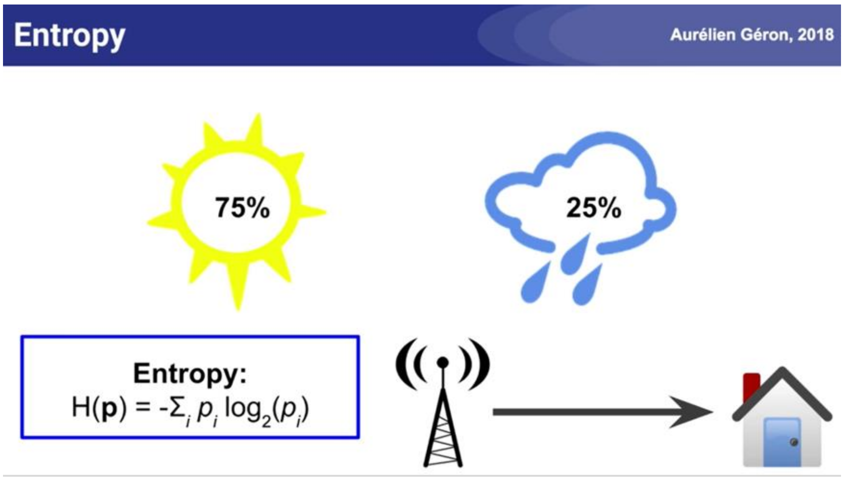
Entropy, Cross-Entropy and KL-Divergence

These concepts come from Claude Shannon's Information Theory. Information Theory's goal is to reliably and efficiently transmit a message consists of bits from a sender to a recipient.

Entropy

Entropy is measuring information when there are uncertain events. It measures the average information you get from each event by considering their probabilities. In a more meaningful way, it is the average amount of information that you get from one sample drawn from a given probability distribution p. It tells you how unpredictable the probability distribution is.



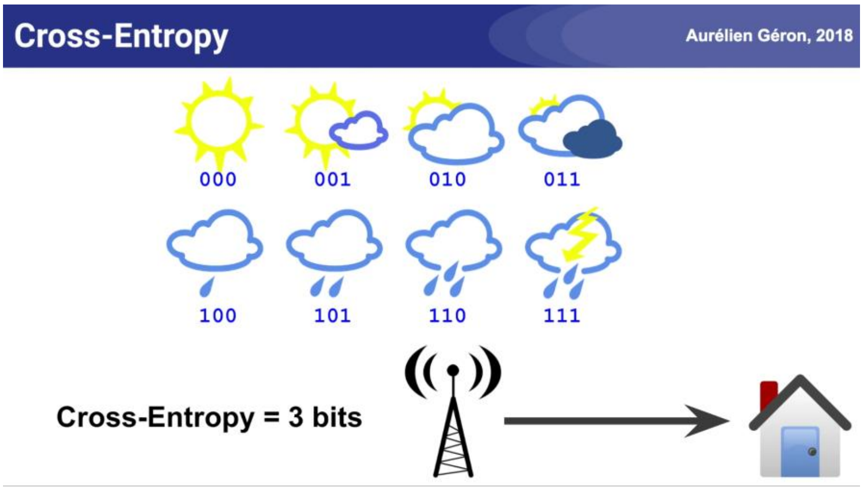
Entropy

Aurélien Géron, 2018

Entropy:
$H(\mathbf{p}) = -\Sigma_i\, p_i \log_2(p_i)$

Cross-Entropy

Cross entropy is basically the average message length. For example, if the weather station encodes each of the 8 possible options using a 3-bit code like the picture below then every message will have 3 bits, so the average message length is 3 bits and that's the cross entropy.

