Information Theory

Pedro Larrañaga, Concha Bielza

Computational Intelligence Group
Artificial Intelligence Department
Universidad Politécnica de Madrid

Machine Learning

Master in Data Science + Master HMDA

Outline

Outline

Entropy

Entropy (Shannon, 1948) quantifies the uncertainty when predicting the value of a random variable

Entropy of a discrete random variable

The entropy of a discrete random variable X, with sample space $\Omega_X = \{x_1, ..., x_n\}$ and pdf given by $p(x)$, is denoted $\mathbb{H}(X)$ and defined as

$$\mathbb{H}(X) = -\sum_{i=1}^{n} p(X = x_i) \log_2 p(X = x_i)$$

- The entropy of a discrete random variable, X, verifies $0 \leq \mathbb{H}(X) \leq \log_2 n$
- The upper bound is calculated from the uniform distribution
- The choice of logarithmic base in the above formula determines the unit of information entropy
  - The most common unit is the bit, which is based on binary logarithms
  - If e is the base, the unit is called nat
  - For decimal logarithms, that is, base 10, the unit is called Hartley

Entropy

Entropy of a Bernouilli distribution

For a Bernoulli distribution with parameter $\theta = p(X = 1)$, the entropy is:

$$\mathbb{H}(X) = -\theta \log_2 \theta - (1 - \theta) \log_2 (1 - \theta)$$

The maximum value of this expression is $\log_2 2 = 1$ and is achieved at point $\theta = 0.5$

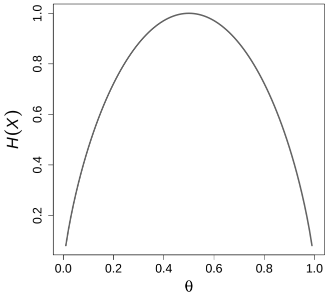

Figure: Entropy function $\mathbb{H}$ of a Bernoulli distribution with parameter θ

Entropy

Entropy of a continuous random variable

The entropy of a continuous random variable X, with pdf given by $f(x)$, is called differential entropy and is defined as

$$h(X) = -\int_{\Omega_X} f(x) \ln f(x) \, dx$$

- For a Gaussian variable $X \sim \mathcal{N}(x|\mu, \sigma)$, the differential entropy is $h(X) = \ln(\sigma\sqrt{2\pi e})$
- This verifies that it has the largest entropy of all random variables of equal variance (Cover and Thomas, 1991)

Outline

Joint entropy

Joint entropy

Given a bidimensional discrete random variable (X, Y), with a bivariate probability mass function $p(x, y)$ where $x \in \Omega_X = \{x_1, ..., x_n\}$ and $y \in \Omega_Y = \{y_1, ..., y_m\}$, the joint entropy is defined by

$$\mathbb{H}(X, Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i, y_j)$$

Synapse type	I	II-III	Layer IV	V	VI
as	0.0595	0.1691	0.1848	0.1428	0.2772
ss	0.0105	0.0209	0.0552	0.0272	0.0528

$$\begin{aligned}
\mathbb{H}(X, Y) = & -(0.0595 \log_2 0.0595 + 0.1691 \log_2 0.1691 + 0.1848 \log_2 0.1848 \\
& + 0.1428 \log_2 0.1428 + 0.2772 \log_2 0.2772 + 0.0105 \log_2 0.0105 + 0.0209 \log_2 0.0209 \\
& + 0.0552 \log_2 0.0552 + 0.0272 \log_2 0.0272 + 0.0528 \log_2 0.0528) \\
& \simeq 2.8219
\end{aligned}$$

Outline

Conditional entropy

Conditional entropy

The conditional entropy of X given Y is defined as $\mathbb{H}(X|Y) = \sum_{j=1}^{m} p(y_j)\mathbb{H}(X|Y = y_j)$ where $\mathbb{H}(X|Y = y_j) = -\sum_{i=1}^{n} p(x_i|y_j) \log_2 p(x_i|y_j)$ is the entropy of X given that $Y = y_j$. After some algebraic manipulations:

$$\mathbb{H}(X|Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i|y_j)$$

Total entropies law

The total entropies law expresses the joint entropy of two variables in terms of the entropy of one of the variables and the conditional entropy of the other variable given the first variable

$$\mathbb{H}(X, Y) = \mathbb{H}(X) + \mathbb{H}(Y|X) = \mathbb{H}(Y) + \mathbb{H}(X|Y)$$

If X and Y are independent variables:

1. $\mathbb{H}(X|Y) = \mathbb{H}(X)$
2. $\mathbb{H}(Y|X) = \mathbb{H}(Y)$
3. $\mathbb{H}(X, Y) = \mathbb{H}(X) + \mathbb{H}(Y)$

Outline

Mutual information

Mutual information

The mutual information $\mathbb{I}(X, Y)$ between two variables X and Y is defined as

$$\mathbb{I}(X, Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X)$$

Replacing entropy and conditional entropy by their respective expressions:

$$\mathbb{I}(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

- Mutual information is interpreted as the reduction in uncertainty about X after observing Y, or, by symmetry, the reduction in uncertainty about Y after observing X
- It holds that $\mathbb{I}(X, Y) \geq 0$
- If X and Y are independent $\Rightarrow \mathbb{I}(X, Y) = 0$

Mutual information

Conditional mutual information

Given three random variables, X, Y and Z with $x \in \Omega_X = \{x_1, ..., x_n\}$, $y \in \Omega_Y = \{y_1, ..., y_m\}$, and $z \in \Omega_Z = \{z_1, ..., z_r\}$, the conditional mutual information of X and Y given Z, $\mathbb{I}(X, Y|Z)$, is defined as

$$\mathbb{I}(X, Y|Z) = \sum_{k=1}^{r} p(z_k)\mathbb{I}(X, Y|Z = z_k)$$

After some algebraic manipulations:

$$\mathbb{I}(X, Y|Z) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r} p(x_i, y_j, z_k) \log_2 \frac{p(x_i, y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)}$$

The conditional mutual information can be expressed in terms of conditional entropies:

$$\mathbb{I}(X, Y|Z) = \mathbb{H}(X|Z) + \mathbb{H}(Y|Z) - \mathbb{H}(X, Y|Z)$$

Outline

Kullback-Leibler divergence

Kullback-Leibler divergence

The Kullback-Leibler divergence (Kullback and Leibler, 1951) is a way of comparing two probability distributions, $p(X)$ and $q(X)$, defined over the same sample space, $\{x_1, ..., x_n\}$. One of the two distributions, $p(X)$, plays the role of a "true" distribution, whereas the other distribution, $q(X)$, is an arbitrary probability distribution

$$\mathbb{KL}(p||q) = \sum_{i=1}^{n} p(x_i) \log_2 \frac{p(x_i)}{q(x_i)}$$

1. The Kullback-Leibler divergence is not a true distance (in the mathematical sense of the term) since it is not symmetric and does not verify the triangle inequality
2. $\mathbb{KL}(p||q) \geq 0$
3. $\mathbb{I}(X, Y) = \mathbb{KL}(p(X, Y)||p(X)p(Y))$

References

- T.M. Cover, J.A. Thomas (1991). *Elements of Information Theory*. Wiley
- S. Kullback and R.A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1), 79-86
- C.E. Shannon (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 3, 379-423