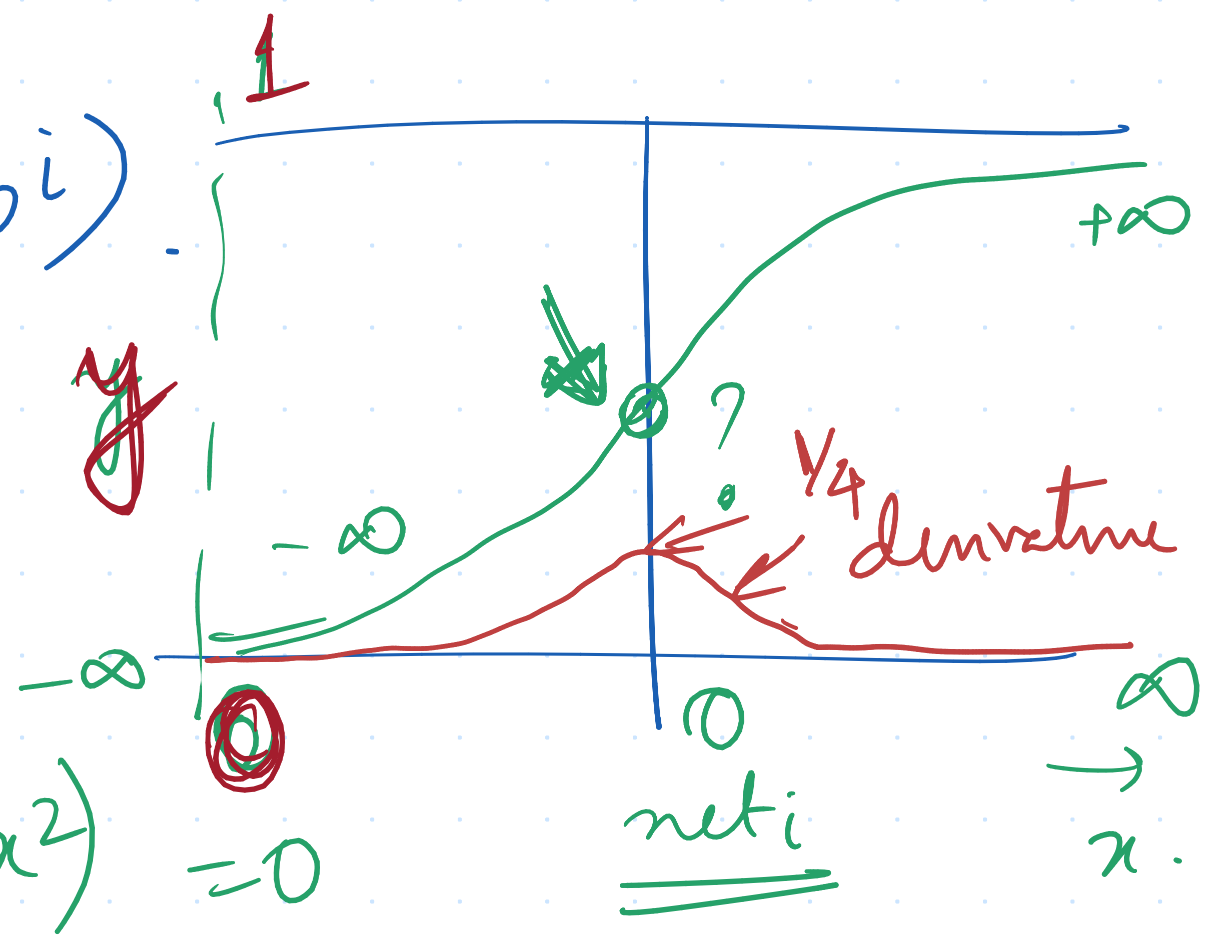
Recap:-

$$o^i = \frac{1}{1+e^{-net_i}}$$

$$\frac{\partial o^i}{\partial net_i} = o^i(1-o^i)$$

$\downarrow$

x



$$\frac{\partial}{\partial x}\left(x(1-x)\right) = 0 \Rightarrow \frac{d}{dx}\left(x - x^2\right) = 0$$

