Entropy

Where does the term entropy come from?

- The term "entropy" was initially discovered by German physicist and mathematician Rudolf Clausius and was used in the field of thermodynamics.
  - Clausius initially defined the entropy of a thermodynamic system as the ratio of the heat absorbed by the system to its absolute temperature when the system is in thermal equilibrium.

- The concept of entropy was later expanded to information theory by Claude Shannon, an American mathematician and electrical engineer, in 1948. Shannon introduced the term "entropy" to quantify the amount of uncertainty or randomness in a message or signal.
  - Shannon was also known as the ‘father of information theory’ as he had invented the field of information theory with the paper of *“A Mathematical Theory of Communication”*.

A Mathematical Theory of Communication By Father of Information Theory

- Claude Shannon demonstrated that all communication systems fundamentally share the same structure for transmitting and receiving information. He showed how messages can be encoded in bits to compress them and transmit them with near-perfect accuracy.

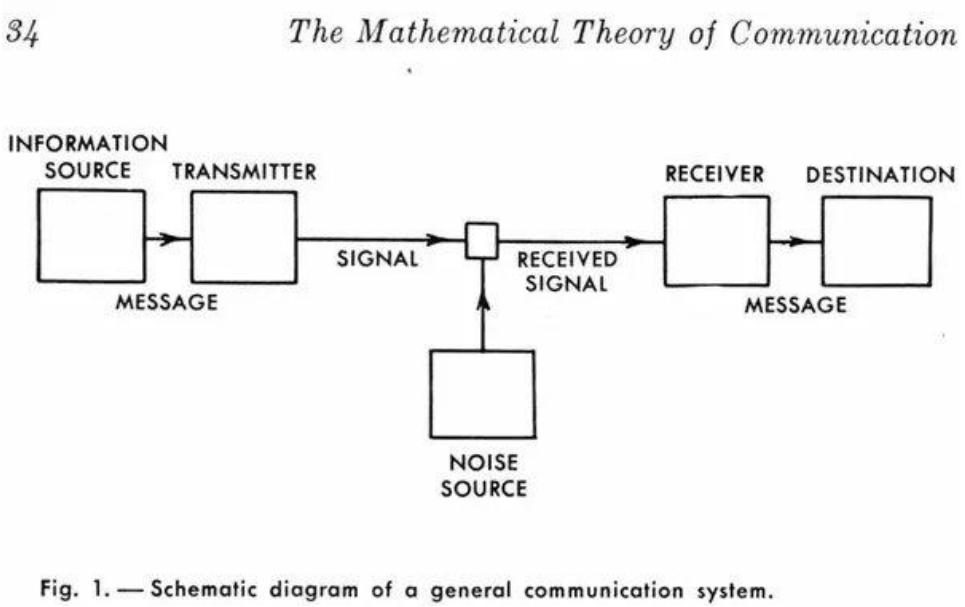
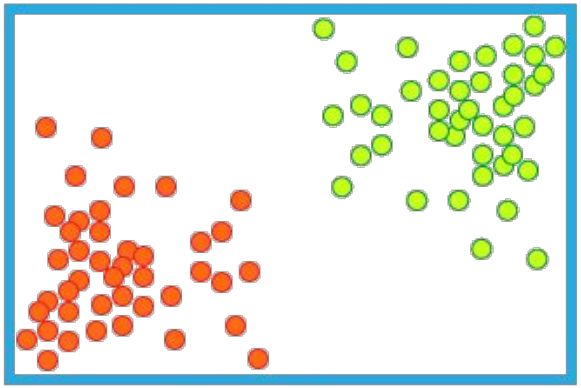


Fig. 1. — Schematic diagram of a general communication system.

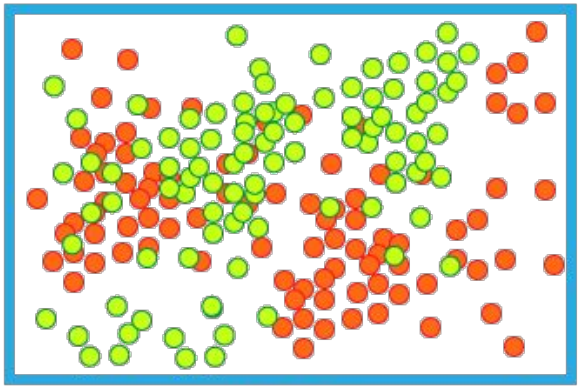
- In his article, he embarked on measuring the statistical nature of "lost information" in telephone line signals. The study aimed to address the problem of how to encode the desired information to be transmitted by a sender most effectively. For this purpose, the concept of information entropy was developed as a means to estimate the information content in a message, which is a measure of the reduced uncertainty caused by the message.

