Probabilistic Graphical Models

Lecture Notes
Fall 2009

September 15, 2009

Byoung-Tak Zhang
School of Computer Science and Engineering &
Cognitive Science, Brain Science, and Bioinformatics
Seoul National University
http://bi.snu.ac.kr/~btzhang/

Chapter 2. Information Theory and Bayesian Inference

2.1 Probability, Information, and Entropy

Definition. $P(x)$ is some probability of event x. Information $I(x)$ of observing the event is defined as

$$I(x) = \log_2 \frac{1}{P(x)}$$

Example. If $P(x) = 1/2$ then $I(x) = 1$ bit. If $P(x) = 1$ then $I(x) = 0$ bit.

An information source generates symbols from the set $S = \{s_1, s_2, \ldots, s_N\}$ with each symbol occurring with a fixed probability $\{P(s_1), P(s_2), \ldots, P(s_N)\}$. For such an information source the amount of information received from each symbol is

$$I(s_i) = \log_2 \frac{1}{P(s_i)}$$

The **average amount of information** received by a symbol is

$$\langle I \rangle = \sum_{i=1}^{N} P(s_i)\, I(s_i) = -\sum_{i=1}^{N} P(s_i) \log_2 P(s_i)$$

which is the definition of (information) **entropy**, $H(S)$, of the source S:

$$H(S) = -\sum_{i=1}^{N} P(s_i) \log_2 P(s_i)$$

Entropy is associated with a measure of disorder in a physical system. In an information system, entropy measures the degree of uncertainty in predicting the symbols generated by the information source. When all the symbols are equally probable ($P(s) = 1/N$), the system has the highest entropy (maximum entropy).

The **maximum entropy** occurs for a source whose symbol probabilities are all equal. To show this, consider two sources S_1 and S_2 with q symbols each. Symbol probabilities $\{P_{1i}\}$ and $\{P_{2i}\}$, $i = 1, \ldots, q$. $\sum_i P_{1i} = \sum_i P_{2i} = 1$.

The difference in entropy

$$H_1 - H_2 = -\sum_{i=1}^{q} [P_{1i} \log_2 P_{1i} - P_{2i} \log_2 P_{2i}]$$

$$\begin{aligned}
H_1 - H_2 &= -\sum_{i=1}^{q} [P_{1i} \log_2 P_{1i} + P_{1i} \log_2 P_{2i} - P_{1i} \log_2 P_{2i} - P_{2i} \log_2 P_{2i}] \\
&= -\sum_{i=1}^{q} \left[P_{1i} \log_2 \frac{P_{1i}}{P_{2i}} + (P_{1i} - P_{2i}) \log_2 P_{2i}\right] \\
&= -\sum_{i=1}^{q} P_{1i} \log_2 \frac{P_{1i}}{P_{2i}} - \sum_{i=1}^{q} (P_{1i} - P_{2i}) \log_2 P_{2i}
\end{aligned}$$

Assuming S_2 as a source with equiprobable symbols, then $H_2 = H = -\log_2 q$. Since $\log_2 P_{2i} = \log_2 \frac{1}{q}$ is independent of i, $\sum_{i=1}^{q}(P_{1i} - P_{2i}) = \sum_{i=1}^{q} P_{1i} - \sum_{i=1}^{q} P_{2i} = 1 - 1 = 0$, the second sum is zero.

$$H_1 - (-\log_2 q) = -\sum_{i=1}^{q} P_{1i} \log_2 \frac{P_{1i}}{P_{2i}}$$

or

$$H_1 - (-\log_2 q) = \sum_{i=1}^{q} P_{1i} \log_2 \frac{P_{2i}}{P_{1i}}$$

Using the inequality $\log_2 x \leq x - 1$, the right side is

$$\begin{aligned} \sum_{i=1}^{q} P_{1i} \log_2 \frac{P_{2i}}{P_{1i}} &\leq \sum_{i=1}^{q} P_{1i} \left(\frac{P_{2i}}{P_{1i}} - 1 \right) \\ &\leq \sum_{i=1}^{q} P_{2i} - \sum_{i=1}^{q} P_{1i} \\ &\leq 0 \end{aligned}$$

Then

$$H_1 - (-\log_2 q) \leq 0$$

The only way the equality can hold is if S_1 is also an equiprobabile source, so that $H_1 = -\log_2 q$. Otherwise, the entropy of S_1 is always going to be less than the source with equiprobable symbols.

2.2 Information Theory and the Brain

Information theory deals with messages, code, and the ability to transmit and receive messages accurately through noisy channels.