$x = o^i$

$$1 - 2x = 0$$

$$\therefore x = \frac{1}{2} = o^i$$

$$o^i = \frac{1}{1+e^{-net_i}} = \frac{1}{2}$$

$net_i = ?$

$$2 = 1 + e^{-net_i}$$

$$\therefore \; e^{-net_i} = 1 = e^0$$

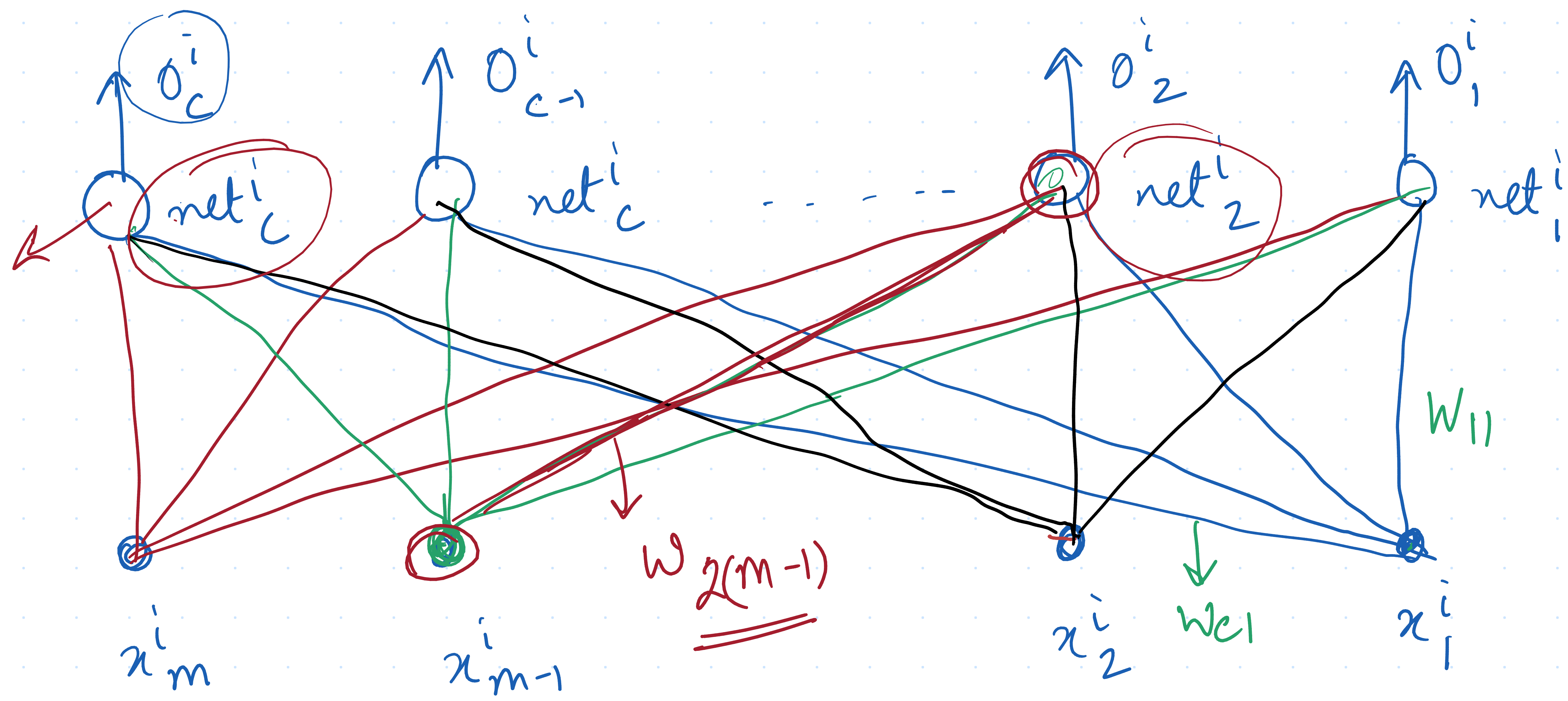
$$\therefore \; \boxed{net_i = 0}$$

Value of derivative of sigmoid which is maximum = ?

$$\boxed{o^i = \frac{1}{2}} \qquad o^i(1-o^i) \Rightarrow \frac{1}{2}\left(1-\frac{1}{2}\right) = \frac{1}{4}$$