What is Information ?

- In Shannon's terms, **information entropy** is the degree of uncertainty, ie, a mathematical concept that **measures the amount of uncertainty** or randomness in a set of possible outcomes.
- In Shannon's terms, information measures surprise, or the amount of uncertainty we overcome. As a basic illustration of this point, he asked us to think about tossing a coin. A fair coin carries equal odds of landing on either side. As Shannon put it, such a coin stores one bit of information.A coin with heads on both sides carries no information.
  - An event with probability 100% is perfectly unsurpring.



Low Entropy



High Entropy

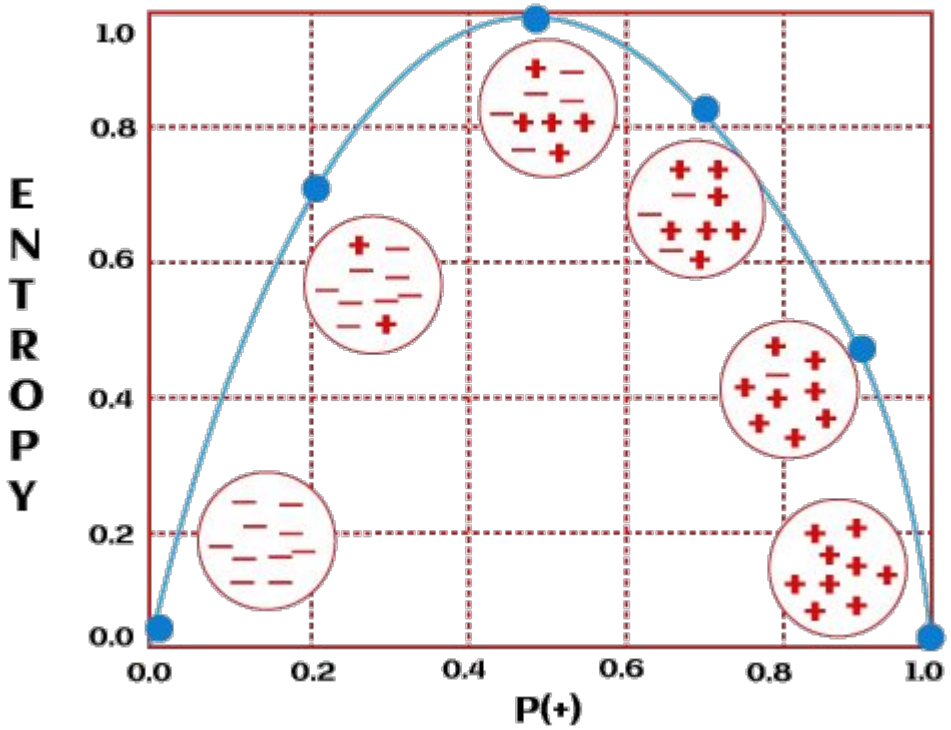
Shannon Entropy

- The entropy formula for an event *X* with *n* possible outcomes and probabilities *p_1, ...,p_n*:

$$H(X) = H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

$$\begin{aligned} \mathrm{H}(X) &= -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \\ &= -\sum_{i=1}^{2} \frac{1}{2} \log_2 \frac{1}{2} \\ &= -\sum_{i=1}^{2} \frac{1}{2} \cdot (-1) = 1 \end{aligned}$$

