Lecture 1: Entropy and mutual information

1 Introduction

Imagine two people Alice and Bob living in Toronto and Boston respectively. Alice (Toronto) goes jogging whenever it is not snowing heavily. Bob (Boston) doesn't ever go jogging.

Notice that Alice's actions give information about the weather in Toronto. Bob's actions give no information. This is because Alice's actions are random and correlated with the weather in Toronto, whereas Bob's actions are deterministic.

How can we quantify the notion of information?

2 Entropy

Definition The *entropy* of a discrete random variable X with pmf $p_X(x)$ is

$$H(X) = -\sum_x p(x) \log p(x) = -\mathbb{E}[\ \log(p(x))\ ] \tag{1}$$

The entropy measures the expected uncertainty in X. We also say that $H(X)$ is approximately equal to how much *information* we learn on average from one instance of the random variable X.

Note that the base of the algorithm is not important since changing the base only changes the value of the entropy by a multiplicative constant.
$H_b(X) = -\sum_x p(x) \log_b p(x) = \log_b(a)[\sum_x p(x) \log_a p(x)] = \log_b(a) H_a(X)$. Customarily, we use the base 2 for the calculation of entropy.

2.1 Example

Suppose you have a random variable X such that:

$$X = \begin{cases} 0 & \text{with prob } p \\ 1 & \text{with prob } 1-p, \end{cases} \tag{2}$$

then the entropy of X is given by

$$H(X) = -p \log p - (1-p)\log(1-p) = H(p) \tag{3}$$

Note that the entropy does not depend on the values that the random variable takes (0 and 1 in this case), but only depends on the probability distribution $p(x)$.

2.2 Two variables

Consider now two random variables X, Y jointly distributed according to the p.m.f $p(x, y)$. We now define the following two quantities.

Definition The *joint entropy* is given by

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y). \tag{4}$$

The joint entropy measures how much uncertainty there is in the two random variables X and Y taken together.

Definition The *conditional entropy* of X given Y is

$$H(X|Y) = -\sum_{x,y} p(x, y) \log p(x|y) = -\mathbb{E}[\ \log(p(x|y))\ ] \tag{5}$$

The conditional entropy is a measure of how much uncertainty remains about the random variable X when we know the value of Y.

2.3 Properties

The entropic quantities defined above have the following properties:

- **Non negativity:** $H(X) \geq 0$, entropy is always non-negative. $H(X) = 0$ iff X is deterministic.
- **Chain rule:** We can decompose the joint entropy as follows:

  $$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X^{i-1}), \tag{6}$$

  where we use the notation $X^{i-1} = \{X_1, X_2, \ldots, X_{i-1}\}$.

  For two variables, the chain rule becomes:

  $$\begin{aligned} H(X, Y) &= H(X|Y) + H(Y) &(7)\\ &= H(Y|X) + H(X). &(8) \end{aligned}$$

  Note that in general $H(X|Y) \neq H(Y|X)$.
- **Monotonicity:** Conditioning always reduces entropy:

  $$H(X|Y) \leq H(X). \tag{9}$$

  In other words "information never hurts".

- **Maximum entropy:** Let $\mathcal{X}$ be set from which the random variable X takes its values (sometimes called the *alphabet*), then

  $$H(X) \leq \log |\mathcal{X}|. \tag{10}$$

  The above bound is achieved when X is uniformly distributed.

- **Non increasing under functions:** Let X be a random variable and let $g(X)$ be some deterministic function of X. We have that:

  $$H(X) \geq H(g(X)), \tag{11}$$

  with equality iff g is invertible.
  Proof: We will the two different expansions of the chain rule for two variables.

  $$\begin{aligned} H(X, g(X)) &= H(X, g(X)) & (12) \\ H(X) + \underbrace{H(g(X)|X)}_{=0} &= H(g(X)) + H(X|g(X)), & (13) \end{aligned}$$

  so we have

  $$H(X) - H(g(X) = H(X|g(X)) \geq 0. \tag{14}$$

  with equality if and only if we can deterministically guess X given $g(X)$, which is only the case if g is invertible. □

3 Continuous random variables

Similarly to the discrete case we can define entropic quantities for continuous random variables.