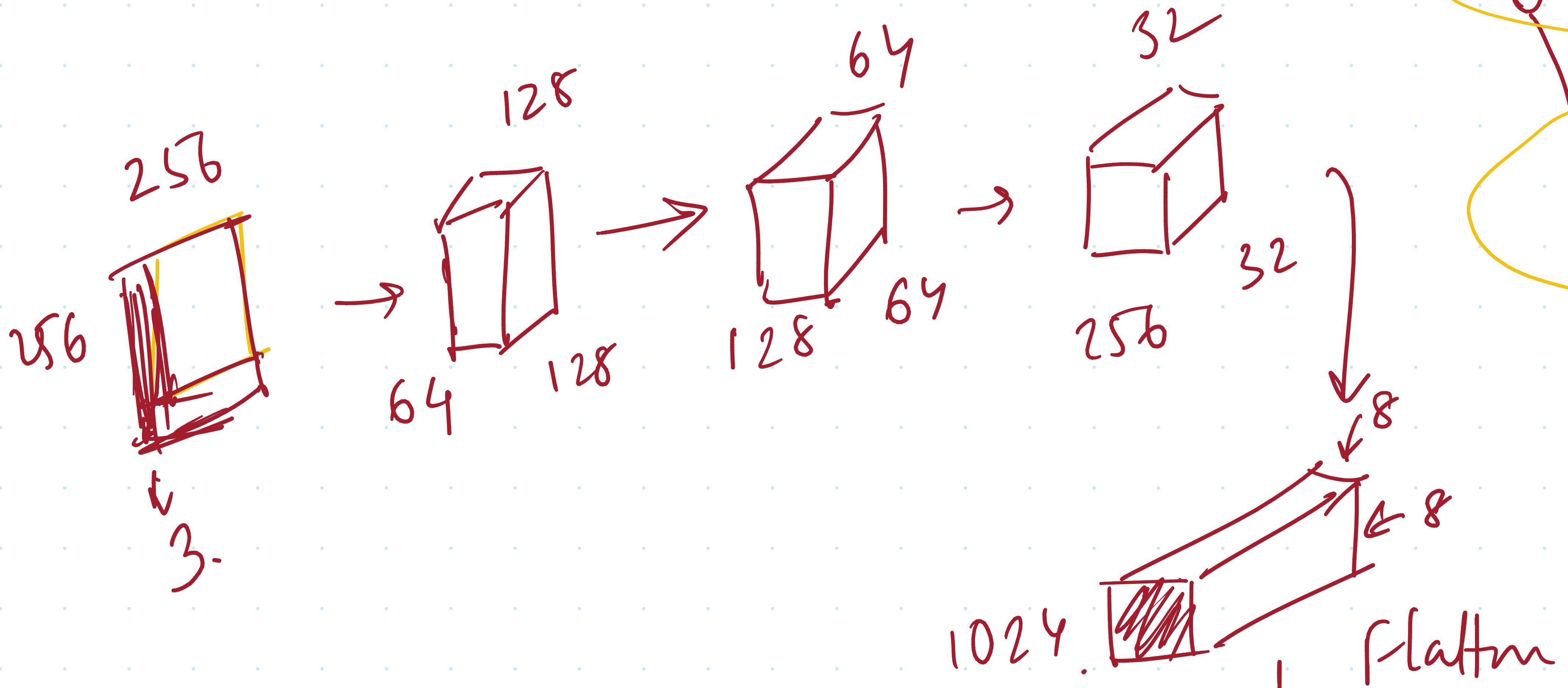
$$\frac{\partial o^i}{\partial net_i} = 0$$

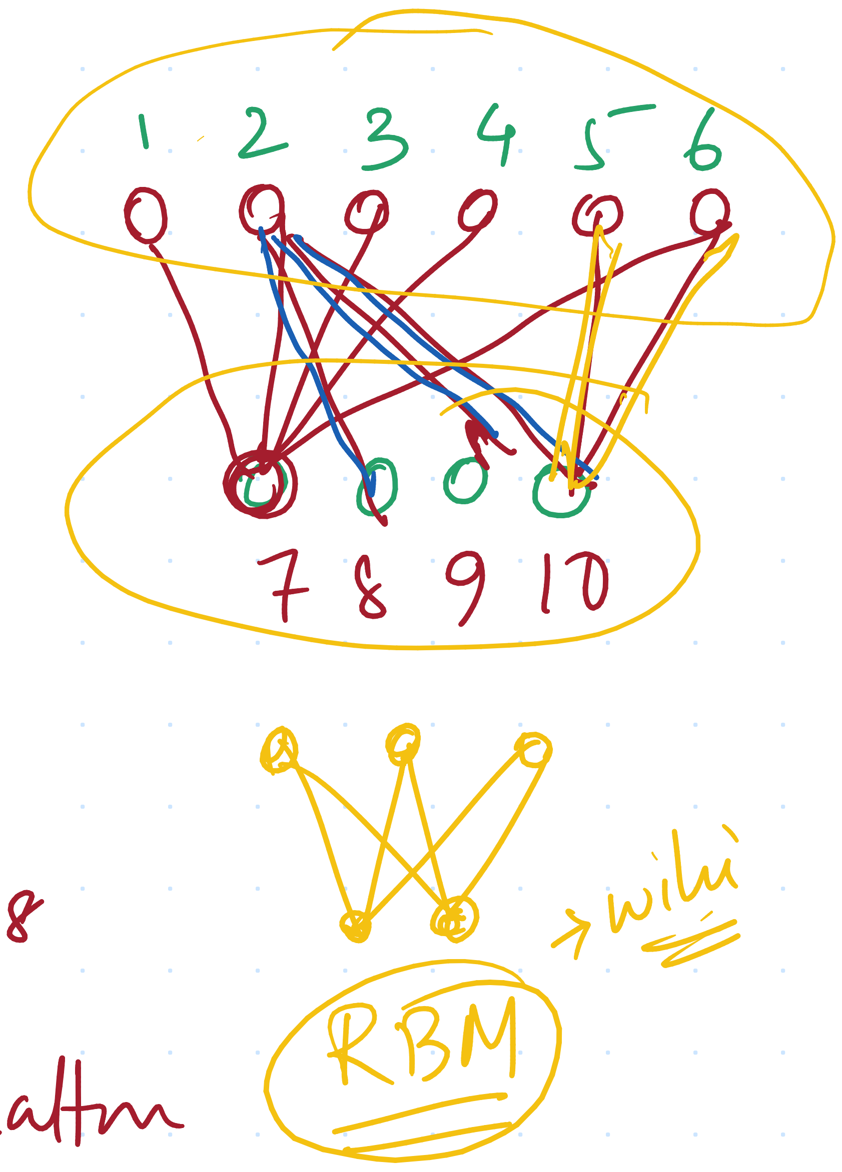
Maximum value of derivative of sigmoid is $\frac{1}{4}$.



