Information theory

LING 572

Fei Xia

Information theory

- Reading: M&S 2.2
- It is the use of probability theory to quantify and measure "information".
- Basic concepts:
  - Entropy
  - Cross entropy and relative entropy
  - Joint entropy and conditional entropy
  - Entropy of the language and perplexity
  - Mutual information

Entropy

- Entropy is a measure of the uncertainty associated with a distribution.

$$H(X) = -\sum_{x} p(x) \log p(x)$$

Here, X is a random variable, x is a possible outcome of X.

- The lower bound on the number of bits that it takes to transmit messages.

- An example:
  - Display the results of a 8-horse race.
  - Goal: minimize the number of bits to encode the results.

An example

- Uniform distribution: p_i=1/8.

$$H(X) = -\sum_x p(x)\log p(x) = -8*(\frac{1}{8}\log_2\frac{1}{8}) = 3\,bits$$

- Non-uniform distribution: (1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64)

$$H(X) = -(\frac{1}{2}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{4} + \frac{1}{8}\log\frac{1}{8} + \frac{1}{16}\log\frac{1}{16} + 4*\frac{1}{64}\log\frac{1}{64}) = 2\,bits$$

(0, 10, 110, 1110, 111100, 111101, 111110, 111111)

➔Uniform distribution has a higher entropy.

➔MaxEnt: make the distribution as "uniform" as possible.

Cross Entropy

- Entropy: $$H(X) = -\sum_x p(x)\log p(x)$$
- Cross Entropy: $$H_c(X) = -\sum_x p(x)\log q(x)$$

Here, p(x) is the true probability;

q(x) is our estimate of p(x).

$$H_c(X) \geq H(X)$$

Relative Entropy

- Also called Kullback-Leibler divergence:

$$KL(p \,||\, q) = \sum p(x) \log_2 \frac{p(x)}{q(x)} = H_c(X) - H(X)$$

- A “distance” measure between probability functions p and q; the closer p(x) and q(x) are, the smaller the relative entropy is.

- KL divergence is asymmetric, so it is not a proper distance metric: $KL(p,q) \neq KL(q,p)$

Joint and conditional entropy

- Joint entropy:

$$H(X,Y) = -\sum_{x}\sum_{y} p(x,y)\log p(x,y)$$

- Conditional entropy:

$$H(Y \mid X) = H(X,Y) - H(X)$$

Entropy of a language (per-word entropy)

- The entropy of a language L:

$$H(L,p) = -\lim_{n\to\infty} \frac{\sum_{x_{1n}} p(x_{1n}) \log p(x_{1n})}{n}$$

- If we make certain assumptions that the language is "nice", then the cross entropy can be calculated as: (Shannon-Breiman-Mcmillan Theorem)

$$H(L,p) = -\lim_{n\to\infty} \frac{\log p(x_{1n})}{n} \approx -\frac{\log p(x_{1n})}{n}$$

Per-word entropy (cont)

- $p(x_{1n})$ can be calculated by n-gram models

- Ex: unigram model

$$p(x_{1n}) = \prod_i p(x_i)$$

$$log\ p(x_{1n}) = \sum_i log\ p(x_i)$$

Perplexity

- Perplexity $PP(x_{1n})$ is $2^{H(L,p)}$.
- Perplexity is the weighted average number of choices a random variable has to make.
- Perplexity is often used to evaluate a language model; lower perplexity is preferred.

Mutual information

- It measures how much is in common between X and Y:

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

- I(X;Y) = KL(p(x,y) || p(x)p(y))

- If X and Y are independent, I(X;Y) is 0.

Summary on Information theory

- Reading: M&S 2.2

- It is the use of probability theory to quantify and measure "information".

- Basic concepts:
  - Entropy
  - Cross entropy and relative entropy
  - Joint entropy and conditional entropy
  - Entropy of the language and perplexity
  - Mutual information

Additional slides

Conditional entropy

$$
\begin{aligned}
H(Y \mid X) \\
&= \sum_x p(x) H(Y \mid X = x) \\
&= -\sum_x p(x) \sum_y p(y \mid x) \log p(y \mid x) \\
&= -\sum_x \sum_y p(x, y) \log p(y \mid x) \\
&= -\sum_x \sum_y p(x, y) \log p(x, y) / p(x) \\
&= -\sum_x \sum_y p(x, y) (\log p(x, y) - \log p(x)) \\
&= -\sum_x \sum_y p(x, y) \log p(x, y) + \sum_x \sum_y p(x, y) \log p(x) \\
&= \sum_x \sum_y p(x, y) \log p(x, y) + \sum_x p(x) \log p(x) \\
&= H(X, Y) - H(X)
\end{aligned}
$$

Mutual information

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