in Machine learning ⇒

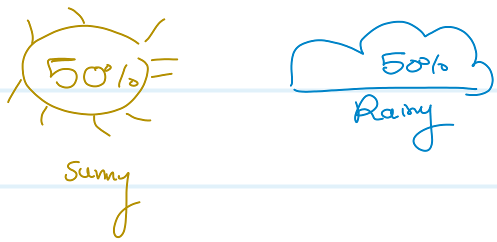
Cross-Entropy mostly used as cost function in traning the classifier

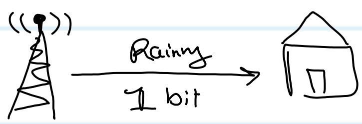
Come from Information theory (Claude Shannon) :

in theorem :

to transfer 1 bit of information = uncertainty divide by 2

eg, if weather





if send 1 bit it divide uncertainty by 2

so there's two options, now there's just one

so mean : it send you in 1 bit of useful info even if it send 8 bits

eg, what if there's 8 states of weather equally likely
(Sunny , Cloudy , Rainy ------)

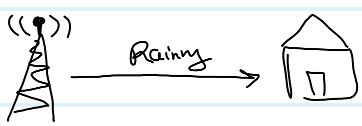
When station send you information ⇒ it divide your uncertainty by factor of $8 = 2^{?} = 2^3$

meaning ③ bits are the actual useful information

it's easy to calculate $\#\text{useful bits} = \log_2\left(\text{uncertainty Reduction Factor}\right)$

$$= \log_2(8)$$

eg, what if probability not equally likely

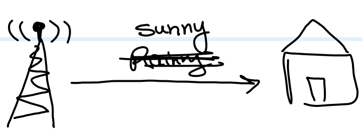
if 

$\Rightarrow$ this mean your uncertainty drop By factor of $\underline{\underline{4}} = 2^{②}$

Notes: uncertainty Reduction = inverse of probability of event

$$= \frac{1}{P}$$

So $\#\text{useful bits} = \log_2\left(\text{uncertainty Reduction}\right) = \log_2\left(\frac{1}{P}\right) = -\log_2(P)$

if 

$\text{uncertainty Reduction} = \frac{1}{0.75} = 1.33$

$\#\text{useful bits} = -\log(0.75) = 0.41$

Not so much useful $\simeq$ as I am already 75% sure that's Sunny Before you tell me

$\Rightarrow$ Note /#useful bits/ can be translated also as /#useful info./

what the average of useful information ?

↳ as there's 75% it's sunny & it'll send (0.41) useful info

↳ also there's 25% it's Rainy & it'll send (2) useful info.

$$\therefore \# \text{avg info} = 0.75(0.41) + 0.25(2) = 0.81 \text{ bits}$$

$$= -p_S \log(p_S) - p_R \log(p_R)$$

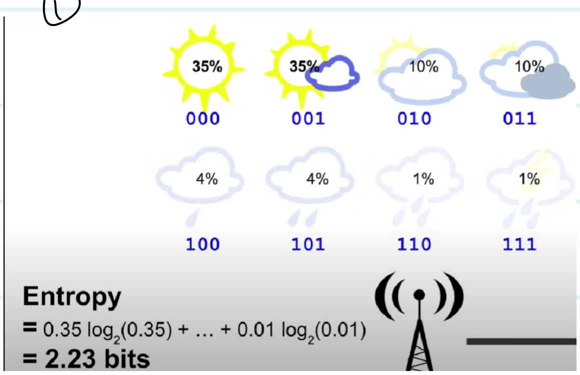
= Entropy ⇒ nice Measure of uncertain Events are

$$\text{Entropy} = H(\vec{p}) = -\sum p_i \log(p_i)$$ = how much on avg information you get → when you Sample event from distribution $\vec{p}$

= how unpredictable the probability distribution

Cross Entropy = avg message length (that actually sent)