Shannon Entropy

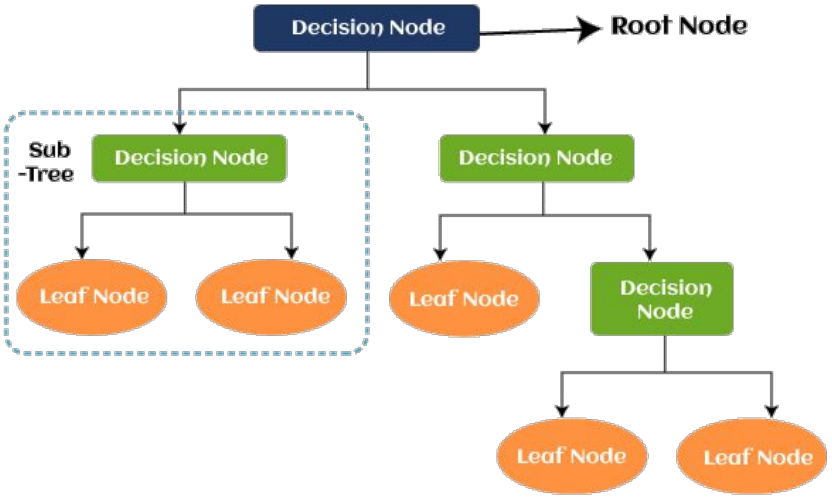


- When entropy becomes 0, then the dataset has no impurity. Datasets with 0 impurities are not useful for learning. Further, if the entropy is 1, then this kind of dataset is good for learning.

entropy is the machine learning metric that measures the unpredictability

Use Cases of Entropy in ML

- Decision trees are used to predict an outcome based on historical data. The decision tree works on the sequence of 'if-then-else' statements and a root which is our initial problem to solve.
- Entropy is used in decision trees as a measure of impurity or **disorder in a set of data**. The goal of a decision tree is to divide a dataset into subsets that are as pure (or homogeneous) as possible, and entropy is a useful tool for determining the **degree of impurity** in a set of data. When **building a decision tree, the algorithm will use entropy to calculate the potential information gain from each split, and will choose the split that results in the greatest reduction in entropy**. This process is repeated recursively until the tree is built.



Example of How Entropy is Used in Decision Tree Construction

1) The first step is to calculate the entropy of the target variable "Playing Golf"

Play Golf(14)	
Yes	No
9	5

So, the entropy of the dataset is: E(PlayGolf) = -(9/14)log2(9/14) - (5/14)log2(5/14) = 0.94

2) For the other attributes, we need to calculate the entropy after each of the split. E(PlayGolf, Outloook), E(PlayGolf, Temperature), E(PlayGolf, Humidity), E(PlayGolf,Windy)

		PlayGolf(14)		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5

E(Sunny) = -(3/5)log2(3/5) - (2/5)log2(2/5) = 0.971

E(Overcast) = -(4/4)log2(4/4) - (0/4)log2(0/4) = 0

E(Rainy) = -(2/5)log2(2/5) - (3/5)log2(3/5) = 0.971

$$E(PlayGolf, Outlook) = P(Sunny)E(Sunny) + P(Overcast)E(Overcast) + P(Rainy)E(Rainy)$$

$$E(PlayGolf, Outlook) = \frac{5}{14}E(3,2) + \frac{4}{14}E(4,0) + \frac{5}{14}E(2,3)$$

$$= \frac{5}{14}0.971 + \frac{4}{14}0.0 + \frac{5}{14}0.971$$

$$= 0.357 * 0.971 + 0.0 + 0.357 * 0.971$$

$$= \mathbf{0.693}$$

Example of How Entropy is Used in Decision Tree Construction

3) 3rd step is to calculate the entropy for other variable called Temperature

E(PlayGolf, Temperature) = P(Hot) E(2,2) + P(Cold) E(3,1) + P(Mild) E(4,2)

		PlayGolf(14)		
		Yes	No	
Temperature	Hot	2	2	4
	Cold	3	1	4
	Mild	4	2	6

$$E(\text{PlayGolf, Temperature}) = \mathbf{4/14} * E(\text{Hot}) + \mathbf{4/14} * E(\text{Cold}) + \mathbf{6/14} * E(\text{Mild})$$

$$E(\text{PlayGolf, Temperature}) = \mathbf{4/14} * E(2, 2) + \mathbf{4/14} * E(3, 1) + \mathbf{6/14} * E(4, 2)$$

$$E(\text{PlayGolf, Temperature}) = 4/14 * -(2/4 \log 2/4) - (2/4 \log 2/4) + 4/14 * -(3/4 \log 3/4) - (1/4 \log 1/4) + 6/14 * -(4/6 \log 4/6) - (2/6 \log 2/6)$$

$$E(\text{PlayGolf, Temperature}) = 5/14 * 1.0 + 4/14 * 1.811 + 5/14 * 0.918$$

$$= \mathbf{0.911}$$

Example of How Entropy is Used in Decision Tree Construction

4) Lets say we have the entropies for all the four attributes, let's go ahead to summarize them as shown in below:

