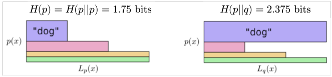
Introduction to Machine Learning

Information Theory
Source Coding and Cross-Entropy



Learning goals

- Know connection between source coding and (cross-)entropy
- Know that the entropy of the source distribution is the lower bound for the average code length

SOURCE CODING AND CROSS-ENTROPY

- For a random source / distribution p, the minimal number of bits to optimally encode messages from is the entropy $H(p)$.
- If the optimal code for a different distribution $q(x)$ is instead used to encode messages from $p(x)$, expected code length will grow.

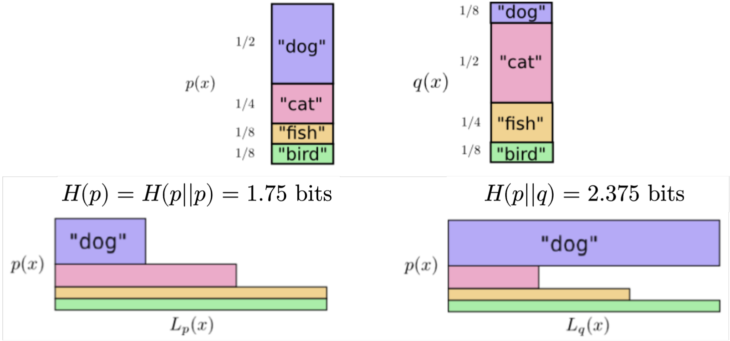


Figure: $L_p(x)$, $L_q(x)$ are the optimal code lengths for $p(x)$ and $q(x)$

SOURCE CODING AND CROSS-ENTROPY / 2

Cross-entropy is the average length of communicating an event from one distribution with the optimal code for another distribution (assume they have the same domain $\mathcal{X}$ as in KL).

$$H(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{1}{q(x)} \right) = - \sum_{x \in \mathcal{X}} p(x) \log \left(q(x) \right)$$

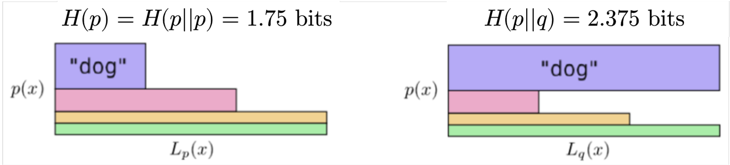
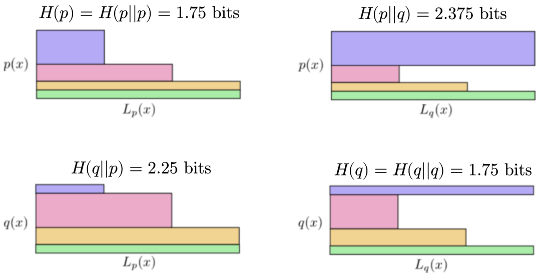


Figure: $L_p(x)$, $L_q(x)$ are the optimal code lengths for $p(x)$ and $q(x)$

We directly see: cross-entropy of p with itself is entropy: $H(p\|p) = H(p)$.

SOURCE CODING AND CROSS-ENTROPY / 3



Credit: Chris Olah

- In top, $H(p||q)$ is greater than $H(p)$ primarily because the blue event that is very likely under p has a very long codeword in q.
- Same, in bottom, for pink when we go from q to p.
- Note that $H(p||q) \neq H(q||p)$.

SOURCE CODING AND CROSS-ENTROPY / 4

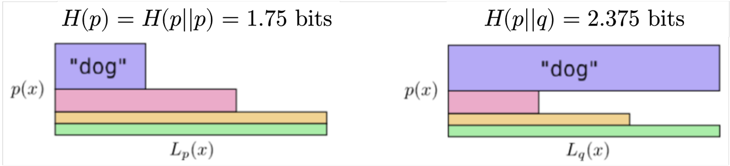


Figure: $L_p(x)$, $L_q(x)$ are the optimal code lengths for $p(x)$ and $q(x)$

- Let x' denote the symbol "dog". The difference in code lengths is:

$$\log\left(\frac{1}{q(x')}\right) - \log\left(\frac{1}{p(x')}\right) = \log\frac{p(x')}{q(x')}$$

- If $p(x') > q(x')$, this is positive, if $p(x') < q(x')$, it is negative.
- The expected difference is KL, if we encode symbols from p:

$$D_{KL}(p\|q) = \sum_{x\in\mathcal{X}} p(x)\cdot\log\frac{p(x)}{q(x)}$$