①

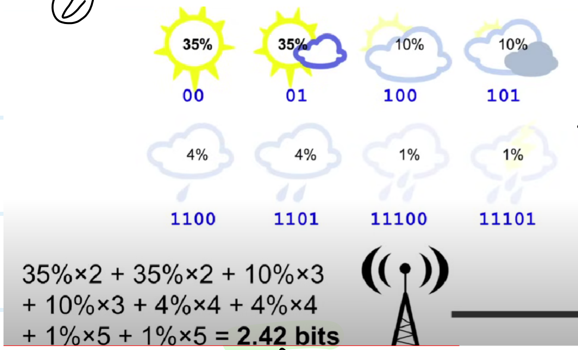


$$\text{Cross Entropy} = \underline{0.35} * 3 + \underline{0.35} * 3 + \underline{0.1} * 3 + \cdots$$

$$= 3 \text{ bits}$$

Other words: We sent 3 bits on avg

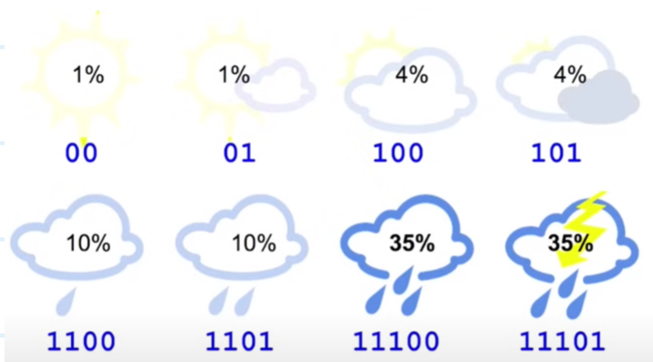
But also 2.23 on avg are useful

② 

← We can do Better By changing Encoding

Cross-Entropy

③ → What if we use Encoding in different location with different distribution



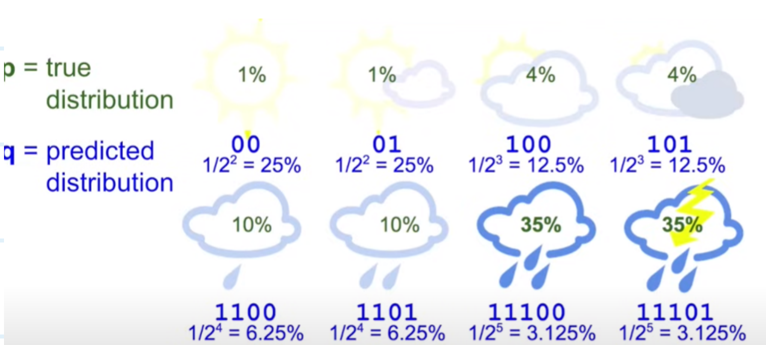
Cross-Entropy $= \underline{4.58}$ bits

Why? as we send ② bits for Sunny weather ≃ we predicted that uncertainty Reduction knowing it's Sunny $= 4$

≃ meaning we predicted 25% Sunny

Very Wrong assumption/predication

So our predicated probability $= \dfrac{1}{\text{uncertainty Red.}} = \dfrac{1}{2^{\#\text{bits}}} = q$



So Cross Entropy $= H(\vec{p}, \vec{q})$

$= -\sum_{i=1} p_i \log_2(q_i)$

Notes : - if Cross Entropy = Entropy $\Longrightarrow$ very efficient

- if Cross Entropy >>> Entropy $\Rightarrow$ Make very wrong Assumptions

- "Cross Entropy" - "Entropy" = Relative Entropy

= Kullback-leibler divergence

= KL Divergence

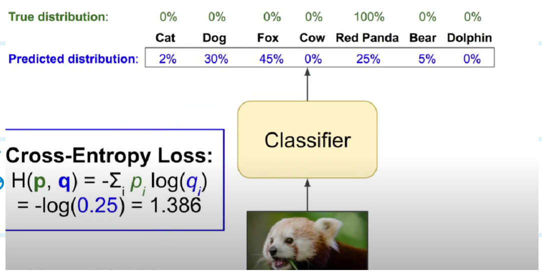
OR Better :

Cross Entropy = Entropy + KL Divergence

OR :

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

Cross Entropy as Cost Function :



natural log

Since it's One hot Encoding

then "Cross Entropy loss" = $-\log(0.25)$

if it predicated "1" $\Rightarrow$ loss = 0

if it predicated "0" $\Rightarrow$ loss = 1