10-704: Information Processing and Learning **Fall 2016**

Lecture 2: August 31

Lecturer: Aarti Singh

Note: *These notes are based on scribed notes from Spring15 offering of this course. LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

2.1 Information Quantities

In the previous class, we defined the information content of a random outcome and average information content of a random variable:

- The **Shannon Information Content** of a random outcome x which occurs with probability $p(x)$ is

$$\log_2 \frac{1}{p(x)}.$$

- The **Entropy** in bits is the average uncertainty of a random variable X, i.e. a weighted combination of the Shannon information content of each value x that random variable X could take, weighed by the probability of that value/outcome:

$$H(X) = \sum_{x \epsilon \mathcal{X}} p(x) \log_2 \frac{1}{p(x)} = -\mathbb{E}_{X \sim p}[\log_2 p(X)]$$

  Here $\mathcal{X}$ is the collection of all values x that X can take. It is also known as the alphabet over which X is defined. Note that we are focusing on discrete random variables for now.

  The entropy will turn out to be a fundamental quantity that characterizes the fundamental limit of compression, i.e. the smallest number of bits to which a source distribution or model given by $p(X)$ can be compressed.

We now define some more information quantities that will be useful:

- The **joint entropy** in bits of two random variables X, Y with joint distribution $p(x, y)$ is

$$H(X, Y) = \sum_{x \epsilon \mathcal{X}, y \epsilon \mathcal{Y}} p(x, y) \log_2 \left(\frac{1}{p(x, y)} \right)$$

- The **conditional entropy** in bits of Y conditioned on X is the average uncertainty about Y after observing X.

$$H(Y|X) = \sum_{x \epsilon \mathcal{X}} p(x) H(Y|X = x) = \sum_{x \epsilon \mathcal{X}} p(x) \sum_{y \epsilon \mathcal{Y}} p(y|x) \log_2 \left(\frac{1}{p(y|x)} \right)$$

- Given two distributions p, q for a random variable X, the **relative entropy** between p and q is

$$D(p||q) = \mathbb{E}_{X\sim p}[\log\left(\frac{1}{q(X)}\right)] - \mathbb{E}_{X\sim p}[\log\left(\frac{1}{p(X)}\right)] = \mathbb{E}_p[\log\left(\frac{p}{q}\right)] = \sum_x p(x)\log\left(\frac{p(x)}{q(x)}\right)$$

  The base of the log can be 2, if information is measured in bits, or e if information is measured in nats. The relative entropy is also known as the **Information divergence** or the **Kullback-Leibler (KL) divergence**.

  The relative entropy is the cost incurred if we used distribution q to encode X, when the true underlying distribution is p. Consider the example from last lecture, where $p(X) \sim \text{uniform}(\{0, 1, \dots, 63\})$ and we need 6 yes/no questions to guess each outcome, hence the average information content or entropy is 6 bits. If instead we consider the model $q(X) \sim \text{uniform}(\{0, 1, \dots, 127\})$, then 7 questions are needed for each outcome. The extra price paid to encode an outcome x using model q when x is generated according to the true model p is 1 bit, which is the relative entropy. We will see below that is also the cost incurred in excess risk, under the negative loss likelihood loss, when the true model is p but the estimated model is q.

- The **Mutual Information** between X and Y is the KL-divergence between the joint distribution and the product of the marginals. Formally:

$$I(X;Y) = D(p(x,y)||p(x)p(y)), \tag{2.1}$$

  where $p(x, y)$ is the joint distribution of X, Y and $p(x), p(y)$ are the corresponding marginal distributions. Thus, $p(x)p(y)$ denotes the joint distribution that would result if X, Y were independent.

  The mutual information quantifies how much dependence there is between two random variables. If $X \perp Y$ then $p(x, y) = p(x)p(y)$ and $I(X;Y) = 0$.

  The mutual information will turn out to be a fundamental quantity that characterizes the fundamental limit of transmission, i.e. the smallest number of bits that can be reliably transmitted through a noisy channel with input X and output Y.

