Contents

Chapter 1

Introduction

1.1 Entropy

Let X be a random variable with probability mass function $p(x)$, then the **entropy** of X is defined as

$$H(X) = \mathbb{E}[-\log(p(X))] = -\sum_{x \in \mathcal{X}} p(x) \log(p(x))$$

which intuitively measures the uncertainty of a single variable. Depending one the base of the logarithm, the entropy is measured in bits, for base 2, nats, for base e. Entropy can also be viewed as the average amount information revealed after sampling X. We can define conditional entropy of X given that $Y = y$ to be

$$H(X|Y = y) = -\sum_{x \in \mathcal{X}} p_{X|Y}(x|y) \lg\left(\frac{p_{XY}(x, y)}{p_Y(y)}\right)$$

and conditional entropy of X given Y is

$$\begin{aligned} H(X|Y) &= \sum_{y \in \mathcal{Y}} p_Y(y) H(X|Y = y) \\ &= -\sum_{y} \sum_{x} p_{XY}(x, y) \lg\left(\frac{p_{XY}(x, y)}{p_Y(y)}\right) \end{aligned}$$

Lastly, the joint entropy to variables is defineds

$$H(X, Y) = \mathbb{E}_{X,Y}[-\log(p_{XY}(X, Y))] = -\sum_{x,y} p_{XY}(x, y) \lg(p_{XY}(x, y))$$

From now on we omit the subscript for the PMFs unless it can not be inferred from the context.

Proposition 1.1 (Chain rule for entropy). *For any two random variables X and Y*

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

furthermore if Z is another random variable then

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

which then can be used to generalize the chain rule

$$H(X_1, \ldots, X_n) = \sum i = 1^n H(X_i|X_{i-1}, \ldots, H(X_1))$$

Proof. For the conditional case

$$\begin{aligned}
H(X|Z) &= -\sum_{x,z} p(x,z)\lg\left(\frac{p(x,z)}{p(z)}\right) \\
H(Y|X,Z) &= -\sum_{x,y,z} p(x,y,z)\lg\left(\frac{p(x,y,z)}{p(x,z)}\right) \\
\implies H(X|Z) + H(Y|X,Z) &= -\sum_{x,y,z} p(x,y,z)\lg\left(\frac{p(x,y,z)}{p(z)}\right) \\
&= H(X,Y|Z)
\end{aligned}$$

1.2 Mutual information

Mutual information is the reduction in entropy due to another random variable.

$$\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= \mathbb{E}_{x,y}\left[\lg\left(\frac{p(X,Y)}{p(X)p(Y)}\right)\right] \\
&= \sum_x \sum_y p(x,y)\lg\left(\frac{p(x,y)}{p(x)p(y)}\right) \\
&= H(Y) - H(Y|X) = I(Y;X)
\end{aligned}$$

Proposition 1.2. *$I(X;Y)$ is zero if and only if X and Y are independent.*

For conditional mutual information we have

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$

Proposition 1.3 (Chain rule for mutual information). *For a random variable Y and random variables $X_1,\dots,X_n$ we have*

$$I(X_1,\dots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_{i-1},\dots,X_1)$$

Proof. We have

$$\begin{aligned}
I(X_1,\dots,X_n;Y) &= H(X_1,\dots,X_n) - H(X_1,\dots,X_n|Y) \\
&= \sum_{i=1}^{n} H(X_i|X_{i-1},\dots,X_1) - H(X_i|X_{i-1},\dots,X_1,Y) \\
&= \sum_{i=1}^{n} I(X_i;Y|X_{i-1},\dots,X_1)
\end{aligned}$$

1.3 Channel Capacity

A *communication channel* is a system in which output depends probabilistically on its input. It is characterized by a probability transition matrix $p(y|x)$. **Capacity** of a communication channel with input X and output Y is defined as

$$C = \max_{p(x)} I(X;Y)$$

1.4 Relative entropy

Relative entropy or *Kullback–Leibler divergence* measures how one probability distribution differs from another.

$$D(p||q) = \mathbb{E}_{p(x)}\left[\lg\left(\frac{p(X)}{q(X)}\right)\right] = \sum_x p(x)\lg\left(\frac{p(x)}{q(x)}\right)$$

Even though it is not a metric, if $D(p||q) = 0 \implies p = q$.

Note that

$$I(X;Y) = \sum_{x,y} p(x,y)\lg\left(\frac{p(x,y)}{p(x)p(y)}\right) = D(p(x,y)||p(x)p(y))$$

Conditional relative entropy is defined as

$$\begin{aligned} D(p(y|x)||q(y|x)) &= \mathbb{E}_{p(x,y)}\left[\lg\left(\frac{p(Y|X)}{q(Y|X)}\right)\right] \\ &= \sum_x p(x)\sum_y p(y|x)\lg\left(\frac{p(y|x)}{q(y|x)}\right) \end{aligned}$$

Similarly we define the following chain rule

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$