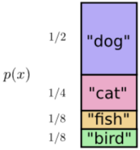
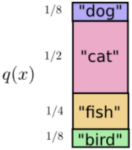
Introduction to Machine Learning

Cross-Entropy, KL and Source Coding





Learning goals

- Know the cross-entropy
- Understand the connection between entropy, cross-entropy, and KL divergence

CROSS-ENTROPY - DISCRETE CASE

- 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.

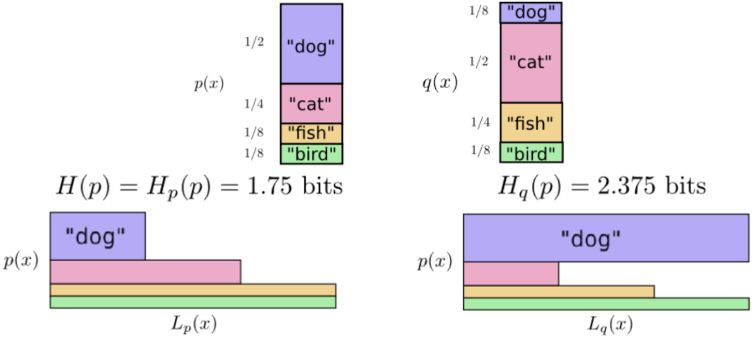


Figure: $L_p(x)$, $L_q(x)$ are the optimal code lengths for $p(x)$ and $q(x)$

CROSS-ENTROPY - DISCRETE CASE

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).

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

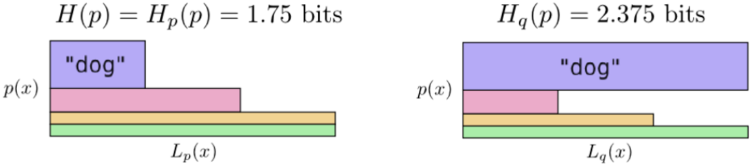
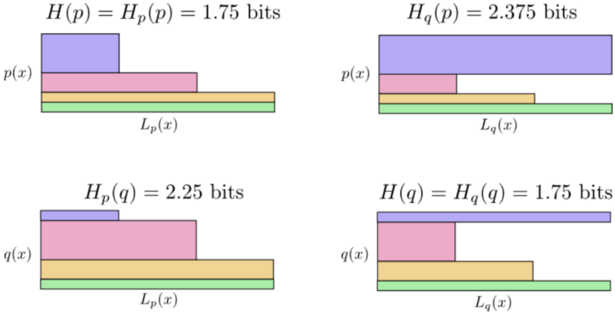


Figure: $L_p(x)$, $L_q(x)$ are the optimal code lengths for $p(x)$ and $q(x)$

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

CROSS-ENTROPY - DISCRETE CASE



Credit: Chris Olah

- In top, $H_q(p)$ 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 $H_q(p) \neq H_p(q)$.

CROSS-ENTROPY - DISCRETE CASE

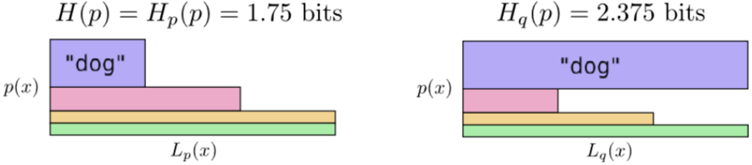


Figure: $L_p(x)$, $L_q(x)$ are the optimal code lengths for $p(x)$ and $q(x)$

- 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:

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

CROSS-ENTROPY - DISCRETE CASE

- Entropy = Avg. nr. of bits if we optimally encode p
- Cross-Entropy = Avg. nr. of bits if we suboptimally encode p with q
- $DL_{KL}(p\|q)$: Difference in bits between the two

We can summarize this also through this identity:

$$H_q(p) = H(p) + D_{KL}(p\|q)$$

This is because:

$$\begin{aligned}
H(p) + D_{KL}(p\|q) &= -\sum_{x\in\mathcal{X}} p(x)\log p(x) + \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \\
&= \sum_{x\in\mathcal{X}} p(x)(-\log p(x) + \log p(x) - \log q(x) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log q(x) = H_q(p)
\end{aligned}$$

CROSS-ENTROPY - CONTINUOUS CASE

For continuous density functions $p(x)$ and $q(x)$:

$$H_p(q) = \int q(x) \ln \left(\frac{1}{p(x)} \right) dx = - \int q(x) \ln \left(p(x) \right) dx$$

- It is not symmetric.
- As for the discrete case, $H_p(q) = h(q) + D_{KL}(q \| p)$ holds.
- Can now become negative, as the $h(q)$ can be negative!

PROOF: MAXIMUM OF DIFFERENTIAL ENTROPY

Claim: For a given variance, the distribution that maximizes differential entropy is the Gaussian.

Proof: Let $g(x)$ be a Gaussian with mean μ and variance σ^2 and $f(x)$ an arbitrary density function with the same variance. Since differential entropy is translation invariant, we can assume $f(x)$ and $g(x)$ have the same mean.

The KL divergence (which is non-negative) between $f(x)$ and $g(x)$ is:

$$\begin{aligned} 0 \leq D_{KL}(f \| g) &= -h(f) + H_g(f) \\ &= -h(f) - \int_{-\infty}^{\infty} f(x) \ln(g(x)) dx \end{aligned} \tag{1}$$

PROOF: MAXIMUM OF DIFFERENTIAL ENTROPY

The second term in (1) is,

$$
\begin{aligned}
\int_{-\infty}^{\infty} f(x)\log(g(x))dx &= \int_{-\infty}^{\infty} f(x)\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)dx \\
&= \int_{-\infty}^{\infty} f(x)\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)dx + \log(e)\int_{-\infty}^{\infty} f(x)\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\
&= -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \log(e)\frac{\sigma^2}{2\sigma^2} = -\frac{1}{2}(\log\left(2\pi\sigma^2\right) + \log(e)) \\
&= -\frac{1}{2}\log\left(2\pi e\sigma^2\right) = -h(g)\,, \qquad (2)
\end{aligned}
$$

where the last equality follows from the normal distribution example of the entropy chapter. Combining (1) and (2) results in

$$h(g) - h(f) \geq 0$$

with equality when $f(x) = g(x)$ (following from the properties of Kullback-Leibler divergence).