Lecture-02: Mutual Information

1 Mutual Information

Definition 1.1. The **joint entropy** (in bits) of a pair of r.v. $(X,Y) \sim p_{X,Y}(x,y)$ is denoted

$$H(X,Y) \triangleq \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y)\log_2 \frac{1}{p_{X,Y}(x,y)} = \mathbb{E}\left[\log_2 \frac{1}{p_{X,Y}(X,Y)}\right].$$

Notice that this is identical to $H(Z)$ with $Z=(X,Y)$.

Definition 1.2. For a pair of r.v. $(X,Y)\sim p_{X,Y}(x,y)$, the **conditional entropy** (in bits) of Y given X is denoted

$$H(Y|X) \triangleq \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y)\frac{1}{\log_2 p_{Y|X}(y|x)} = \mathbb{E}\left[\frac{1}{\log_2 p_{Y|X}(Y|X)}\right].$$

Notice that this equals entropy of the conditional distribution $p_{Y|X}(y|x)$ averaged over x.

Definition 1.3. For a pair of r.v. $(X,Y)\sim p_{X,Y}(x,y)$, the **mutual information** (in bits) between X and Y is denoted

$$I(X;Y) \triangleq \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y)\log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} = \mathbb{E}\left[\log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right].$$

Lemma 1.4. *Basic properties of joint entropy and mutual information:*

1. *(chain rule of entropy)* $H(X,Y)=H(X)+H(Y|X)$. *If* X *and* Y *are independent,* $H(X,Y)=H(X)+H(Y)$.

   Proof. Take the expectation of $\log_2 \frac{1}{p_{X,Y}(x,y)} = \log_2 \frac{1}{p_X(x)} + \log_2 \frac{1}{p_{Y|X}(y|x)}$ and note that $p_{Y|X}(y|x)=p_Y(y)$ for all x,y if X and Y are independent. □

2. *(mutual information) The mutual information satisfies*

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

   Proof. Take the expectation of $\log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} = \log_2 \frac{1}{p_X(x)} + \log_2 \frac{1}{p_Y(y)} - \log_2 \frac{1}{p_{X,Y}(x,y)}$ and apply the chain rule as needed. Also, symmetry follows from swapping X,Y and x,y in the sum because $p_{X,Y}(x,y)=p_{Y,X}(y,x)$. □

Example 1.5. Let $\mathcal{X}=\mathcal{Y}=\{0,1\}$ and $p_{X,Y}(x,y)=\frac{\rho}{2}\mathbb{1}_{\{x\neq y\}}+\frac{(1-\rho)}{2}\mathbb{1}_{\{x=y\}}$. It follows that $p_X(x)=p_Y(y)=\frac{1}{2}$, and hence $H(X)=H(Y)=1$. Since $p_{Y|X}(y|x)\in\{\rho,1-\rho\}$, it follows that $H(Y|X)=\mathcal{H}(\rho)$. Thus, we have $I(X;Y)=H(Y)-H(Y|X)=1-\mathcal{H}(\rho)$. The conditional distribution $p_{Y|X}$ called the **binary symmetric channel** with error probability ρ and denoted by BSC(ρ).

Definition 1.6. The **Kullback-Liebler (KL) divergence** (in bits) between distributions $p(x)$ and $q(x)$, defined on the same support $\mathcal{X}$, is denoted

$$D(p\|q) \triangleq \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)},$$

where we assume $0 \log_2 \frac{0}{q} = 0$ for $q \in [0,1]$ and $p \log_2 \frac{p}{0} = \infty$ for $p > 0$. Thus, $D(p\|q) = \infty$ if there is any $x \in \mathcal{X}$ such that $p(x) > 0$ and $q(x) = 0$.

Remark 1.7. The divergence is non-negative and equal to 0 iff $p(x) = q(x)$ for all $x \in \mathcal{X}$. Thus, it behaves something like a metric on the space of distributions. It is not exactly a metric, however, because it is not symmetric.

Example 1.8. For $\mathcal{X} = \{0,1\}$, let $p(1) = r$ define a Bernoulli(r) distribution and $q(1) = s$ define a Bernoulli(s) distribution. Then, the divergence between p and q is given by

$$\mathcal{D}(r\|s) \triangleq r \log_2 \frac{r}{s} + (1-r) \log_2 \frac{1-r}{1-s}.$$

Example 1.9. Let X be the number of ones in a length-n vector of i.i.d. Bernoulli(s) random variables. Then, the probability the resulting vector has exactly rn ones is given by

$$\mathbb{P}(X = rn) = \binom{n}{rn} s^{rn} (1-s)^{n(1-r)}.$$

Using the results from the previous example, one can see that this equals

$$\frac{1 + O(\frac{1}{nr(1-r)})}{\sqrt{2\pi nr(1-r)}} 2^{n\mathcal{H}(r)} 2^{n[r \log_2 s + (1-r) \log_2 (1-s)]} = \frac{1 + O(\frac{1}{nr(1-r)})}{\sqrt{2\pi nr(1-r)}} 2^{-n\mathcal{D}(r\|s)}.$$

Thus, we see that exponential decay rate is determined by the divergence between a Bernoulli(r) distribution and a Bernoulli(s) distribution. This example highlights the connection between information theory and the theory of large deviations.