2.2 Connection to Maximum Likelihood Estimation

Suppose $X = (X_1, \dots, X_n)$ are generated from a distribution p (for example $X_i \sim p$ i.i.d.). In maximum likelihood estimation, we want to find a distribution q from some family $\mathcal{Q}$ such that the likelihood of the data is maximized.

$$\arg\max_{q\in\mathcal{Q}} q(X) = \arg\max_{q\in\mathcal{Q}} \log q(X) = \arg\min_{q\in\mathcal{Q}} -\log q(X)$$

In machine learning, we often define a loss function. In this case, the loss function is the negative log loss: $\text{loss}(q, X) = -\log q(X) = \log(1/q(X))$. The expected value of this loss function is the risk: $\text{Risk}(q) = E_p[\log(1/q(X))]$. We want to find a distribution q that minimizes the risk. However, notice that minimizing the risk with respect to a distribution q is exactly minimizing the relative entropy between p and q. This is because:

$$\text{Risk}(q) = \mathbb{E}_p[\log(1/q(X))] = \mathbb{E}_p\left[\log\frac{p(X)}{q(X)}\right] + \mathbb{E}_p\left[\log\frac{1}{p(X)}\right] = D(p||q) + \text{Risk}(p)$$

As we will see below, the relative entropy is always non-negative, and hence the risk is minimized by setting q equal to p. Thus the minimum risk $R^\star = \text{Risk}(p) = H(p)$, the entropy of distribution p. The excess risk, $\text{Risk}(q) - R^\star$ is precisely the relative entropy between p and q.

2.3 Fundamental Limits in Information Theory

The **source coding** model is as follows:

$$\text{Source Model} \xrightarrow{X_1..X_n} \boxed{\text{Compressor}} \xrightarrow{b_1,\ldots,b_m bits} \boxed{\text{Decompresser}} \rightarrow \text{Receiver}$$

Let the data source be generated according to some distribution $p(X)$. The rate of a source code is defined as the average number of bits used to encode one source symbol , i.e. $\mathbf{E}_p[\frac{\text{codelength}}{\text{\#src symbols}}] = \mathbf{E}_p[m/n]$. If the rate of a code is less than the source entropy $H(X)$, that is $\mathbb{E}_p[\frac{\text{codelength}}{\text{\#src symbols}}] < H(X)$ then perfect reconstruction is not possible. A distribution cannot be compressed below its entropy without loss. We will state and prove it rigorously later in the course.

The **channel coding** model is as follows:

$$\xrightarrow{b_1,\ldots,b_m bits} \boxed{\text{Channel Encoder}} \xrightarrow{X_1,\ldots,X_n} \boxed{\text{Channel } p(y|x)} \xrightarrow{Y_1,\ldots,Y_n} \boxed{\text{Channel Decoder}} \rightarrow$$

The rate of a channel code is defined as the average number of bits transmitted per channel use, i.e. $\mathbf{E}_p[\frac{\text{\#src symbols}}{\text{codelength}}] = \mathbf{E}_p[m/n]$. If the rate of a code is greater than the channel capacity $C := \max_{p(X)} I(X,Y)$, then perfect reconstruction is not possible.

The **inference** problem is similar to the channel coding problem except we do not design the encoder:

In the density estimation setting with $p_\theta, \theta \epsilon \Theta$:

$$\theta \rightarrow \boxed{\text{Channel: } p_\theta(X)} \xrightarrow{X_1,\ldots,X_n} \boxed{\text{Decoder}} \xrightarrow{\hat{\theta}}$$

We can denote the estimated model as $q = p_{\hat{\theta}}$. Under log loss:

$$\text{Excess Risk}(q) = \text{Risk}(q) - \text{Risk}(p) = D(p||q)$$

Fundamental limits of inference problems are often characterized by minmax lower bounds, i.e. the smallest possible excess risk that any estimator can achieve for a class of models. For the density estimation problem, the minmax excess risk is $\mathrm{inf_q sup_{p\epsilon\mathcal{P}}\ D(p||q)}$ and we will show that this is equal to the capacity C of the corresponding channel. This would imply that for all estimators q, $\mathrm{sup_{p\epsilon\mathcal{P}}\ D(p||q) \geq C}$.