$m > c$ # of neurons in input > # of predicted class.

output for class c, $\underline{c = 1 \text{ to } c}$





$m \leftarrow \underline{1024 \times 8 \times 8}$

$\downarrow$

$C \leftarrow$ 10 neurons

10-class classification

In general

$\underline{m >> C}$

Softmax

$$o_c^i = S(NET^i)_c = \frac{e^{net_c^i}}{\sum_{k=1}^{C} e^{net_k^i}}$$

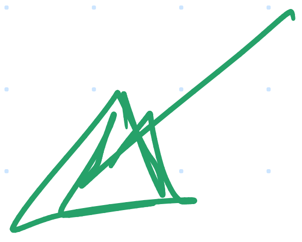
$$\begin{array}{c} 0 \\ 1 \\ 2 \end{array}\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} \xrightarrow{\text{Softmax}} \begin{bmatrix} 0.09 \\ 0.665 \\ 0.245 \end{bmatrix} \xrightarrow{\text{argmax}} 1$$

$$\left\{ \frac{e^{-1} \to 0.367}{\underline{e^{-1}+e^{1}+e^{0}}}, \frac{e^{1} \to 2.718}{\underline{e^{-1}+e^{1}+e^{0}}}, \frac{e^{0} \to 1}{\underline{e^{-1}+e^{1}+e^{0}}} \right\}$$

4.086.

* In Softmax, all the output ≥ 0

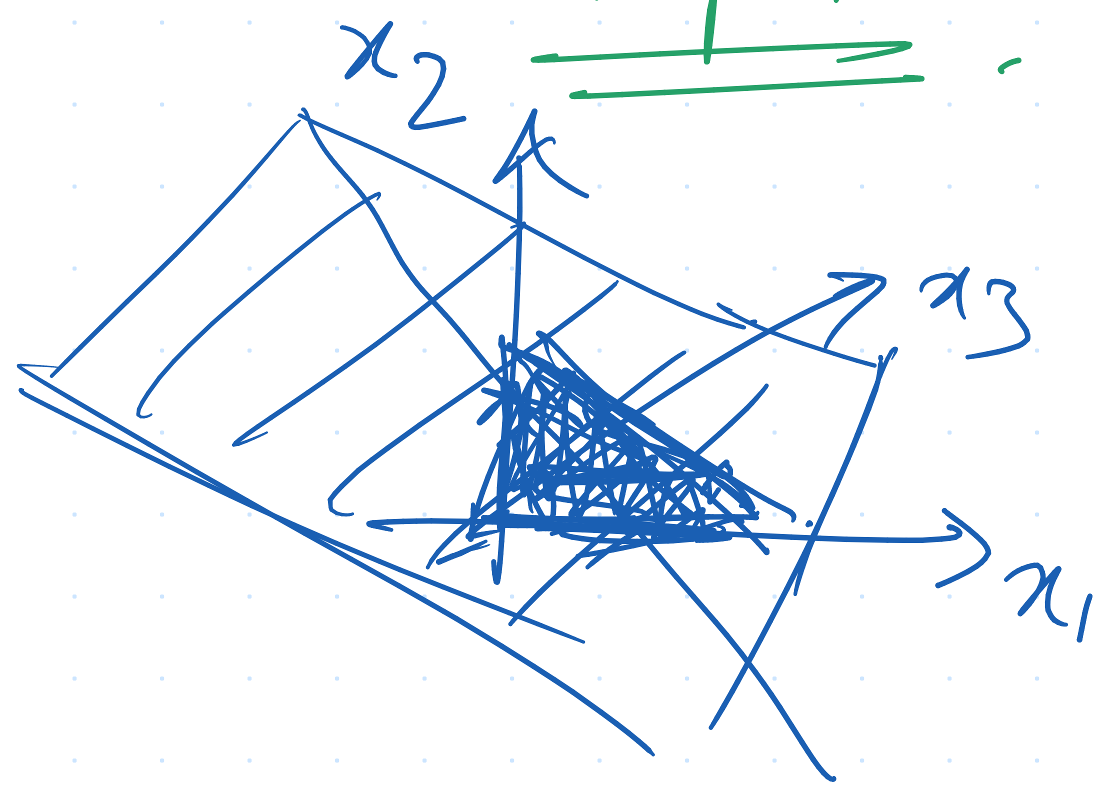
* The sum of all the elements in the softmax $= ? = 1$.



Lies in a $(k-1)$ - dimensional simplex.

$$x_1 + x_2 + x_3 = 1$$

$$\underbrace{x_1 + x_2}_{\text{2-dimensional simplex.}} = 1 - x_3$$



Notations :-

$i = 1, 2, 3, \dots N$ (N - i/p - o/p pairs).

i- runs over the training data. $\{x^i\}_1^N$

$$\mathcal{D} = \left\{ \begin{array}{l} [x_1^1, x_2^1, \dots, x_m^1] \rightarrow [O_1^1, O_2^1, \dots, O_c^1] \\ [x_1^2, x_2^2, \dots, x_m^2] \rightarrow [O_1^2, O_2^2, \dots, O_c^2] \\ \vdots \\ [x_p^N, x_2^N, \dots, x_m^N] \rightarrow [O_1^N, O_2^N, \dots, O_c^N] \end{array} \right\}$$

One-hot vector.