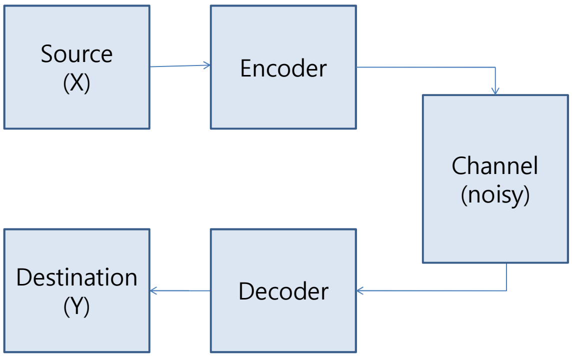

Figure 1. Information transmission from source to destination through a communication channel

Examples.

X: Images produced by the camera at KBS

Y: Images generated on TV at home

Channel: TV network (or cable TV)

X: Speech spoken by the speaker at radio station
Y: Speech heard by the listener
Channel: radio network

X: Sentences spoken by cell phone user 1 (mom)
Y: Sentences understood by cell phone user 2 (daughter)
Channel: cell phone communication network

X: Sentences spoken by my friend (Bob)
Y: Sentences understood by me (or my brain)
Channel: air + my brain

X: Sentences I heard in a scene of a Harry Porter movie (my recognition)
Y: Sentences I can remember in a week from the Harry Porter movie (my memory)
Channel: my brain

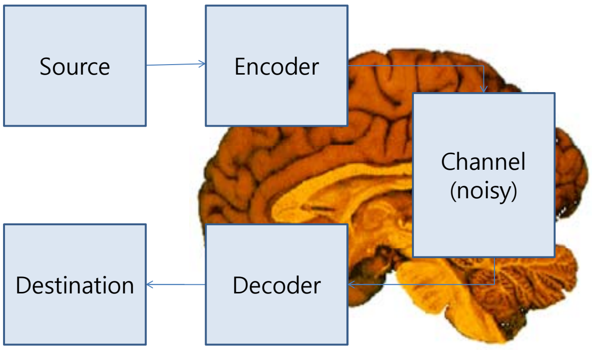

Figure 2. Brain as an information channel

X: Images of movie scenes (vision)
Y: Sentences (dialogue) of the movie scenes (language)
Channel: my brain

X: Sentences (dialogue) of the movie scenes (language)
Y: Images of movie scenes (vision, mental imagery)
Channel: my brain

A **random variable** X is a function mapping the sample space of a random process to the real numbers. For coin tossing, the sample space is {0, 1} and a random variable X can take a value of 1 (heads) or 0 (tails). The probability of the event, $P_X(x)$, is described by a probability mass function (pmf) in discrete random variables.

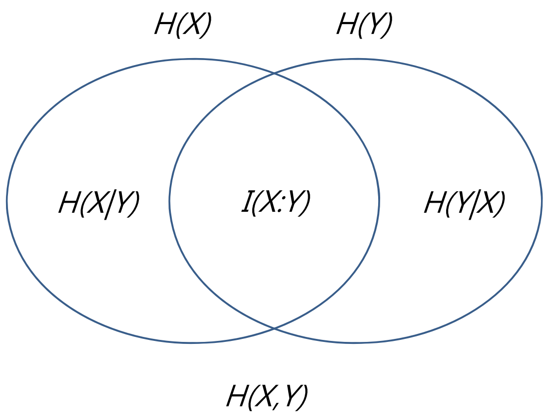

Figure 1. Joint entropy, conditional entropy, and mutual information.

Joint Entropy $H(X,Y)$

- The joint entropy measures how much entropy is contained in a joint system of *two random variables*.

$$H(X,Y) = -\sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i,y_j)\log_2 P(x_i,y_j)$$

$$H(X,Y) \le H(X) + H(Y)$$

Conditional Entropy $H(Y|X)$

- $H(Y|X)$: uncertainty about Y knowing X
- Entropy of Y given a specific $X = x_i$

$$H(Y|X = x_i) = -\sum_{j=1}^{M} P(y_j|x_i)\log_2 P(y_j|x_i)$$

- Conditional entropy is the average over *all* the possible outcomes of X

$$H(Y|X) = -\sum_{i=1}^{N} P(x_i)H(Y|X = x_i) = -\sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i)P(y_j|x_i)\log_2 P(y_j|x_i)$$

$$H(Y|X) = -\sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i,y_j)\log_2 P(y_j|x_i)$$

