Entropy and KL Divergence

From https://www.youtube.com/watch?v=KHVR587oW8I

Entropy

- Entropy can be thought of as **average surprise**
- Before you toss a coin, suppose someone predicts it lands on tails
  - You would be surprised, but not that surprised
- Before you roll a dice, suppose someone predicts it lands on 1
  - You would be a lot more surprised
- To formalize this, we have an idea of a "surprise" function that indicates how surprised you are to see a certain state s:
  - $h(s)$ should be a function of p_s
  - Should be high for small values of p_s and low for large values
  - If we see three events we are surprised at, the surprise should be 3x as much
    - However, the probabilities were multiplied instead of added together
    - Therefore, when probabilities multiply, surprise should add
- This leads to $h(s) = -\log p_s$ for a single state s
- The expected surprise for a whole distribution would be:

$$H = \sum_s -p_s \log(p_s) = \mathbb{E}\left[-\log p_s\right]$$

Cross-Entropy

- When we see real world events, we don't have the exact probability model that we can use to compute entropy
- Instead, we have an internal model, and we compute surprise based on that
- Cross-entropy is $H(P, Q)$, which is the average surprise you will get by observing a random variable governed by P, but you believe in Q

$$H(P, Q) = \sum_s -p_s \log q_s$$

- If P and Q are the same, this just becomes normal entropy
  - $H(P, Q) \geq H(P)$, so you can never be less surprised than if you had just known the true model directly
- This is **not** symmetric

KL Divergence

- How can we quantify the difference between two distributions P and Q?
  - We have cross entropy, but this is already shifted up by a baseline for how surprised we are from just viewing the model P directly
- We can get a more general difference by taking:

$$D_{KL}(P||Q) = H(P, Q) - H(P) = \sum_s p_s \log\left(\frac{p_s}{q_s}\right)$$