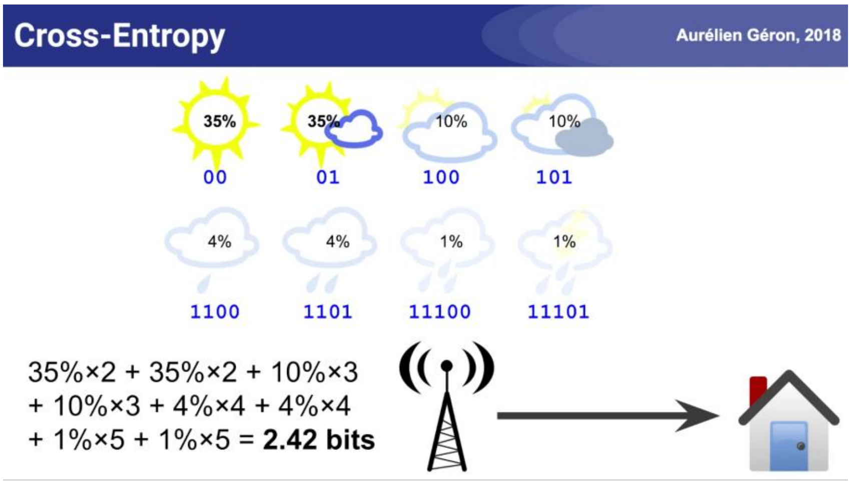


Cross-Entropy

Aurélien Géron, 2018

Cross-Entropy = 3 bits

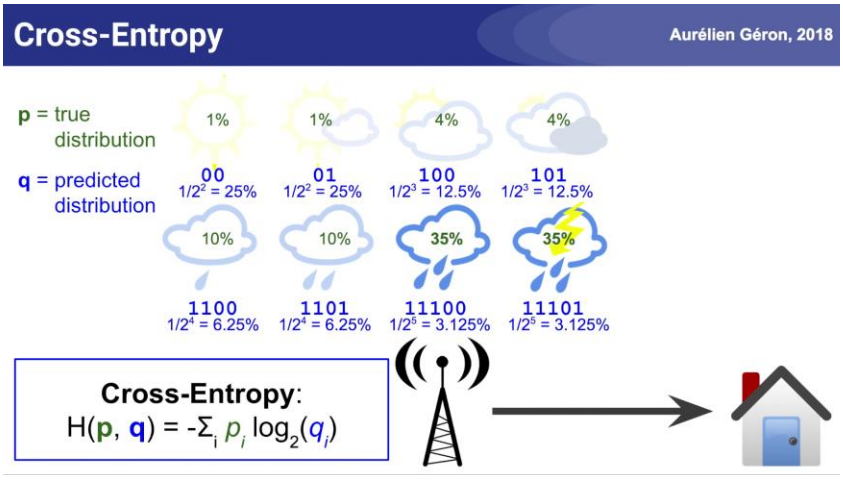
But when the length of messages differs and there is a probability involved, the formula changes.



Cross-Entropy

Aurélien Géron, 2018

35%×2 + 35%×2 + 10%×3 + 10%×3 + 4%×4 + 4%×4 + 1%×5 + 1%×5 = **2.42 bits**



Cross-Entropy

Aurélien Géron, 2018

p = true distribution

q = predicted distribution

Cross-Entropy:
$H(\mathbf{p}, \mathbf{q}) = -\Sigma_i \, p_i \log_2(q_i)$

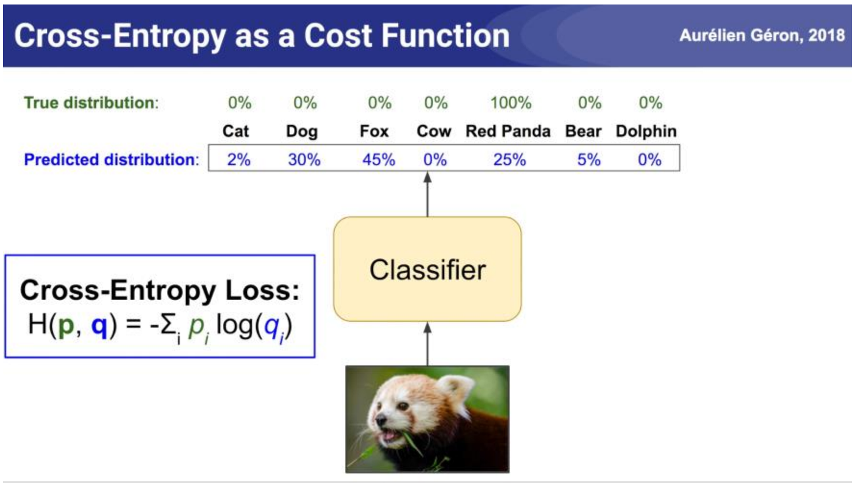
KL Divergence

If our predictions are perfect that is the predicted distribution is equal to the true distribution, then the cross entropy is simply equal to the entropy. But if the distributions differ, then the cross entropy will be greater than the entropy by some number of bits. This amount by which the cross-entropy exceeds the entropy is called relative entropy, or Kullback-Leibler Divergence (KL Divergence).

Cross-entropy = Entropy + KL Divergence

Cross Entropy in Machine Learning

For example, we classify the animals. It is a supervised learning, so we have a true distribution and prediction distribution. We can use the cross-entropy between these two distributions as a cost function. This is called the cross-entropy loss or log-loss. It uses natural logarithm rather than the binary logarithm.



Cross-Entropy as a Cost Function

Aurélien Géron, 2018

	Cat	Dog	Fox	Cow	Red Panda	Bear	Dolphin
True distribution:	0%	0%	0%	0%	100%	0%	0%
Predicted distribution:	2%	30%	45%	0%	25%	5%	0%

Cross-Entropy Loss:
$H(\mathbf{p}, \mathbf{q}) = -\Sigma_i\, p_i \log(q_i)$

References:

https://www.youtube.com/watch?v=ErfnhcEV1O8

Yağmur Uzun