$\alpha(P_b, P_a)$

$MAE = \frac{|e|}{n}$

$MSE = \frac{e}{N}$

n.t	P
1	0.8, 0.9
1	→ 0.1 / 0.2

learn fast

Optimizing Parameters $\alpha / \frac{d J(\theta)}{d w}$

* Chose loss function wisely

Entropy

Randomness / Chaos (Measurement of these)

~~Information~~ Information

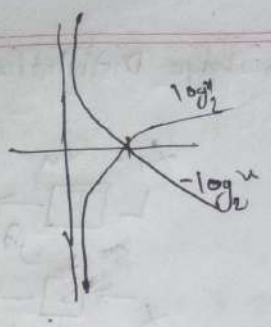
P(E) = 1 [where we knows the output]

→ least Entropy

P(E) = ? [we are not sure about output]

↳ Higher Entropy

		Bit length
A	0.3	$-\log_2^{0.3}$
B	0.6	$-\log_2^{0.6}$
C	0.05	$-\log_2^{0.05}$
D	0	$-\log_2^{0}$
F	0.05	$-\log_2^{0.05}$



(- to ~~avoid~~ get positive results)

Average bit length

Eg)

Bit length$_A$ = $(0.3 \times -\log_2^{0.3})$ + $(0.6 \times -\log_2^{0.6})$ + . . . + $(0.05 \times -\log_2^{0.05})$

Cake 10 Coke 3 Sandwich 20

$cost_{Avg} = \frac{10+3+20}{3} = 11$

take, cake, cake, coke, S, S, S

$= \frac{2\times10}{6} + \frac{3}{6} + \frac{3\times20}{6}$

$= \left(\frac{2}{6}\right)\times10 + \left(\frac{1}{6}\right)\times3 + \left(\frac{3}{6}\right)\times20$

$=$

((equivalent to

Average bit length $= -\sum_{i=1}^{n} p_i \log_2^{p_i}$

Entropy ⇒ Measurement of cost, Information, Randomness

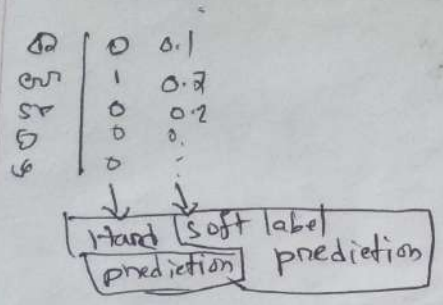
$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ → Entropy 0

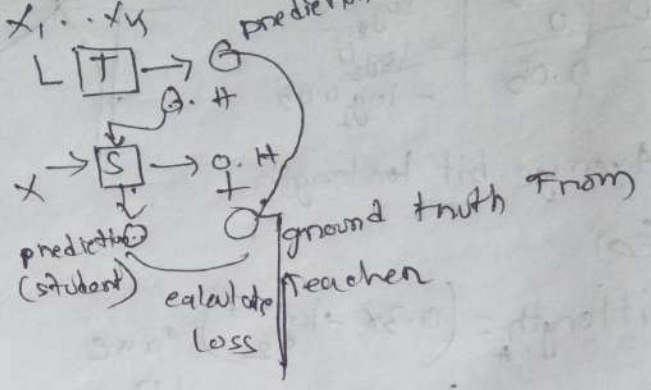
0.2
0.3
0.3
0.2

Entropy ?

So We developed cross-Entropy [what is the distance between these entropy, or Amount of Avg bit length]

Dark knowledge (knowledge Distillation)





$\alpha L_{0.H} + (1-\alpha) L_T$

$L_{0.H}$ → cross Entropy

L_T → KL Divergence

Dr. Sifat Momen (Sfm1)

Kacchi	0.1
Burger	0.3
Pizza	0.2
Fried chicken	0.4

$\Sigma = 1$

H(Sifat Bhai) = 1.85

Dr. Nabebel Mohammed (Nbm)

Kacchi	0.35
Burger	0.1
Pizza	0.1
Fried Chicken	0.3

$\Sigma = 1$

H(Nabeel) = 1.67 (slightly less Diversity)

[checks Avg bit length equation)

What is the avg code length if Dr. sibat uses my code book.

↳ 0.1*1 + 0.3*63.32 + 0.2*3.32 + 0.4*1.73 = 2.435

(cross Entropy)

Equation:

$$-\sum_{i=1}^{l} \underbrace{P_i^s}_{\text{sibl}} \underbrace{\log_2 P_i^N}_{\text{mabeel}} \quad \text{[cross Entropy]} \Rightarrow 6.2.45 \quad [\text{Extra } 0.6]$$

$$-\sum_{i=1}^{l} \underbrace{P_i^s}_{\text{sibat}} \underbrace{\log_2 P^s}_{\text{sibat}} \quad \text{[Not cross Entropy]} \Rightarrow 1.85$$

$$KL(S,N) = \underbrace{-\sum_{i=1}^{l} P_i^s \log P_i^N}_{\text{cross Entropy}} - \underbrace{\left(-\sum_{i=1}^{q} P_i^s \log P_i^s\right)}_{\text{Entropy}}$$