Definition 1.10. A function $f : \mathbb{R} \to \mathbb{R}$ is called **convex** on the interval (a,b) if, for all $x_1, x_2 \in (a,b)$ and $\lambda \in [0,1]$,

$$f(\lambda x_1 + (1-\lambda) x_2) \leqslant \lambda f(x_1) + (1-\lambda) f(x_2).$$

It is called **strictly convex** if equality holds only if $\lambda = 0$ or $\lambda = 1$. For a (strictly) convex function f, the function $-f$ is called (strictly) **concave**.

Lemma 1.11 (Jensen's Inequality). *If f is convex and X is a real random variable, then*

$$\mathbb{E}[f(X)] \geqslant f(\mathbb{E}[X]).$$

If f is strictly convex, the equality occurs iff $X = \mathbb{E}[X]$ with probability 1. The inequality is simply reversed if f is (strictly) concave.

Proof. Since f is convex, any tangent line to its graph must lower bound the function. Thus, for any $x_0 \in \mathbb{R}$, there is a constant $a \in \mathbb{R}$ such that the linear function $a(x - x_0) + f(x_0)$ lower bounds $f(x)$. If we choose $x_0 = \mathbb{E}[X]$, then it follows that

$$\mathbb{E}[f(X)] \geqslant \mathbb{E}[a(X - x_0) + f(x_0)] = a\mathbb{E}[X] - a\mathbb{E}[X] + f(\mathbb{E}[X]) = f(\mathbb{E}[X]).$$

If f is strictly convex, then equality in the tangent lower bound occurs only at $x = x_0$. Thus, Jensen's inequality is strict unless $X = \mathbb{E}[X]$ with probability 1. □

Theorem 1.12 (Non-Negativity of Divergence). *For $x \in \mathcal{X}$, let $p(x)$ and $q(x)$ be two discrete distributions. Then, $D(p\|q) \geqslant 0$, with equality iff $p(x) = q(x)$ holds for all $x \in \mathcal{X}$. This result holds even if $\sum_x q(x) < 1$.*

Proof. Let $A = \text{supp}(p) \triangleq \{x \in \mathcal{X} : p(x) > 0\}$ be the support of $p(x)$. Then,

$$-D(p\|q) = \sum_{x \in A} p(x) \log_2 \frac{q(x)}{p(x)} \leqslant \log_2 \sum_{x \in A} p(x) \frac{q(x)}{p(x)} \leqslant \log_2 \sum_{x \in \mathcal{X}} q(x) = 0.$$

The first inequality holds with equality iff $p(x) = cq(x)$ for all $x \in A$. The second inequality holds iff $\sum_{x \in A} q(x) = 1$. From these, we see that $c = 1$ and $q(x) = p(x) = 0$ for $x \in \mathcal{X} \setminus A$. □

Theorem 1.13 (Convexity of Divergence). *The divergence $D(p\|q)$ is convex in the pair (p, q). Thus, for two pairs, (p_1, q_1) and (p_2, q_2), we have*

$$D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \leqslant \lambda D(p_1\|q_1) + (1-\lambda)D(p_2\|q_2)$$

for all $\lambda \in [0,1]$.

Proof. For non-negative numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, we can form distributions $p = (\frac{a_1}{\sum_i a_i}, \ldots, \frac{a_n}{\sum_i a_i})$ and $q = (\frac{b_1}{\sum_i b_i}, \ldots, \frac{b_n}{\sum_i b_i})$ on the support $[n]$. From the non-negativity of Divergence, we get

$$0 \leqslant D(p\|q) = \sum_{i=1}^{n} \frac{a_i}{\sum_i a_i} \log_2 \frac{a_i}{b_i} - \log_2 \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i},$$

where the equality holds iff $\frac{a_i}{b_i}$ is constant for $i \in [n]$. This inequality is called the log-sum inequality

$$\sum_{i=1}^{n} a_i \log_2 \frac{a_i}{b_i} \geqslant (\sum_{i=1}^{n} a_i) \log_2 \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i},$$

where equality holds iff $\frac{a_i}{b_i}$ is constant for $i \in [n]$.

One can apply this to the LHS of (1) to derive (1). □

Lemma 1.14. *More properties of entropy and mutual information:*

1. *$I(X;Y) \geqslant 0$ with equality iff X and Y are independent.*

   Proof. First, we observe that $I(X;Y) = D(p_{X,Y}\|p_X p_Y)$. By Theorem 1.12, this divergence is zero iff $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for all x, y, but this is precisely the definition of independence. □

2. *$H(Y|X) \leqslant H(Y)$ with equality iff X and Y are independent.*

   Proof. Since $H(Y) - H(Y|X) = I(X;Y) = D(p_{X,Y}\|p_X p_Y) \geqslant 0$ iff X and Y are independent, this follow directly from the previous statement. □

3. *The entropy $H(p)$ is concave in p and the uniform distribution is the unique maximum.*

   Proof. Given $p(x)$ defined on $\mathcal{X}$, let $q(x) = 1/|\mathcal{X}|$ for all $x \in \mathcal{X}$. Then, we see that

   $$D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{1/|\mathcal{X}|} = -H(p) + \log_2 |\mathcal{X}|.$$

   Solving for $H(p)$, we see that $H(p)$ is concave in p because $D(p\|q)$ is convex in p. The uniform distribution gives the unique maximum because $D(p\|q) \geqslant 0$ with equality iff $p(x)$ is uniform. □

Remark 1.15. From $I(X;Y) = D(p_{X,Y}\|p_X p_Y)$, we see that the mutual information measures the difference between a joint distribution $p_{X,Y}$ and the product of its marginals $p_X p_Y$.