E(PlayGolf, Outloook) = 0.693

E(PlayGolf, Temperature) = 0.911

E(PlayGolf, Humidity) = 0.788

E(PlayGolf,Windy) = 0.892

5) The next step is to calculate the information gain for each of the attributes. The attribute with the l**argest information gain is used for the split.** *Gain(S,T) = Entropy(S) – Entropy(S,T)*

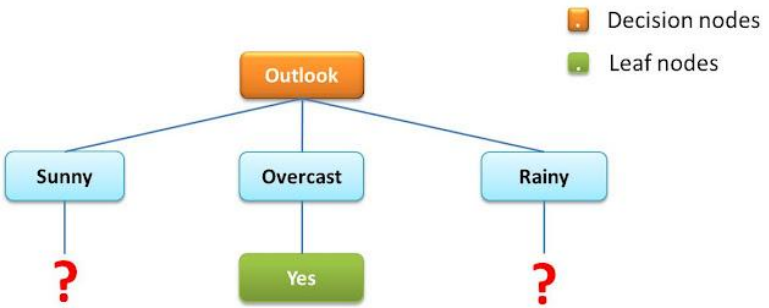
***Gain(PlayGolf, Outlook) = Entropy(PlayGolf) – Entropy(PlayGolf, Outlook)*= 0.94 – 0.693 = 0.247**

Gain(PlayGolf, Temperature) = Entropy(PlayGolf) – Entropy(PlayGolf, Temparature)= 0.94 – 0.911 = 0.029

Gain(PlayGolf, Humidity) = Entropy(PlayGolf) – Entropy(PlayGolf, Humidity)= 0.94 – 0.788 = 0.152

Gain(PlayGolf, Windy) = Entropy(PlayGolf) – Entropy(PlayGolf, Windy)= 0.94 – 0.892 = 0.048

Example of How Entropy is Used in Decision Tree Construction



Decision Tree after first split

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No

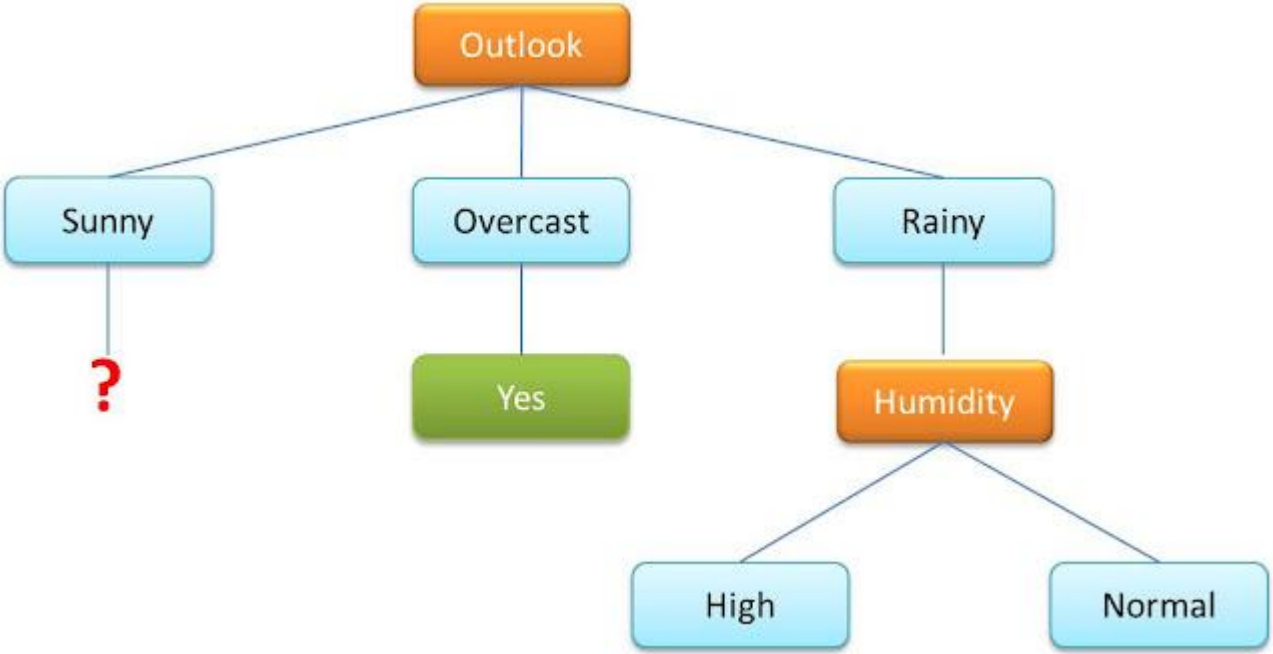
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