$KL(S,N)$ not neeeessary equals to $KL(N,S)$

$\hat{P}$

~~#~~ $x \rightarrow [m] \rightarrow s \rightarrow \text{softmax} \rightarrow \hat{P}$

$$\hat{P} = \begin{bmatrix} \hat{P}_{DN} \\ \hat{P}_e \\ \hat{P}_{TB} \\ \hat{P}_N \end{bmatrix} \begin{matrix} \hat{P}_1 \\ \hat{P}_2 \\ \hat{P}_3 \\ \hat{P}_4 \end{matrix} \qquad P = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{matrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{matrix}$$

PN - 1
e - 2
TB - 3
N - 4

$$KL(P,\hat{P}) = -\sum_{i=1}^{4} P_i \log \hat{P}_i - \left(-\sum_{i=1}^{4} P_i \log P_i\right)$$

$$\Downarrow$$

$$(-(0\log 0 + 0 + 1\log 1 + \dots)) = 0$$

(This will be eliminated)

Simplified (when Ground Truth is One Hot Encoding)

$$KL(P,\hat{P}) = -\sum_{i=1}^{4} P_i \log \hat{P}_i$$

$$= -1.\log \hat{P}_3$$

$$= -\log \hat{P}_3$$

log loss = Cross Entropy loss = logarithmic loss

= Negative log loss

$$\vec{x} \rightarrow \boxed{m} \rightarrow \vec{S} \rightarrow \text{softmax} \rightarrow \hat{P} \rightarrow L(P,\hat{P})$$

$$\vec{S} = \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix}$$

$$\text{Softmax}(\vec{S}) = \begin{bmatrix} \hat{P}_1 \\ \hat{P}_2 \\ \hat{P}_3 \\ \hat{P}_4 \end{bmatrix}$$

$$\sum_{i=1}^{4} \hat{P}_i = 1$$

$$0 \le \hat{P}_i \le 1$$

$$\Downarrow$$

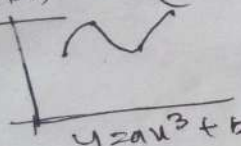
$$\hat{P}_i = \frac{e^{S_i}}{e^{S_1} + e^{S_2} + e^{S_3} + e^{S_4}} = \frac{e^{S_i}}{\sum_{j=1}^{4} e^{S_j}}$$

Alternatively

$$P_i = \frac{|s_i|}{\sum_{i=1}^{4} |s_j|}$$ [But often $-x, x$ both it will get same result $|-x| = x$, $|x| = x$]

$$\vec{P}_i = \frac{s_i^2}{\sum_{i=1}^{4} s_i^2}$$ [Same $x^2 = 4$, $(-x)^2 = 4$]

Regularization (It reduces complexity of model, prevents overfitting)



$L = (t, p) = \text{mae}(\ ) + R(\theta)\uparrow$ [If overfit, $R(\theta)\uparrow$ then $a \to 0, b \to 0$ i.e close to zero $y = ax^3 + bx + c$ becomes simple model

Also Dropout → does not same things (off some neuron)

Softmax function with Temperature:

$$P_i = \frac{e^{s_i}}{\sum_{j=1}^{4} e^{s_j}} \Rightarrow P_i = \frac{e^{(s_i/T)}}{\sum_{j=1}^{4} e^{(s_i/T)}}$$

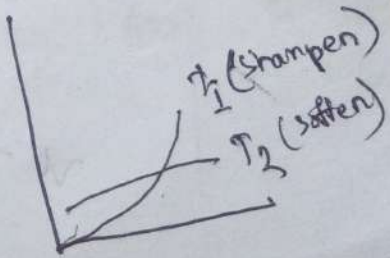
s_1	1
s_2	2
s_3	3
s_4	4

$e' = e$

	$T=1$	$T=2$
P_1	0.03	0.10
P_2	0.09	0.17
P_3	0.24	0.28
P_4	0.64	0.46

$\Sigma = 1$

For $T=1, 2$



x_1	x_2	x_3 ...	x_l
t_1	t_2	t_3 ...	t_l

$\downarrow OH(t_1)$ $\downarrow OH(t_2)$

$x \rightarrow \boxed{m} \rightarrow \vec{S} \rightarrow \text{softmax} \rightarrow \vec{P}$

$\hookrightarrow \mathcal{L} = -\log\left(P_{\text{argmax}}(OH(t_i))\right)$

$\text{arg max}\left(\begin{matrix}1\\2\\3\\5\end{matrix}\right) = 4$

$m_{Teacher} \doteq m$

x_1	x_2	x_3 ...	x_i

$\vec{S}_L = m_{teacher}(x_i)$

$\hookrightarrow$

x_1	x_2	x_3 ...

$\vec{S}_L = m_{teacher}(x_i)$

$\hookrightarrow$ Temperature controlled softmax

$\rightarrow$ Teacher

P_L

$x \rightarrow m_{student} \rightarrow \vec{S}^{student} \hookrightarrow \text{softmax}\left(\frac{\vec{S}^{student}}{\beta}\right)$

$\downarrow$

$\vec{P}^{student}$