

Information Theory Hints — CMSC 858L

Here are a few facts about entropy, mutual information, and variation of information. Throughout, X is a discrete random variable (r.v.) that takes value x from the set \mathcal{X} with probability p_x . We sometimes write $p(x)$ for p_x when convenient. Y is a similar random variable over \mathcal{Y} .

Entropy

Definition. The *entropy* of a random variable X is

$$(1) \quad H(X) := - \sum_{x \in \mathcal{X}} p_x \log p_x.$$

Entropy is a function of the probability distributions. It can be thought of as a measure of the *uncertainty* in the outcome of r.v. X . When X has positive probability for only one value, then $H(X) = 0$. $H(X)$ is maximum when X has uniform probability for all its possible values. The base of the log is usually taken to be 2, and in this case, the units of entropy are called “bits.” Other bases only change the entropy by a constant scaling factor. The *joint entropy* of two random variables is just the entropy of their joint distribution, so that $H(X, Y) = - \sum_{x, y \in \mathcal{X} \times \mathcal{Y}} p_{xy} \log p_{xy}$.

Definition. The *conditional entropy* of X given Y is defined as

$$(2) \quad H(X | Y) := - \sum_{y \in \mathcal{Y}} p_y \sum_{x \in \mathcal{X}} p(x | Y = y) \log p(x | Y = y).$$

Here, $p(x | Y = y)$ is standard conditional probability. It is the probability that r.v. $X = x$ given that r.v. $Y = y$. Let X_y be a random variable that takes on value $x \in \mathcal{X}$ with probability $p(x | Y = y)$. Then we can rewrite this definition as $\mathbb{E}_{y \in Y} H(X_y)$, where $\mathbb{E}_{y \in Y}$ is the expectation over random variable Y . In other words, $H(X | Y)$ is uncertainty you expect in X if you are told the value of Y . We can rewrite the conditional entropy as $H(X|Y) = H(X, Y) - H(Y)$ as follows:

$$\begin{aligned} (3) \quad H(X, Y) &= - \sum_{x, y} p_{xy} \log p_{xy} \\ (4) \quad &= - \sum_{x, y} p_{xy} \log [p_y p(x | Y = y)] && \text{def. of cond. prob.} \\ (5) \quad &= - \sum_{x, y} (p_{xy} \log p_y) - \sum_{x, y} (p_{xy} \log p(x | Y = y)) && \text{expand } \log(ab) \\ (6) \quad &= - \sum_y (p_y \log p_y) - \sum_{x, y} (p_{xy} \log p(x | Y = y)) && \text{marginal prob.} \\ (7) \quad &= H(Y) + H(X|Y). \end{aligned}$$

Note also that $H(X, X) = H(X)$ and therefore that $H(X|X) = 0$.

Mutual Information

Definition. The *mutual information* between X and Y is

$$(8) \quad I(X; Y) = H(X) - H(X | Y).$$

This is the decrease in uncertainty in X when the value of Y is known. It is symmetric: $I(X; Y) = I(Y; X)$. Using (3), we have $I(X; Y) = H(X) + H(Y) - H(X, Y)$. Further, $I(X; X) = 0$. We can also rewrite $I(X; Y)$ directly in terms of the probability distributions:

$$(9) \quad I(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{xy} \log \frac{p_{xy}}{p_x p_y}$$

We use the convention that $0 \log(0/q) = 0$ and $p \log(p/0) = \infty$. Written as in (9), you can start to see another intuitive way of looking at $I(X; Y)$: it is a “distance” between the joint distribution $\{p_{xy}\}_{x \in \mathcal{X}, y \in \mathcal{Y}}$ and the product distribution $\{p_x p_y\}_{x \in \mathcal{X}, y \in \mathcal{Y}}$. If X and Y are independent, then $p_{xy} = p_x p_y$ and $I(X; Y) = 0$, which is intuitively right. However, mutual information is not a metric.

Variation of Information

Definition. The *variation of information* is a measure of the distance between two clusterings (partitions of elements). A clustering with clusters X_1, X_2, \dots, X_k is represented by a random variable X with $\mathcal{X} = \{1, \dots, k\}$ such that $p_i = |X_i|/n$, where $i \in \mathcal{X}$ and $n = \sum_i |X_i|$. The variation of information between two clusterings X and Y so represented is defined to be:

$$(10) \quad \text{VI}(X, Y) := H(X) + H(Y) - 2I(X; Y)$$

Intuition. The variation of information is a metric. $\text{VI}(X, Y)$ measures how much knowing the cluster assignment for an item in clustering X reduces the uncertainty about the item's cluster in clustering Y . We can rewrite the definition to make this more clear:

$$(11) \quad \text{VI}(X, Y) = H(X) + H(Y) - I(X; Y) - I(X; Y)$$

$$(12) \quad = H(X) + H(Y) - H(X) + H(X | Y) - H(Y) + H(Y | X)$$

$$(13) \quad = H(X | Y) + H(Y | X).$$

We can rewrite VI in terms of the entropies of X , Y and the joint entropy of X, Y :

$$(14) \quad \text{VI}(X, Y) = H(X) + H(Y) - 2[H(X) + H(Y) - H(X, Y)]$$

$$(15) \quad = 2H(X, Y) - H(X) - H(Y).$$

Useful Information Theory Formulas

Useful Entropy Formulas.

$$(16) \quad H(X) \geq 0$$

$$(17) \quad H(X) \geq H(X | Y) \text{ for all } Y$$

$$(18) \quad H(X_1, X_2, \dots, X_k) \leq \sum_{i=1}^k H(X_i)$$

$$(19) \quad H(X) \leq \log |\mathcal{X}|$$

$$(20) \quad H(X, X) = H(X)$$

$$(21) \quad H(X|X) = 0$$

$$(22) \quad H(X, Y) = H(Y) + H(X|Y)$$

Useful Mutual Information Formulas.

$$(23) \quad I(X; Y) = H(X) + H(Y) - H(X, Y)$$

$$(24) \quad I(X; Y) = H(Y) - H(Y|X)$$

$$(25) \quad I(X; Y) = I(Y; X)$$

$$(26) \quad I(X; Y) \geq 0$$

Useful Variation of Information Formulas.

$$(27) \quad \text{VI}(X, Y) = H(X) + H(Y) - 2I(X; Y)$$

$$(28) \quad \text{VI}(X, Y) = H(X | Y) + H(Y | X)$$

$$(29) \quad \text{VI}(X, Y) = 2H(X, Y) - H(X) - H(Y)$$