Definition The *differential entropy* of a continuous random variable X with p.d.f $f(x)$ is

$$h(X) = -\int f(x) \log f(x) dx = -\mathbb{E}[\ \log(f(x))\ ] \tag{15}$$

Definition Consider a pair of continuous random variable (X, Y) distributed according to the joint p.d.f $f(x, y)$. The *joint entropy* is given by

$$h(X, Y) = -\int\int f(x, y) \log f(x, y) dx dy, \tag{16}$$

while the *conditional entropy* is

$$h(X|Y) = -\int\int f(x, y) \log f(x|y) dx dy. \tag{17}$$

3.1 Properties

Some of the properties of the discrete random variables carry over to the continuous case, but some do not. Let us go through the list again.

- **Non negativity doesn't hold:** $h(X)$ can be negative.
  Example: Consider the R.V. X uniformly distributed on the interval $[a, b]$. The entropy is given by
  $$h(X) = -\int_a^b \frac{1}{b-a} \log \frac{1}{b-a} dx = \log(b-a), \tag{18}$$
  which can be a negative quantity if $b - a$ is less than 1.

- **Chain rule** holds for continuous variables:
  $$\begin{aligned} h(X,Y) &= h(X|Y) + h(Y) && (19) \\ &= h(Y|X) + h(X). && (20) \end{aligned}$$

- **Monotonicity:**
  $$h(X|Y) \leq h(X) \tag{21}$$
  The proof follows from the non-negativity of mutual information (later).

- **Maximum entropy:** We do not have a bound for general p.d.f functions $f(x)$, but we do have a formula for power-limited functions. Consider a R.V. $X \sim f(x)$, such that
  $$\mathbb{E}[x^2] = \int x^2 f(x) dx \leq P, \tag{22}$$
  then
  $$\max h(X) = \frac{1}{2} \log(2\pi e P), \tag{23}$$
  and the maximum is achieved by $X \sim \mathcal{N}(0, P)$.

  To verify this claim one can use standard Lagrange multiplier techniques from calculus to solve the problem $\max h(f) = -\int f \log f dx$, subject to $\mathbb{E}[x^2] = \int x^2 f dx \leq P$.

- **Non increasing under functions:** Doesn't necessarily hold since we can't guarantee $h(X|g(X)) \geq 0$.

4 Mutual information

Definition The *mutual information* between two discreet random variables X, Y jointly distributed according to $p(x, y)$ is given by

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} && (24) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y). && (25) \end{aligned}$$

We can also define the analogous quantity for continuous variables.

Definition The *mutual information* between two continuous random variables X, Y with joint p.d.f $f(x, y)$ is given by

$$I(X;Y) \quad = \quad \int\!\!\int f(x,y)\log\frac{f(x,y)}{f(x)f(y)}dxdy. \tag{26}$$

For two variables it is possible to represent the different entropic quantities with an analogy to set theory. In Figure 4 we see the different quantities, and how the mutual information is the uncertainty that is common to both X and Y.

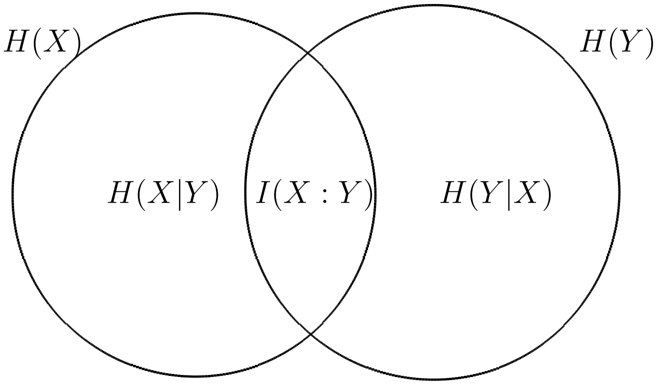

Figure 1: Graphical representation of the conditional entropy and the mutual information.

4.1 Non-negativity of mutual information

In this section we will show that

$$I(X;Y) \geq 0, \tag{27}$$

and this is true for both the discrete and continuous cases.

Before we get to the proof, we have to introduce some preliminary concepts like Jensen's inequality and the relative entropy.

Jensen's inequality tells us something about the expected value of a random variable after applying a convex function to it.

We say a function is *convex* on the interval $[a, b]$ if, $\forall x_1, x_2 \in [a, b]$ we have:

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2). \tag{28}$$

Another way stating the above is to say that the function always lies below the imaginary line joining the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$. For a twice-differentiable function $f(x)$, convexity is equivalent to the condition $f''(x) \geq 0,\ \forall x \in [a, b]$.

Lemma *Jensen's inequality* states that for any convex function $f(x)$, we have

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x]). \tag{29}$$

The proof can be found in [Cover & Thomas].

