Information theory: Basic definitions

Yuval Filmus

July 24, 2018

For a random variable X on a countable domain, recall that we defined the entropy of X by

$$H(X) = \sum_i \Pr[X = i] \log \frac{1}{\Pr[X = i]},$$

where log is base 2. We also defined $h(p)$ as the entropy of a Bernoulli p random variable.

1. Show that $H(X) \geq 0$ always, and $H(X) = 0$ iff X is constant.
2. Show that if X is uniformly distributed over a finite domain of size n, then $H(X) = \log n$.
3. Conditional entropy: Define

   $$H(X|Y) = \mathop{\mathbb{E}}_{y \sim Y}[H(X|Y = y)],$$

   where "$X|Y = y$" is the distribution of X conditioned on the event $Y = y$.

   (a) Show that

   $$H(X|Y) = \sum_{x,y} \Pr[X = x, Y = y] \log \frac{1}{\Pr[X = x|Y = y]}.$$

   (b) Chain rule: show that

   $$H(X, Y) = H(X|Y) + H(Y),$$

   where $H(X, Y)$ is the entropy of the random variable (X, Y).

   (c) When is $H(X|Y) = 0$?
4. Mutual information: Define

   $$I(X; Y) = H(X) + H(Y) - H(X, Y).$$

   (a) Show that $I(X; Y) = I(Y; X)$ and $I(X; X) = H(X)$.

   (b) Show that $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.

(c) Show that

$$I(X;Y) = \sum_{x,y} \Pr[X = x, Y = y] \log \frac{\Pr[X = x, Y = y]}{\Pr[X = x] \Pr[Y = y]}.$$

(d) Use convexity of $\log(1/z)$ and Jensen's inequality to deduce $I(X;Y) \geq 0$.

(e) Conclude that $H(X|Y) \leq H(X)$ and $H(X,Y) \leq H(X) + H(Y)$.

(f) When is $I(X;Y) = 0$? When is $I(X;Y) = H(X)$? When is $H(X|Y) = H(X)$? When is $H(X,Y) = H(X) + H(Y)$?

(g) Guess the definition of $I(X;Y|Z)$.

5. Concavity of entropy: Let $\vec{p}, \vec{q}$ be two vectors of length n. Using the concavity of $z \log(1/z)$, show that for $0 \leq t \leq 1$,

$$tH(\vec{p}) + (1-t)H(\vec{q}) \leq H(t\vec{p} + (1-t)\vec{q}),$$

where $H(\vec{p})$ is the entropy of the random variable X given by $\Pr[X = i] = p_i$.

6. Data processing inequality: Let f be an arbitrary function.

(a) Use the chain rule to show that $H(X, f(X)) = H(X)$.

(b) Use the chain rule again to conclude that $H(f(X)) \leq H(X)$.

(c) Deduce that $H(f(X)|Y) \leq H(X|Y)$.

(d) Use the chain rule to show that $H(Y|X) = H(Y|f(X)) - I(X;Y|f(X))$.

(e) Deduce that $H(Y|X) \leq H(Y|f(X))$.

7. Kullback–Leibler (KL) divergence: Given two probability distributions on the same domain, define

$$D(p\|q) = \sum_t p(t) \log \frac{p(t)}{q(t)}.$$

(a) Use convexity of $\log(1/z)$ to show that $D(p\|q) \geq 0$. When is $D(p\|q) = 0$?

(b) Let p be a probability distribution on $\{1, \ldots, n\}$, and let u be the uniform distribution on $\{1, \ldots, n\}$. Calculate $D(p\|u)$, and deduce that $H(p) \leq \log n$.

(c) Express $I(X;Y)$ using KL divergence.

(d) Show that KL divergence isn't symmetric.

Pinsker's inequality relates KL divergence to total variation distance: $d_{TV}(p,q) \leq \sqrt{2D(p\|q)}$.

Exercise: Let $n \geq 1$ and $k \leq n/2$. Let $X_1, \ldots, X_n$ be the uniform distribution over $\{\vec{v} \in \{0,1\}^n : \sum_i v_i \leq k\}$.

1. Show that $\Pr[X_i = 1] \leq k/n$.
2. Deduce that $H(X_1, \ldots, X_n) \leq nh(k/n)$.
3. Conclude that $\sum_{\ell=0}^{k} \binom{n}{\ell} \leq 2^{nh(k/n)}$.
4. Bonus: Try getting a similar estimate on $\binom{n}{k}$ using Stirling's formula.