$\{1, 2, 3\} \rightarrow \left\{ \begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix} \right\}$ y

3-class.

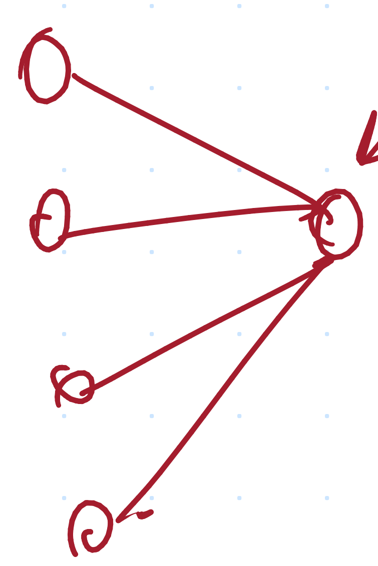
Cat, Dog, Elephant. → classification.

1 2 3.

MSE.

Categorical ?
Cross-Entropy Loss ?

classification problem.



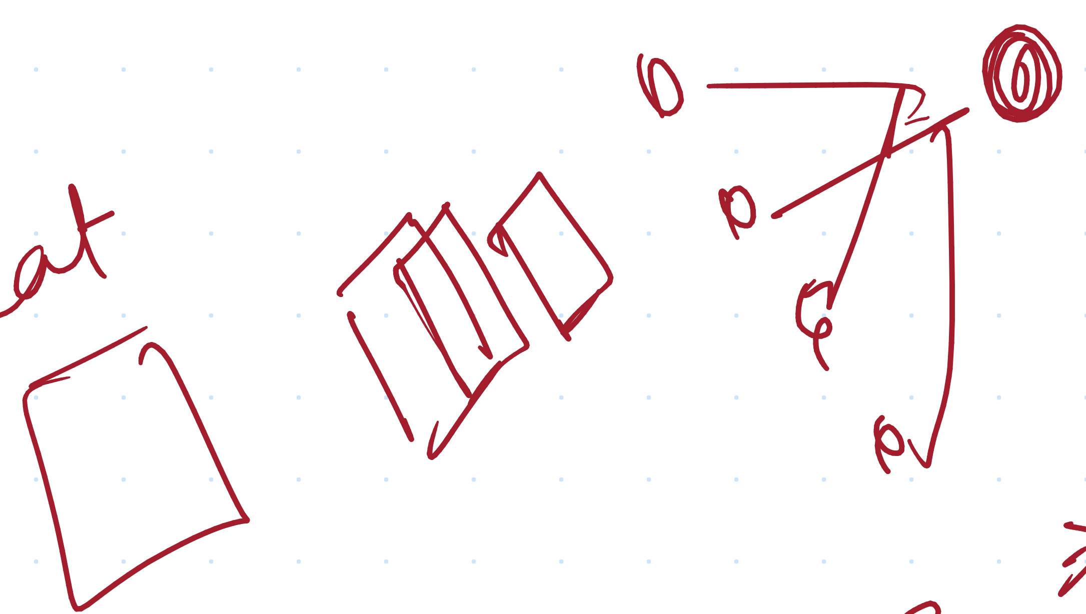
1, 2, 3. $\hat{y}$ y = 1,2,3.

L1, L2.

Regression problem.

(1 – 3)

1, 2, 3,.



sigmoid.

Cat

(0, 1) ×3 ⇒ (0, 3)

scale ⇒ 3.

(0, 3)

L2
L1 MSE GT.

1 ↔ 1

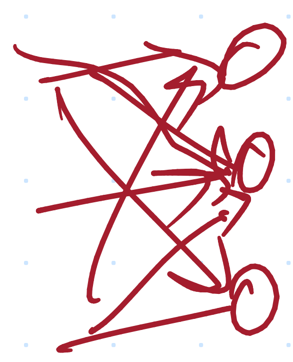
2 ↔ 2

3 ↔ 3.

[1.5] ⇒ ?.

1, 2, 3

argmax. ✓



$\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}$.

Cat-Cross Entropy?

1 2 3 4

MSE, L1

Cat 2 → 2, 3 ~ Dog ~ Pig ⟷ Elephant

Img

Cat	Dog	Pig	Elephant
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Kgs, Km.

1 Kg = ? Km

Dist

Kg Km.

x

2 Kg = 1 Km

Man.

100.

0

Regression

scalar.

A, B, C, D

scalar.

ordinal.

