Information Theory

Bruno A. Olshausen

November 16, 2010

What is information theory?

- Information theory was invented by Claude Shannon in the late 1940's. The goal of information theory is to quantify the amount of information contained in a signal, as well as the capacity of a channel or communication medium for sending information.

- Information theory is used by engineers to design and analyze the communication systems—telephone networks, modems, radio communication, etc. In neuroscience, information theory is used to quantify the amount of information conveyed by a neuron, or a population of neurons, as well as the efficiency of neural representation.

Source entropy

- The amount of potential information contained is a signal is termed the *entropy*, usually denoted by H, which is defined as follows:

$$H(X) = -\sum_X P(X) \log P(X) \tag{1}$$

- We can essentially think of this as a weighted average of $\log \frac{1}{P(X)}$. The quantity $\frac{1}{P(X)}$ expresses the amount of "surprise" in an event X—i.e., if the probability is low, then there is alot of surprise, and consequently alot of information is conveyed by telling you that X happened.

- The reason we are taking the log of the surprise is so that the total amount of surprise from independent events is additive. Logs convert multiplication to addition, so $\log \frac{1}{P(X_1\ X_2)} = \log \frac{1}{P(X_1)} + \log \frac{1}{P(X_2)}$.

- Thus, entropy essentially measures the "average surprise" or "average uncertainty" of a random variable. If the distribution $P(X)$ is highly peaked around one value, then we will rarely be surprised by this variable, hence it would be incapable of conveying much information. If on the other hand $P(X)$ is uniformly distributed, then we will be most surpised on average by this variable, hence it could potentially convey alot of information.

Mutual Information

- Mutual information quantifies the amount of information shared between two variables. It is defined as follows:

$$\begin{aligned} I(X,Y) &= H(Y) - H(Y|X) \quad (2) \\ &= H(X) - H(X|Y) \quad (3) \end{aligned}$$

- $H(Y|X)$ denotes the *conditional entropy*, which expresses the average amount of uncertainty in Y given that you know the value of X. It is defined as follows:

$$H(Y|X) = -\sum_X P(X) \sum_Y P(Y|X) \ \log P(Y|X) \quad (4)$$

- Thus, $I(X,Y)$ measures the average reduction in uncertainty in the value of Y given that you know X, or vice-versa. If the two variables have a deterministic relationship, then $I(X,Y) = H(X) = H(Y)$. Alternatively, if there is no relationship between X and Y, then $I(X,Y) = 0$.

Channel Capacity

- Channel capacity, C, measures the maximum amount of information that can be sent over a channel (e.g., a wire). It is defined as follows:

$$C = \max_{P(X)} I(X,Y) \quad (5)$$

  where X is the input to the channel and Y is the output.

- Because errors occur in transmission, the relationship between X and Y will not be completely deterministic. Mutual information measures how much information about X gets through the channel, but it will depend on the input distribution $P(X)$. Channel capacity is just the mutual information maximized with respect to the input distribution, $P(X)$.

- Shannon's noisy channel coding theorem says that you can transmit information with arbitrarily small probability of error at a rate R which is less than the channel capacity C.

- Thus, if you wish to achieve error free communication, the source entropy $H(X)$ must be less than the channel capacity C. Error free communication above the channel capacity is impossible.