We will state and prove these results formally later in the course. Information theory will help us identify these fundamental limits of data compression, transmission and inference; and in some cases also demonstrate that the limits are achievable. The design of efficient encoders / decoders / estimators that achieve these limits is the common objective of Signal Processing and Machine Learning algorithms.

2.4 Useful Properties of Information Quantities

1. Entropy is always non-negative: $H(X) \geq 0, H(X) = 0 \Leftrightarrow X$ is constant
   Proof: $0 \leq p(x) \leq 1$ implies that $\log \frac{1}{p(x)} \geq 0$ ■

   For example, consider a binary random variable $X \sim \mathit{Bernoulli}(\theta)$ Then $\theta = 0$ or $\theta = 1$, then the outcome is certain (a constant) and implies that $H(X) = 0$. If $\theta = \frac{1}{2}$, then $H(X) = 1$ (which is the maximum entropy for a binary random variable since the distribution is uniform).

2. $H(X) \leq \log|\mathcal{X}|$ where $\mathcal{X}$ is the set of all outcomes with non-zero probability. Equality is achieved iff X is uniform.

   Proof: Let u be the uniform distribution over X, i.e. $u(x) = \frac{1}{|\mathcal{X}|}$ and let $p(x)$ be the probability mass function for X.

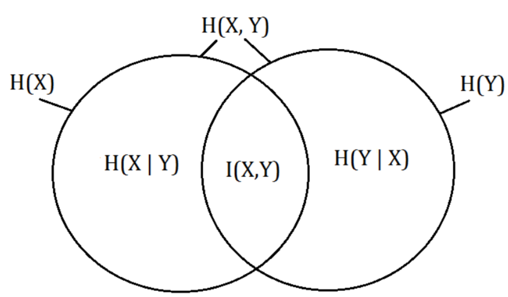
$$D(p||u) = \sum p(x) \log \frac{p(x)}{u(x)} = \log|\mathcal{X}| - H(X)$$

$$0 \leq D(p||u) = \log|\mathcal{X}| - H(X) \quad \text{by non negativity of relative entropy (stated and proved below)}$$

■

3. Chain Rule: $H(X, Y) = H(X) + H(Y|X)$

4. The following relations hold between entropy, conditional entropy, joint entropy, and mutual information:



   (a) H(X,Y) = H(X) + H(Y|X) = H(Y) + H(Y|X)
   (b) I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = I(Y,X)
   (c) I(X,Y) = H(X,Y) - H(X|Y) - H(Y|X)
   (d) I(X,Y) = H(X) + H(Y) - H(X,Y)

5. (**Gibbs Information Inequality**) $D(p||q) \geq 0, = 0$ if and only if $p(x) = q(x)$ for all x.
   Proof: Define the support of p to be $\mathcal{X} = \{x : p(x) > 0\}$

$$\begin{aligned}
-D(p||q) &= -\sum_{x \epsilon \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x \epsilon \mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \\
&\leq \log \sum_{x \epsilon \mathcal{X}} p(x) \frac{q(x)}{p(x)} \\
&= \log \sum_{x \epsilon \mathcal{X}} q(x) \leq \log 1 = 0
\end{aligned}$$

The first inequality is Jensen's inequality[1] since log is concave. Because log is strictly concave we have equality in the first inequality only if p is a constant distribution or if $\frac{q(x)}{p(x)}$ is a constant c, for all x (i.e. if $q(x) = cp(x)$). The second inequality is tight only when that constant $c = 1$ since $\sum_{x \epsilon \mathcal{X}} p(x) = 1$. ∎

6. As a corollary, we get that $I(X, Y) = D(p(x, y)||p(x)p(y)) \geq 0$ and $= 0$ iff X, Y are independent, that is, $p(x, y) = p(x)p(y)$.

7. Conditioning cannot increase entropy, i.e. information always helps.

$$H(X|Y) \leq H(X)$$

with equality iff X and Y are independent.

Proof: $0 \leq I(X; Y) = H(X) - H(X|Y)$ ∎

[1] For a concave function, the (weighted) average of function values at two points is less than function value at (weighted) average of the two points.