Information Theory Cheat Sheet

Simon DeDeo, on behalf of Team LSSP

March 1, 2015

This is a quick cheat-sheet to the formulae that define the fundamental information-theoretic quantities potentially of use to you in your Hackathon. See the main handout at `http://santafe.edu/~simon/it.pdf` for a more detailed guide, and consult your notes from the opening lectures for more. You may also find it useful to recall the Bayesian reasoning material, including `http://santafe.edu/~simon/br.pdf`.

You're always working with a probability distribution over possibilities, which we'll write as $P(X)$, where X, a variable, refers to one of N outcomes. For example, in the case of the game rock-paper-scissors, a player has a particular distribution over one of three outcomes, so X is $\{r, p, s\}$, and N is three. You might sometimes work with two different variables that are somehow connected to each other. A classic example from class would be the move you made now, X, and the move you made last round, Y. We can write $P(X|Y)$ as "the probability of X conditional on Y". Here we'll say that the Y variable has M outcomes; in the case of the previous sentence, $N = M = 3$.

As an explicit example, $P(r|s)$ is the probability of choosing r (rock) given that you last played s (scissors); compute it by counting the number of times you played scissors following by rock, divided by the number of times you played scissors.[1]. You can imagine all sorts of distributions, including $P(X|Z)$, where Z might be defined as the choice of your opponent in the previous round.

All information-theoretic quantities are measured in bits. The most basic is "the **surprise** of outcome i",

$$S(i) = \log_2 \frac{1}{P(X = i)} \tag{1}$$

With that prelude, we have the fundamental formula for **entropy**, a.k.a. **uncertainty**, a.k.a. "average surprise",

$$H(X) = \sum_{i=1}^{N} P(X = i) \log_2 P(X = i), \tag{2}$$

and we have **conditional entropy**, or "the uncertainty in X given that you know Y" as

$$H(X|Y) = \sum_{j=1}^{M} P(Y = j) \left(\sum_{i=1}^{N} P(X = i|Y = j) \log_2 P(X = i|Y = j) \right) \tag{3}$$

[1]Don't get freaked out if your data ends with you playing scissors; don't count this data point, since you don't know what happened next.

and the **mutual information** as

$$MI(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{4}$$

or, the drop in uncertainty in X once you learn the value of Y.

In many cases of interest, we care about a single variable, X, but we're interested in how the probabilities for the different outcomes might differ. An example would be the idea that men have a probability over $\{r, p, s\}$, call it $P(X)$, that is different from that for women, $Q(X)$. For example, famously, men prefer to lead with rock, so $P(r) > Q(r)$.

Given that definition, we have two quantities. The **Kullback-Leibler (KL) divergence**, or "the average surprise you get if you're expecting X to be drawn from the distribution P, but it's actually drawn from distribution Q" is

$$KL(P,Q) = \sum_{i=1}^{N} P(X=i) \log_2 \frac{P(X=i)}{Q(X=i)}. \tag{5}$$

Remember that $KL(P,Q)$ is, generally, not equal to $KL(Q,P)$ (see the main handout to recall why).

Finally, we have the **Jensen-Shannon Distance** (JSD), or "the amount of information that one sample gives you about whether or not you're dealing with distribution P or Q".

$$\begin{aligned} JSD(P,Q) &= \sum_{i=1}^{N} P(X=i) \log_2 \frac{P(X=i)}{\frac{1}{2}(P(X=i)+Q(X=i))} \\ &\quad + Q(X=i) \log_2 \frac{Q(X=i)}{\frac{1}{2}(P(X=i)+Q(X=i))} \\ &= \frac{1}{2}\left[KL\left(P, \frac{1}{2}(P+Q)\right) + KL\left(Q, \frac{1}{2}(P+Q)\right)\right] \end{aligned}$$

The JSD sounds a bit strange, but it's the best way to measure "how different" two distributions are. Unlike KL, it's symmetric, meaning that $JSD(P,Q)$ is equal to $JSD(Q,P)$. We used it to measure how different "Democrat speech" was from "Republican speech", where we defined that as the distribution over words for the Democrats, P, and over Republicans, Q.