The Rainy attribute could be split using High and Normal attributes and that would give us the tree below.

Outlook	Temperature	Humidity	Windy	Play Golf
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

Initial Split using Outlook : The Sunny and the Rainy attributes needs to be split. The Rainy outlook can be split using either Temperature, Humidity or Windy.

Example of How Entropy is Used in Decision Tree Construction



Cross Entropy Loss (Log Loss, Logistic Loss)

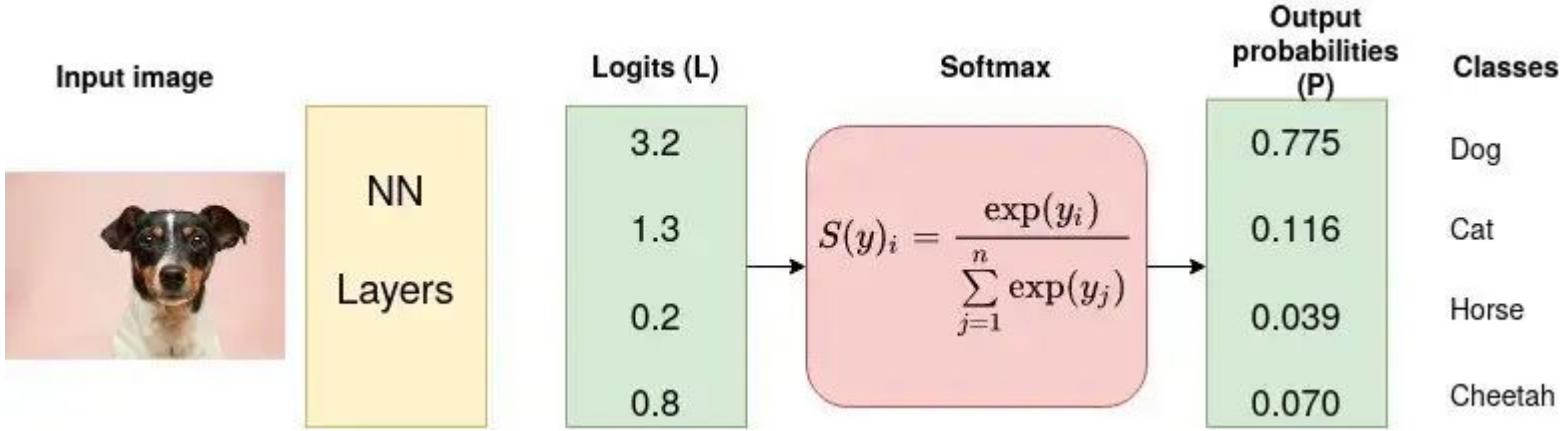
- Entropy is the number of bits required to transmit a randomly selected event from a probability distribution.
- **The cross entropy is the average number of bits needed to encode data coming from a source with distribution p when we use model q.**
- Cross-entropy is is used to measure the **dissimilarity** between the predicted probability distribution and the true probability distribution.In simple terms, cross-entropy loss is a measure of the difference between the predicted class probabilities and the true class probabilities. The goal of the training process is to minimize the cross-entropy loss, which means to make the predicted class probabilities as close as possible to the true class probabilities.
- **H(P, Q): Where H() is the cross-entropy function, P may be the target distribution and Q is the approximation of the target distribution.**Cross-entropy can be calculated using the probabilities of the events from P and Q, as follows:H(P, Q) = — sum x in X P(x) * log(Q(x)