(M, F)

nominal

categorical

one hot.

classification

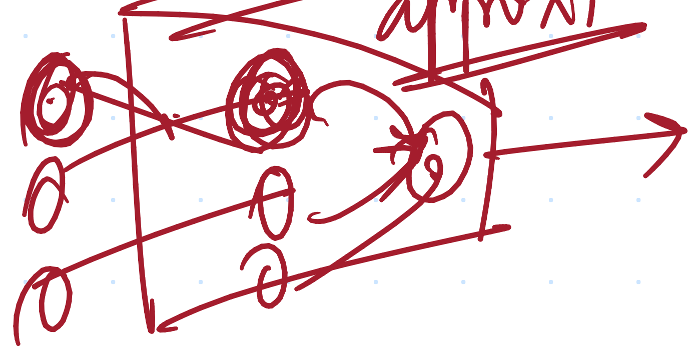
$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

(1, 2)

1.03

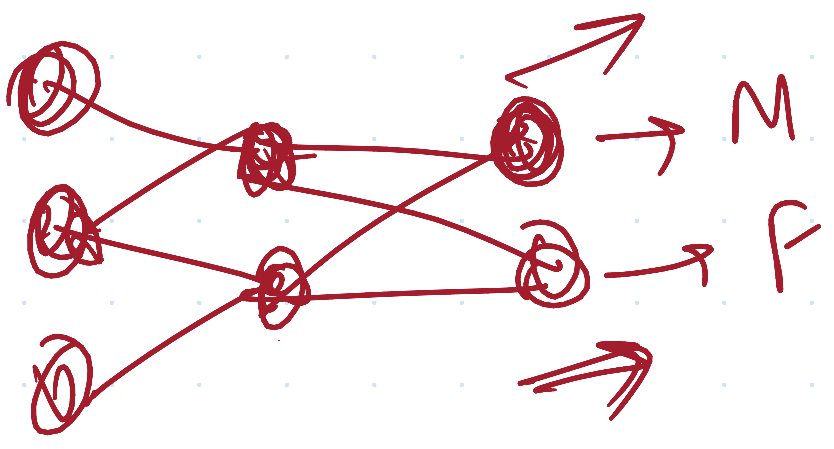
1.05

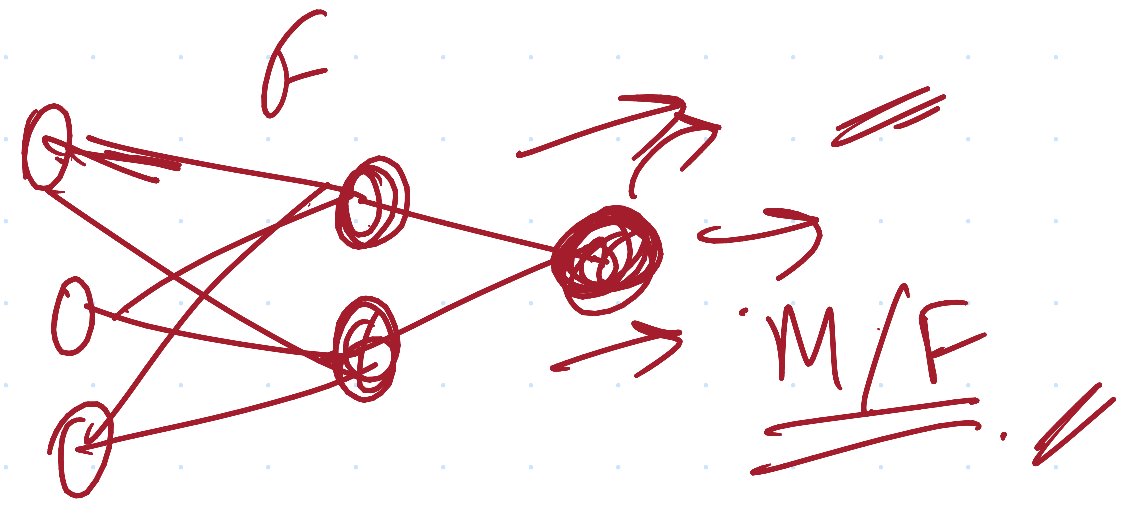
Functional approxi



L1

MSE.





Softmax

$$o^i_c = \frac{e^{net^i_c}}{\sum_{k=1}^{c} e^{net^i_k}}$$, ith input pattern

$$\ln o^i_c = net^i_c - \ln\left(\sum_{k=1}^{c} e^{net^i_k}\right)$$

Derivative w.r.t. net^i_c :-

$$\frac{1}{o^i_c} \cdot \frac{\partial o^i_c}{\partial net^i_c} = 1 - \frac{e^{net^i_c}}{\sum_{k=1}^{c} e^{net^i_k}}$$

$\langle e^{net^i_1} \to 0, e^{net^i_2} \to 0, \ldots e^{net^i_c}, \ldots e^{net^i_k} \to 0 \rangle$

$$\frac{1}{o^i_c} \frac{\partial o^i_c}{\partial net^i_c} = 1 - o^i_c$$

Case 1 :-

when class c for o & net are the same.