Note that an analogue of Jensen's inequality exists for *concave* functions where the inequality simply changes sign.

Relative entropy A very natural way to measure the distance between two probability distributions is the *relative entropy*, also sometimes called the Kullback-Leibler divergence.

Definition The *relative entropy* between two probability distributions $p(x)$ and $q(x)$ is given by

$$D(p(x)||q(x)) = \sum_x p(x) \log \frac{p(x)}{q(x)}. \tag{30}$$

The reason why we are interested in the relative entropy in this section is because it is related to the mutual information in the following way

$$I(X;Y) = D(p(x,y)||p(x)p(y)). \tag{31}$$

Thus, if we can show that the relative entropy is a non-negative quantity, we will have shown that the mutual information is also non-negative.

Proof of non-negativity of relative entropy: Let $p(x)$ and $q(x)$ be two arbitrary probability distributions. We calculate the relative entropy as follows:

$$\begin{aligned}
D(p(x)||q(x)) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\
&= -\sum_x p(x) \log \frac{q(x)}{p(x)} \\
&= -\mathbb{E}\left[\log \frac{q(x)}{p(x)}\right] \\
&\geq -\log\left(\mathbb{E}\left[\frac{q(x)}{p(x)}\right]\right) \qquad \text{(by Jensen's inequality for concave function } \log(\cdot)\text{)} \\
&= -\log\left(\sum_x p(x)\frac{q(x)}{p(x)}\right) \\
&= -\log\left(\sum_x q(x)\right) \\
&= 0.
\end{aligned}$$

□

4.2 Conditional mutual information

Definition Let X, Y, Z be jointly distributed according to some p.m.f. $p(x, y, z)$. The *conditional mutual information* between X, Y given Z is

$$\begin{aligned} I(X;Y|Z) &= -\sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \qquad (32) \\ &= H(X|Z) - H(X|YZ) \\ &= H(XZ) + H(YZ) - H(XYZ) - H(Z). \end{aligned}$$

The conditional mutual information is a measure of how much uncertainty is shared by X and Y, but not by Z.

4.3 Properties

- **Chain rule:** We have the following chain rule

$$I(X \ ; \ Y_1 Y_2 \dots Y_n) = \sum_{i=1}^{n} I(X; Y_i | Y^{i-1}), \qquad (33)$$

where we have used again the shorthand notation $Y^{i-1} = \{Y_1, Y_2, \dots, Y_{i-1}\}$.

- **No monotonicity:** Conditioning can either increase or decrease the mutual information between two variables, so

$$I(X;Y|Z) \ngeq I(X;Y), \qquad \text{and} \qquad I(X;Y|Z) \nleq I(X;Y). \qquad (34)$$

To illustrate the last point, consider the following two examples where conditioning has different effects. In both cases we will make use of the following equation

$$\begin{aligned} I(X;YZ) &= I(X;YZ) \\ I(X;Y) + I(X;Z|Y) &= I(X;Z) + I(X;Y|Z). \qquad (35) \end{aligned}$$

Increasing example: If we have some X, Y, Z such that $I(X;Z) = 0$ (which means X and Z are independent variables), then equation (35) becomes:

$$I(X;Y) + I(X;Z|Y) = I(X;Y|Z), \qquad (36)$$

so $I(X;Y|Z) - I(X;Y) = I(X;Z|Y) \geq 0$, which implies

$$I(X;Y|Z) \geq I(X;Y). \qquad (37)$$

Decreasing example: On the other hand if we have a situation in which $I(X;Z|Y) = 0$, equation (35) becomes:

$$I(X;Y) = I(X;Z) + I(X;Y|Z), \qquad (38)$$

which in implies that $I(X;Y|Z) \leq I(X;Y)$.

So we see that conditioning of the mutual information can both increase or decrease it depending on the situation.

5 Data processing inequality

For three variables X, Y, Z one situation which is of particular interest is when they form a Markov chain: $X \to Y \to Z$. This relation is implies that the probability distribution $p(x, z|y) = p(x|y)p(z|y)$ which in turn implies that $I(X; Z|Y) = 0$ like in the example above.

This situation often occurs when we have some input X that gets transformed by a channel to give an output Y and then we want to apply some processing to obtain a signal Z as illustrated below.

$$X \to \boxed{\text{Channel}} \to Y \to \boxed{\text{Processing}} \to Z$$

In this case we have the *data processing inequality*:

$$I(X; Z) \leq I(X; Y). \tag{39}$$

In other words, processing cannot increase the information contained in a signal.