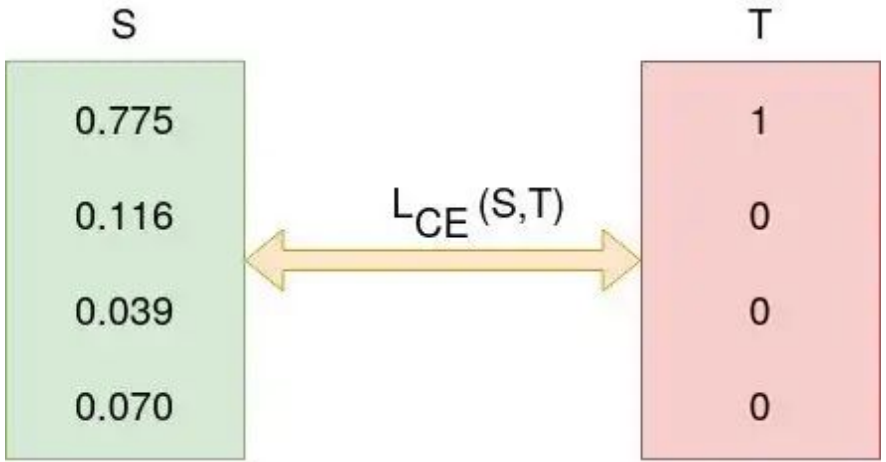
The definition may be formulated using the Kullback–Leibler divergence $D_{\mathrm{KL}}(p \parallel q)$, divergence of p from q (also known as the *relative entropy* of p with respect to q).

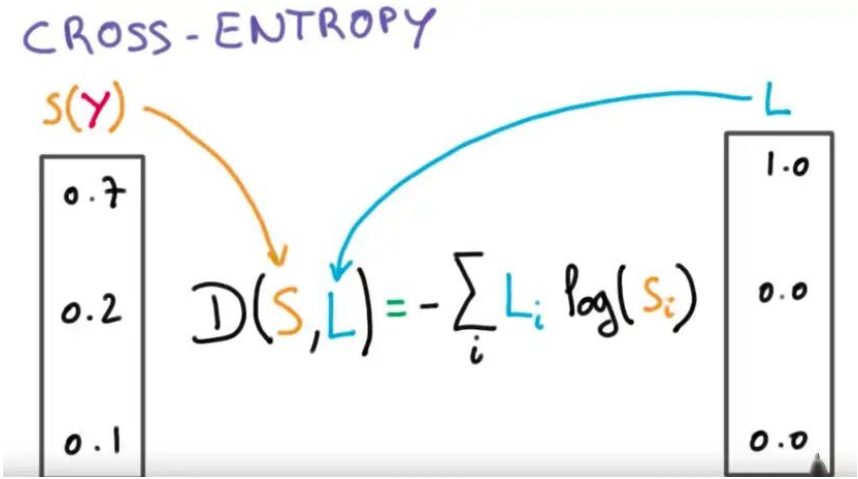
$$H(p, q) = H(p) + D_{\mathrm{KL}}(p \parallel q),$$