$$\frac{\partial o^i_c}{\partial net^i_c} = o^i_c(1-o^i_c)$$

Case 2: when class c' and net^i_c are different from c & 0

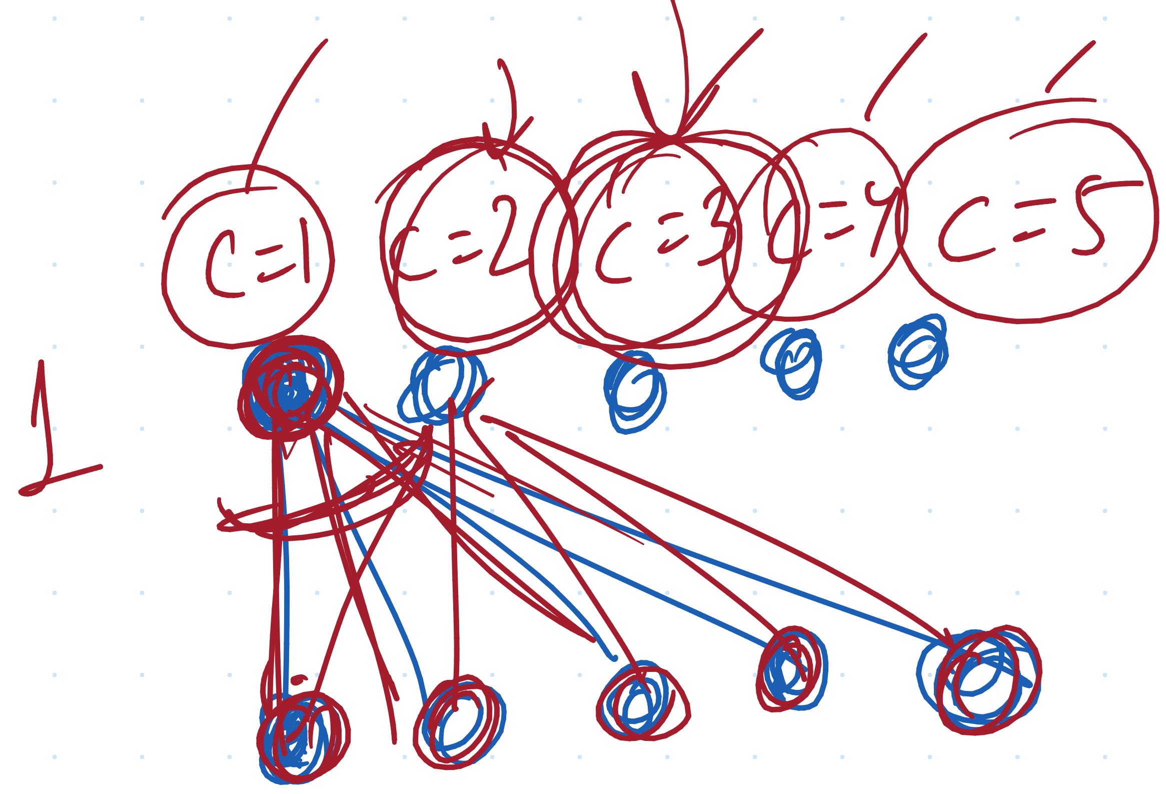
$$\underline{\ln o^i_c} = \underline{net^i_c} - \ln\left(\sum_{k=1}^{c} e^{net^i_k}\right)$$

Derivative w.r.t $net^i_{c'}$

$$\frac{1}{o^i_c}\,\frac{\partial o^i_c}{\partial net^i_{c'}} = 0 - \frac{net^i_{c'}}{\sum_{k=1}^{c} net^i_k} = -o^i_{c'}$$

$$\boxed{\frac{\partial o^i_c}{\partial net^i_{c'}} = -o^i_c\, o^i_{c'}}$$

when the classes are unequal.

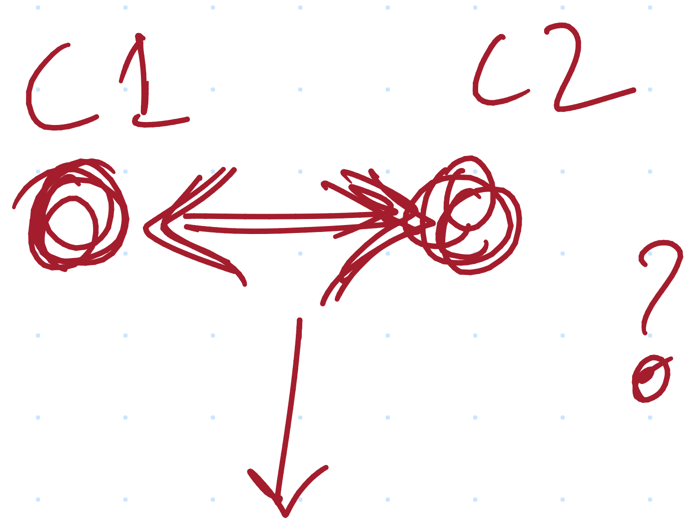


$O^i \rightarrow$ output vector

$x^i \rightarrow$ input vector.

$O^i_c \rightarrow$ component of $O^i \rightarrow$ probability of x^i belonging to the class c $(c = 1, 2, 3 \ldots C)$.

$C \rightarrow$ components.

