Information theory

Course of Machine Learning
Master Degree in Computer Science
University of Rome ``Tor Vergata''

Giorgio Gambosi

a.a. 2017-2018

Information

Let X be a discrete random variable:

- define a measure $h(x)$ of the information (surprise) of observing $X = x$
- requirements:
    - likely events provide low surprise, while rare events provide high surprise: $h(x)$ is inversely proportional to $p(x)$
    - X, Y independent: the event $X = x, Y = y$ has probability $p(x)p(y)$. Its surprise is the sum of the surprise for $X = x$ and for $Y = y$, that is, $h(x, y) = h(x) + h(y)$ (information is additive)

this results into $h(x) = -\log x$ (usually base 2)

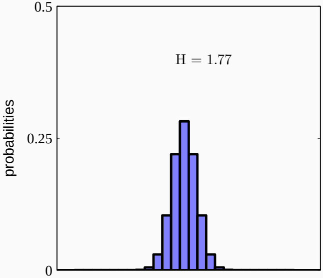
Entropy

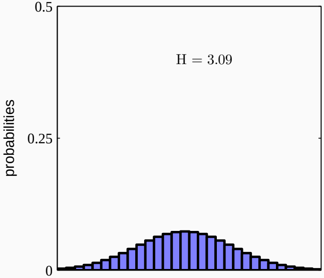
A sender transmits the value of X to a receiver: the expected amount of information transmitted (w.r.t. $p(x)$) is the entropy of X

$$H(x) = -\sum_x p(x) \log_2 p(x)$$

- lower entropy results from more sharply peaked distributions
- the uniform distribution provides the highest entropy

Entropy is a measure of disorder.





- $p(x) \in [0,1]$ implies $p(x) \log_2 p(x) \leq 0$ and $H(X) \geq 0$
- $H(X) = 0$ if there exists x such that $p(x) = 1$

Maximum entropy

Given a fixed number k of outcomes, the distribution $p_1, \ldots, p_k$ with maximum entropy is derived by maximizing $H(X)$ under the constraint $\sum_{i=1}^{k} p_i = 1$. By using Lagrange multipliers, this amounts to maximizing

$$-\sum_{i=1}^{k} p_i \log_2 p_i + \lambda \left(\sum_{i=1}^{k} p_i - 1 \right)$$

Setting the derivative of each p_i to 0,

$$0 = -\log_2 p_i - \log_2 e + \lambda$$

results into $p_i = 2^{\lambda} - e$ for each i, that is into the uniform distribution $p_i = \dfrac{1}{k}$ and $H(X) = \log_2 k$

Entropy, some properties

$H(X)$ is a lower bound on the expected number of bits needed to encode the values of X

- trivial approach: code of length $\log_2 k$ (assuming uniform distribution of values for X)
- for non-uniform distributions, better coding schemes by associating shorter codes to likely values of X

Entropy, continuous case

Differential entropy

X is a continuous r.v.: divide the domain in bins of width Δ. Then, for each bin, there exists x_i such that

$$\int_{i\Delta}^{(i+1)\Delta} p(x)dx = p(x_i)\Delta$$

The probability of a point in the i-th bin is then $p(x_i)\Delta$, and

$$H_\Delta = -\sum_i p(x_i)\Delta \ln(p(x_i)\Delta) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta$$

The differential entropy is defined as

$$H(X) = \lim_{\Delta \to 0} -\sum_i p(x_i)\Delta \ln p(x_i) = -\int p(x) \ln p(x) dx$$

Maximum differential entropy

Let X be a continuous r.v. with given mean μ and variance σ^2.

- The distribution of X with maximum entropy is the gaussian distribution $\mathcal{N}(\mu, \sigma^2)$.

Conditional entropy

Let X, Y be a continuous r.v. : for a pair of values x, y the additional information needed to specify y if x is known is $-\ln p(y|x)$.

The expected additional information needed to specify the value of Y if we assume the value of X is known is the conditional entropy of Y given X

$$H(Y|X) = -\int\int p(x,y) \ln p(y|x) dx dy$$

Clearly, since $\ln p(y|x) = \ln p(x,y) - \ln p(x)$

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

that is, the information needed to describe (on the average) the values of X and Y is the sum of the information needed to describe the value of X plus that needed to describe the value of Y is X is known.

KL divergence

Assume the distribution $p(x)$ of X is unknown, and we have modeled is as an approximation $q(x)$.

If we use $q(x)$ to encode values of X we need an average length $-\int p(x)\ln q(x)dx$, while the minimum (known $p(x)$) is $-\int p(x)\ln p(x)dx$.

The additional amount of information needed, due to the approximation of $p(x)$ through $q(x)$ is the Kullback-Leibler divergence

$$\begin{aligned} KL(p||q) &= -\int p(x)\ln q(x)dx + \int p(x)\ln p(x)dx \\ &= -\int p(x)\ln\frac{q(x)}{p(x)}dx \end{aligned}$$

$KL(p||q)$ measures the difference between the distributions p and q.

- $KL(p||p) = 0$
- $KL(p||q) \neq KL(q||p)$: the function is not symmetric, it is not a distance (it would be $d(x,y) = d(y,x)$)

Applying KL divergence

- $\mathbf{x} = (x_1, \ldots, x_n)$, dataset generated by a unknown distribution $p(x)$
- we want to infer the parameters of a probabilistic model $q_\theta(x|\theta)$
- approach: minimize

$$\begin{aligned} KL(p||q_\theta) &= -\int p(x) \ln \frac{q(x|\theta)}{p(x)} dx \\ &\approx -\frac{1}{n} \sum_{i=1}^{n} \ln \frac{q(x_i|\theta)}{p(x_i)} \\ &= \frac{1}{n} \sum_{i=1}^{n} (\ln p(x_i) - \ln q(x_i|\theta)) \end{aligned}$$

First term is independent of θ, while the second one is the negative log-likelihood of $\mathbf{x}$. The value of θ which minimizes $KL(p||q_\theta)$ also maximizes the log-likelihood.

Mutual information

- Measure of the independence between X and Y

$$I(X,Y) = KL(p(X,Y)||p(X),p(Y)) = -\int\int p(x,y)\ln\frac{p(x)p(y)}{p(x,y)}dxdy$$

additional encoding length if independence is assumed

- We have:

$$\begin{aligned}
I(X,Y) &= -\int\int p(x,y)\ln\frac{p(x)p(y)}{p(x,y)}dxdy \\
&= -\int\int p(x,y)\ln\frac{p(x)p(y)}{p(x|y)p(y)}dxdy \\
&= -\int\int p(x,y)\ln\frac{p(x)}{p(x|y)}dxdy \\
&= -\int\int p(x,y)\ln p(x)dxdy + \int\int p(x,y)\ln p(x|y)dxdy \\
&= H(X) - H(X|Y)
\end{aligned}$$

- Similarly, it derives $I(X,Y) = H(Y) - H(Y|X)$