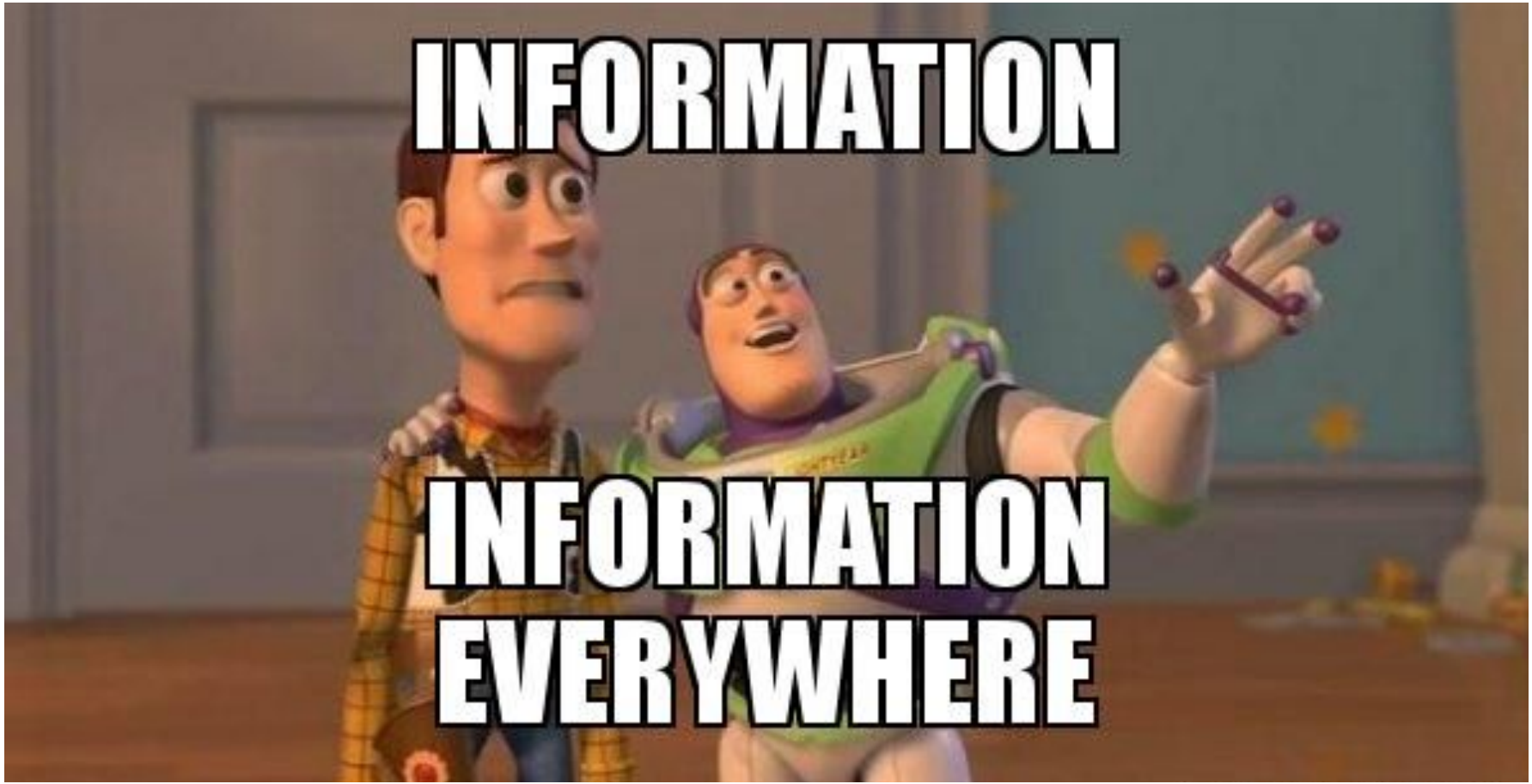
where $H(p)$ is the entropy of p.

$$KL(P||Q) = \sum p_i(x) log(\frac{p_i(x)}{q_i(x)})$$


Input image
NN
Layers
Logits (L)
3.2
1.3
0.2
0.8
Softmax
$S(y)_i = \frac{\exp(y_i)}{\sum_{j=1}^{n} \exp(y_j)}$
Output probabilities (P)
0.775
0.116
0.039
0.070
Classes
Dog
Cat
Horse
Cheetah


S
0.775
0.116
0.039
0.070
L_{CE} (S,T)
T
1
0
0
0


CROSS-ENTROPY
S(Y)
L
0.7
0.2
0.1
$D(S,L) = -\sum_i L_i \log(S_i)$
1.0
0.0
0.0


INFORMATION
INFORMATION
EVERYWHERE

REFERENCES

https://medium.com/unpackai/cross-entropy-loss-in-ml-d9f22fc11fe0#:~:text=Cross%2Dentropy%20can%20be%20calculated,*%20log(Q(x))

https://towardsdatascience.com/entropy-is-a-measure-of-uncertainty-e2c000301c2c

https://www.quora.com/How-would-you-explain-Shannons-information-theory-in-laymans-terms

https://en.wikipedia.org/wiki/Cross_entropy

https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

https://kindsonthegenius.com/blog/how-to-build-a-decision-tree-for-classification-step-by-step-procedure-using-entropy-and-gain/