

Information theory

Siddharth Bhat

Contents

1	Preliminary definitions	5
2	Variational auto encoders	9

Chapter 1

Preliminary definitions

Definition 1. *Entropy(H): The entropy of a random variable X with probability distribution $p : X \rightarrow \mathbb{R}$ is defined as:*

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

Definition 2. *Conditional entropy($H(X|Y)$): The conditional entropy of a random variable X with respect to another variable Y is defined as:*

$$\begin{aligned} H(X|Y) &\equiv - \sum_{y \in Y} p(y) H(X|Y = y) \\ &= \sum_{y \in Y} p(y) \sum_{x \in X} -p(x|y) \log p(x|y) \\ &= \sum_{y \in Y} \sum_{x \in X} -p(y)p(x|y) \log p(x|y) \\ &= \sum_{y \in Y} \sum_{x \in X} -p(y \wedge x) \log p(x|y) \end{aligned}$$

Definition 3. *Kullback-Leibler divergence $D(X||Y)$: The Kullback-Leibler divergence of $X \sim p$ with respect to $X' \sim q$ is:*

$$D(X||X') \equiv \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

Note that $D(X||X')$ is *not symmetric*.

Intuition: extra cost of encoding X if we thought the distribution were X' .

Useful extremal case to remember: Assume X' has $q(x) = 0$ for some letter $x \in X$. In this case, $D(X||X')$ would involve a term $\frac{p(x)}{0}$, which is ∞ . This is intuitively sensible, since X' has no way to represent x , and hence X' is *infinitely far away from encoding* X . However, In this same case, one could have that X is able to encode all of X' .

Definition 4. *Mutual information: $I(X;Y)$: This is the relative entropy between the joint and product distributions.*

$$\begin{aligned} I(X;Y) &\equiv D(p(x,y)||p(x)p(y)) \equiv H(X) - H(X|Y) \\ &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

Theorem 5. *Proof of equivalence of two definitions of mutual information:*

Proof.

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

□

Some notes about mutual information:

- $I(X;Y) = I(Y;X)$. That is, I is symmetric.
- Since $I(X;Y) = H(X) - H(X|Y)$, one can view it as the *reduction in uncertainty* of X , after knowing Y . Alternatively, (to avoid double negatives), it's the *gain in certainty* of X after knowing Y . Another way of saying this is, what is the expected reduction in the number of yes/no questions to be answered to isolate the value of X on knowing the value of Y .
- $I(X;Y) = 0$ iff X, Y are independent. That is, knowing X reduces no uncertainty about Y .
- $I(X;X) = H(X)$. So, knowing X allows us to reduce our uncertainty of X by $H(X)$. ie, we completely know X , since we have *reduced our uncertainty of X* which was initially $H(X)$, by $H(X)$.

Theorem 6. *Chain rule for entropy: Let $X_1, X_2, \dots, X_n \sim p(x_1, x_2, \dots, x_n)$ Then:*

$$H(X_1, X_2, \dots, X_n) = \sum_i H(X_i | X_{i-1}, X_{i-2}, \dots, X_1)$$

Proof.

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1)$$

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3|X_1) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1)$$

induction for the rest

□

Definition 7. *Conditional mutual information: Conditional mutual information of random variables X and Y given Z is:*

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

Theorem 8. *Chain rule for information:*

$$I(X_1, X_2, \dots, X_n; Z) = \sum_i I(X_i; Z|X_1, X_2, \dots, X_{i-1})$$

Proof. TODO: finish □

Chapter 2

Variational auto encoders