Entropy and Cross Entropy

Dev Singh

May 1, 2024

Overview

1. **Entropy**

2. **Cross-Entropy**

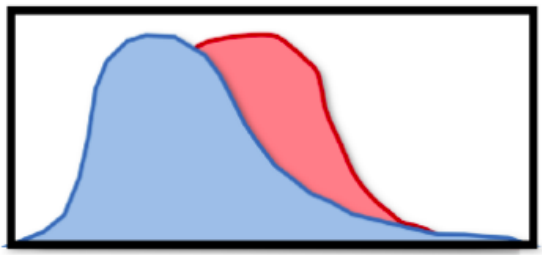
What is Entropy?

- Has roots in physics
- Fundamentally a measure of randomness.
  - Measures "amount of information" present in a variable.
- Entropy $\propto \frac{1}{\text{Information Gain}}$.
- In ML, represents the amount of uncertainty about the meaning of the data.
- Key measurement in ML and is used to optimize networks.
  - Used in the training of decision trees to make optimized decisions and increase accuracy.
  - Computer Vision - used to measure and optimize the representations that the model learns compared to the data present in the original image.

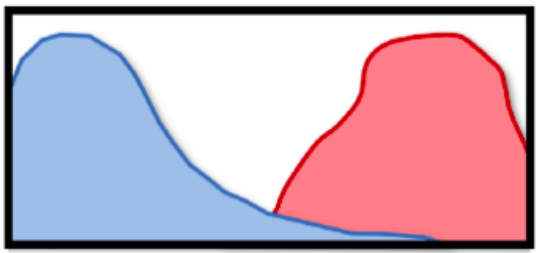
Formal Definition of Entropy

- Defined as the number of bits needed to encode data from source distribution p with model distribution q.
- An event x contains more information if it is more "surprising".
- Information $h(x)$ can be calculated for x, given the probability of the event $P(x)$ as $h(x) = -\log(P(x))$
- Entropy $h(x)$ can be calculated for discrete states x in set X with their probability $P(x)$ as $H(x) = -\sum_{i=0}^{|X|}(P(x_i) \cdot \log_2(P(x_i)))$

Visual Example of Entropy



Low information gain
High entropy



High information gain
Low entropy

What is Cross-Entropy?

- Builds upon entropy and compares the number of bits required to represent some information x with the source distribution p and model distribution q.
- In ML, considered between the dataset distribution and the model's weight distribution.
- Cross Entropy is calculated using the probabilities of events x in set X from distributions P and Q with function $H(x)$ as $H(P, Q) = -\sum_{i=0}^{|X|} P(x_i) \cdot \log_2(Q(x_i))$.

Applying Cross-Entropy in ML

- Kullback-Leiber Divergence (KLD) or Binary Cross Entropy (BCE) can also be used instead of Cross-Entropy, even though they both measure similar quantities.
  - This is due to the mini-batch nature of training ML models - entropy global truth and minibatch entropy can depart to make the metric unusable.
- Very often used to compare the probability distributions of classification outputs.
  - Classification outputs are probabilities that each category is the correct answer.
  - Using BCE or KLD, this probability distribution can be compared to the certain output from the ground truth in the dataset.
  - In this case, P is the category given in the dataset, and Q is the distribution output by the model with $H(P, Q)$ determining the loss.
- Can also be used in unsupervised tasks.
  - When the goal of the model is to create accurate representations of the dataset, compare the distribution of the model P to the per-pixel distribution of the dataset Q with $H(P, Q)$.

The End