Information theory

Entropy

- Consider a random variable X on the set $\{a, b, c, d\}$, with probabilities $P(X = a) = p_a$, $P(X = b) = p_b, \ldots$.
- What is the optimal number of bits to encode the possible values of X?

- Since there are 4 possibilities, we can use 00 for a, 01 for b, 10 for c and 11 for d; i.e. 2 bits.
- If $p_a = p_b = p_c = p_d = \frac{1}{4}$, then on average we expect to use 2 bits to transmit a message containing just the value of X.
- Should we adopt the same encoding scheme if if $p_a = \frac{1}{2}, p_b = \frac{1}{4}, p_c = \frac{1}{8} = p_d$?

- Intuitively we should use fewer bits to encode the more frequently occurring values, and more bits to encode the less frequently occurring ones.
- Eg., we can use 0 for a, 10 for b, 110 for c and 111 for d. Note that we cannot use shorter codes for b, c or d because we need to be able to unambiguously parse a concatenation of the strings, eg. 1110110 decodes uniquely into dac.
- With this encoding scheme, on average we use

$$\left(\frac{1}{2} \times 1\right) + \left(\frac{1}{4} \times 2\right) + \left(\frac{1}{8} \times 3\right) + \left(\frac{1}{8} \times 3\right) = 1.75$$

bits.

Definition

The entropy, $H(X)$ of a discrete random variable is given by

$$H(X) = -\sum_i p_i \log p_i,$$

where we adopt the convention that $0 \log 0 = 0$.

If we use base 2 for the logarithm, the units of entropy are given in *bits*; if the natural logarithm is used, the units are called *nats*.

- Thus, when $p_a = \frac{1}{2}, p_b = \frac{1}{4}, p_c = p_d = \frac{1}{8}$, then

$$H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = 1.75$$

  bits, which is the same as the average number of bits we computed earlier with our encoding scheme.
- In fact, Shannon's source coding theorem (1948) (or noiseless coding theorem) tells us that we cannot do better; i.e. we cannot find a lossless encoding scheme that uses on average fewer bits than the entropy of X; i.e. entropy gives us a lower bound.

- Recall for HW 2 that entropy is maximized when X is a uniform distribution; for n classes, we need

$$\log_2 n$$

bits on average to transmit X, and this is the most bandwidth required amongst all possible distributions of X.
- In contrast, if we know that $p_i = 1$ for some i, then $H(X) = 0$, and we do not need any bandwidth for transmission since we already know the outcome!

Cross entropy

Definition

The cross entropy of two discrete distributions p and q, such that $q_i = 0 \implies p_i = 0$, is given by

$$H(p, q) = -\sum_i p_i \log q_i.$$

If $q_i = 0$ for some i but $p_i > 0$, then $H(p, q) = \infty$.

We can also write $H(X, Y)$ instead when we have two random variables X and Y with distributions p and q respectively.

- We know that $H(p, q) \geq H(p, p)$ for all q and equality occurs when $q = p$.
- Recall that cross entropy loss is used in logistic/softmax regression, where p denotes the target distribution (typically $p_i = 1$ for some i and 0 otherwise; this is the one-hot encoding of $t = i$), and q is the prediction of the model.
- Thus cross entropy gives a measure of how dissimilar q is from p.
- It is not symmetric; i.e. $H(p, q) \neq H(q, p)$ in general.

Kullback-Leibler (KL) divergence (or relative entropy)

Definition

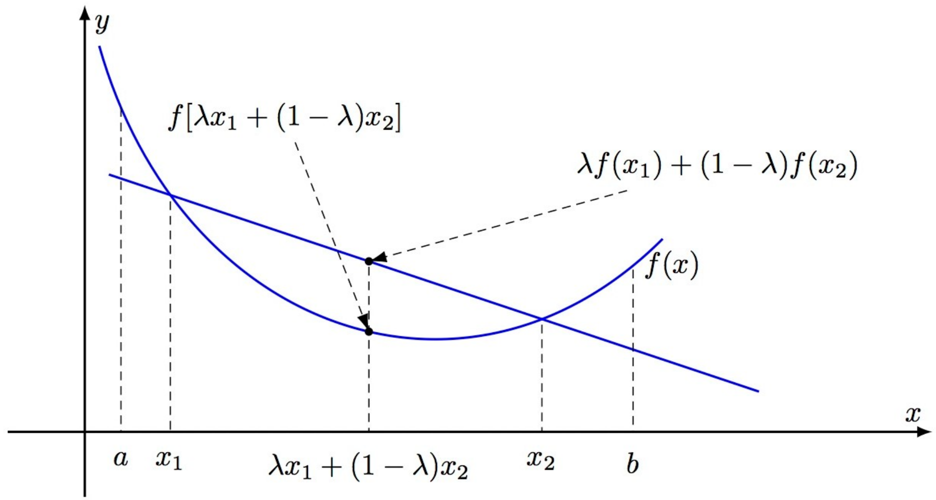
The KL divergence of two discrete distributions p and q such that $q_i = 0 \implies p_i = 0$, is given by

$$D_{KL}(p|q) = H(p,q) - H(p,p) = \sum_i p_i \log \frac{p_i}{q_i}.$$

If $q_i = 0$ for some i but $p_i > 0$, then $H(p,q) = \infty$.

- KL divergence measures the number of extra bits required to transmit X with distribution p, as compared to the optimal code, when we use the sub-optimal coding scheme associated with distribution q.
- As with cross entropy, it is not symmetric.
- We can use source coding theorem to infer that KL divergence is always non-negative, but there is a more direct proof using Jensen's inequality.

Convex functions



Definition

A function $\phi : (a, b) \to \mathbb{R}$ is convex if for all $x, y \in (a, b)$,
$\phi(\lambda x + (1-\lambda)y) \leq \lambda\phi(x) + (1-\lambda)\phi(y)$ for all $\lambda \in (0, 1)$.

Proposition

If a function ϕ is convex, then it is continuous.

Not all convex functions are differentiable, eg. $\phi(x) = |x|$, but we have the following proposition.

Proposition

If a function ϕ has a non-negative second derivative on (a, b), then it is convex.

Jensen's inequality

Theorem

Let $\phi(x)$ be a convex function. If μ is a probability measure, and $f(x)$ and $\phi(f(x))$ are integrable, then

$$\phi\left(\int f(x)\,\mathrm{d}\mu(x)\right) \leq \int \phi(f(x))\,\mathrm{d}\mu(x).$$

Example

$$
\begin{aligned}
Var(X) \geq 0 &\iff \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \geq 0 \\
&\iff \mathbb{E}\left[X^2\right] \geq (\mathbb{E}[X])^2
\end{aligned}
$$

We know the second line is true by applying Jensen's inequality with $\phi(x) = x^2$ and $f(x) = x$.

Proposition

For any two distributions p and q, $D_{KL}(p|q) \geq 0$, and is equal to 0 when $p = q$.

Proof.

$$
\begin{aligned}
D_{KL}(p|q) &= \sum_i p_i \left(-\log \frac{q_i}{p_i} \right) \quad \text{(sum runs over all i such that } p_i > 0\text{)} \\
&\geq -\log \sum_i p_i \left(\frac{q_i}{p_i} \right) \quad \text{(by Jensen's inequality)} \\
&= -\log \sum_i q_i \geq 0.
\end{aligned}
$$

■