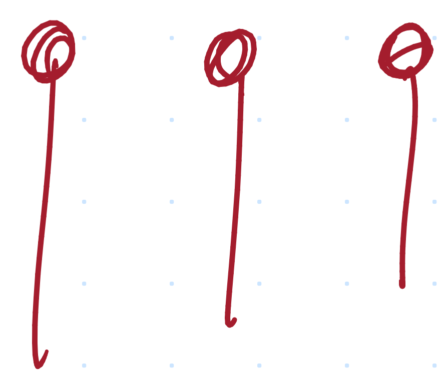


Softmax?

2-class classification problem.

$$p(C_1) + p(C_2) = 1$$

$$p(C_1) = 1 - p(C_2).$$



softmax.

When there are 2-classes the distribution is Bernoulli

$$\mathcal{L} = \prod_{i=1}^{N} (O^i)^{t^i} (1 - O^i)^{(1-t^i)} \qquad t^i = 1 \text{ or } 0.$$

$\rightarrow$ Bernoulli Dist.

$$\log \mathcal{L} = \log\left(\prod_{i=1}^{N} (o^i)^{t_i} (1-o^i)^{(1-t_i)}\right)$$

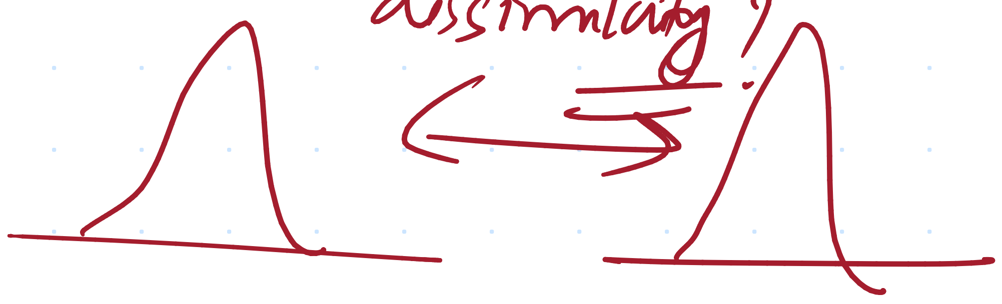
$$LL = \sum_{i=1}^{N} \{t^i \log(o^i) + (1-t_i) \log(1-o^i)\}$$

$$-\text{Log Likelihood} = \underline{\underline{-\sum_{i=1}^{N} \{t^i \log(o^i) + (1-t_i) \log(1-o^i)\}}}$$

$\downarrow$

Binary Cross Entropy Loss.

$$D_{KL}(p\|q) = \sum_{k=1}^{n} p_k \log\left(\frac{p_k}{q_k}\right)$$



Distance ⇒ ??

Symmetric?

$$p_k \log\left(\frac{p_k}{q_k}\right) \approx q_k \log\left(\frac{q_k}{p_k}\right)$$

$x_1 \longleftrightarrow x_2$

symmetric X

$$\underline{\underline{0.1}} \log\left(\frac{0.1}{0.6}\right) \Longleftrightarrow 0.6 \log\left(\frac{0.6}{0.1}\right)$$

$$= 10^{-1} \times \cancel{\log(6)} \Longleftrightarrow 0.6 \cancel{\log(6)}$$

$-1 \approx 6.$

$$D_{KL}(p \| q) = \sum_{k=1}^{n} p_k \log\left(\frac{p_k}{q_k}\right)$$

$$\boxed{D_{KL}(p \| q) = H(p,q) - H(p)}$$

Cross entropy

Entropy

?

$$D_{KL}(p \| q) = \underline{\sum_{k=1}^{n} p_k \log p_k} - \sum_{k=1}^{n} p_k \log q_k .$$

Information Gain ↓ Entropy ↑

State - highly unlikely → probability ↓ ↓

Information ↑

Entropy ↓ (?)

(4) Tomorrow there is a high chance of sunny weather, since, last 3-months have been sunny & extremely hot.

$P(S) = \uparrow \approx 0.9$ which

(↓) Lower information

High entropy (↑)

high prob. (↑)

(*) But what if I say that previous 3 months is extremely dry & hot & sunny but tomorrow there is a maybe high chance of rain.

$P(R) \uparrow$

Prob. low. $\leftarrow$ $P(S) \downarrow$

Entropy $\downarrow$

IG $\uparrow$

$P(S) \downarrow$

Weather Report $\uparrow$ ~ New information (*) (*)

Likelihood L of observations in case of Softmax :-

For N no. of i/p - o/p pairs :-

$$\mathcal{L} = \prod_{i=1}^{N} \prod_{k=1}^{C} (o_k^i)^{t_k^i} \quad ; \; t_k^i \to 1/0$$

$$\log \mathcal{L} = \log \prod_{i=1}^{N} \prod_{k=1}^{C} (o_k^i)^{t_k^i}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{C} t_k^i \log o_k^i$$

$$\boxed{-LL = -\sum_{i=1}^{N} \sum_{k=1}^{C} t_k^i \log o_k^i}$$

Categorical Cross Entropy Loss.