Relations between Joint and Conditional Entropies

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

Mutual Entropy or **Mutual Information** $I(X:Y)$

- Information *between two random variables* or *two sets of random variables* is defined as the correlation entropy or mutual entropy, also known as mutual information. For two random variables X and Y with joint entropy $H(X,Y)$, the information shared between the two is

$$I(X:Y) = H(X) + H(Y) - H(X|Y)$$

- Mutual information is the difference between the entropy of X and the conditional entropy of X given Y:

$$I(X:Y) = H(X) - H(X|Y)$$
$$I(X:Y) = H(Y) - H(Y|X)$$

- Properties

$$I(X:Y) = I(Y:X)$$
Note: $H(X|Y) \neq H(Y|X)$

- To derive the functional form of mutual information, define the *mutual probability* as

$$P(x_i:y_j) = \frac{P(x_i,y_j)}{P(x_i)P(y_j)}$$

- Then, the mutual information is given as

$$I(X:Y) = -\sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i,y_j)\log_2 P(x_i:y_j) = \sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i,y_j)\log_2 \frac{P(x_i,y_j)}{P(x_i)P(y_j)}$$

- $I(X:Y) = 0$ iff $P(x_i,y_j) = P(x_i)P(y_j)$.

2.3 Cross Entropy

Cross Entropy $H(P,Q)$

- The cross entropy for *two probability distributions* $P(X)$ and $Q(X)$ over the *same random variable* is defined as

$$H(P,Q) = -\sum_{i=1}^{N} P(x_i)\log Q(x_i)$$

- The cross entropy measures the average number of bits needed to identify an event from a set of possibilities, if a coding scheme is used based on a *given probability distribution Q*, rather than the *"true" distribution P*.

Relative Entropy (Kullback-Leibler Divergence)

- The relative entropy or KL divergence between two probability distributions $P(X)$ and $Q(X)$ that are defined over the same random variable is defined as

$$KL(P||Q) = \sum_{i=1}^{N} P(x_i)\log\frac{P(x_i)}{Q(x_i)}$$

- The relative entropy satisfies Gibb's inequality

$$KL(P||Q) \geq 0$$

with equality only if $P = Q$.

- Relation to cross entropy: Note that

$$KL(P||Q) = \sum_{i=1}^{N} P(x_i)\log\frac{P(x_i)}{Q(x_i)} = -H(P) + H(P,Q)$$

$$H(P,Q) = H(P) + KL(P||Q)$$

Minimizing the KL divergence of Q from P with respect to Q is equivalent to minimizing the cross-entropy of P and Q. This is called the *principle of minimum cross-entropy* (MCE) or Minxent.

- Relation to mutual information: Note that

$$KL(P||Q) = \sum_{i=1}^{N} P(x_i)\log\frac{P(x_i)}{Q(x_i)}$$

Substituting $P(x_i) = P(x_i,y_i)$ and $Q(x_i) = P(x_i)P(y_i)$ we get

$$KL(P||Q) = KL(P(X,Y)||P(X)P(Y)) = \sum_{i=1}^{N}\sum_{j=1}^{M} P(x_i,y_j)\log_2\frac{P(x_i,y_j)}{P(x_i)P(y_j)} = I(X:Y)$$

Mutual information is a relative entropy between joint probability $P(X,Y)$ of two random variables X and Y and the product of their marginal probabilities, $P(X)P(Y)$.

2.4 Bayesian Inference

Bayes' rule

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

$P(x)$: prior probability
$P(x|y)$: posterior probability
$P(y|x)$: likelihood
$P(y)$: evidence

Derivation of Bayes' rule

$$P(x,y) = P(x)\,P(y|x)$$
$$= P(y)\,P(x|y)$$
$$P(x|y) = \frac{P(x)\,P(y|x)}{P(y)}$$

Example*:* Use of Bayesian inference

$P(\text{disease} \mid \text{symptom})$: hard to compute (hard to know)
$P(\text{symptom} \mid \text{disease})$: easy to compute (well-known)
The hard part can be inferred from the easy part:

$$P(\text{disease} \mid \text{symptom}) = \frac{P(\text{symptom} \mid \text{disease})P(\text{disease})}{P(\text{symptom})}$$

Bayesian Inference and KL Divergence

- Bayes' theorem suggests how to update the current (prior) probability distribution for X from $P(x|I)$ to a new (posterior) probability distribution $P(x\,|y, I)$ if some new data $Y = y$ is observed:

$$P(x|y,I) = \frac{P(y|x)P(x|I)}{P(y|I)}$$

The entropy of prior distribution is

$$H(P(X|I)) = -\sum_{i=1}^{N} P(x_i|I)\log P(x_i|I)$$

The entropy of posterior distribution by observing $Y=y$ is

$$H(P(X|y,I)) = -\sum_{i=1}^{N} P(x_i|y,I)\log P(x_i|y,I)$$

The amount of information gain about X by observing $Y=y$ can be measured by the KL divergence

$$KL(P(X|y,I)||P(X|I)) = \sum_{i=1}^{N} P(x_i|y,I)\log\frac{P(x_i|y,I)}{P(x_i|I)}$$

This is the expected number of bits that would have been added to the message length if we used the original code based on $P(x_i|I)$ instead of a new code based on $P(x_i|y,I)$.