

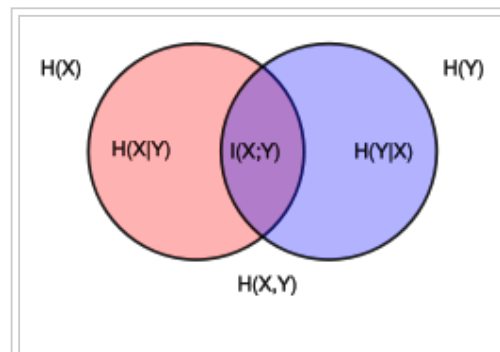
# Mutual information

From Wikipedia, the free encyclopedia

In probability theory and information theory, the **mutual information** (MI) or (formerly) **transinformation** of two random variables is a measure of the variables' mutual dependence. Not limited to real-valued random variables like the correlation coefficient, MI is more general and determines how similar the joint distribution  $p(X,Y)$  is to the products of factored marginal distribution  $p(X)p(Y)$ . MI is the expected value of the pointwise mutual information (PMI). The most common unit of measurement of mutual information is the bit.

## Contents

- 1 Definition of mutual information
- 2 Relation to other quantities
- 3 Variations of mutual information
  - 3.1 Metric
  - 3.2 Conditional mutual information
  - 3.3 Multivariate mutual information
    - 3.3.1 Applications
  - 3.4 Normalized variants
  - 3.5 Weighted variants
  - 3.6 Adjusted mutual information
  - 3.7 Absolute mutual information
  - 3.8 Mutual information for discrete data
- 4 Applications of mutual information
- 5 See also
- 6 Notes
- 7 References



Individual ( $H(X), H(Y)$ ), joint ( $H(X,Y)$ ), and conditional entropies for a pair of correlated subsystems  $X, Y$  with mutual information  $I(X; Y)$ .

## Definition of mutual information

Formally, the mutual information of two discrete random variables  $X$  and  $Y$  can be defined as:

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left( \frac{p(x, y)}{p(x) p(y)} \right),$$

where  $p(x,y)$  is the joint probability distribution function of  $X$  and  $Y$ , and  $p(x)$  and  $p(y)$  are the marginal probability distribution functions of  $X$  and  $Y$  respectively.

In the case of continuous random variables, the summation is replaced by a definite double integral:

$$I(X; Y) = \int_Y \int_X p(x, y) \log \left( \frac{p(x, y)}{p(x) p(y)} \right) dx dy,$$

where  $p(x,y)$  is now the joint probability *density* function of  $X$  and  $Y$ , and  $p(x)$  and  $p(y)$  are the marginal probability density functions of  $X$  and  $Y$  respectively.

If the log base 2 is used, the units of mutual information are the bit.

Intuitively, mutual information measures the information that  $X$  and  $Y$  share: it measures how much knowing one of these variables reduces uncertainty about the other. For example, if  $X$  and  $Y$  are independent, then knowing  $X$  does not give any information about  $Y$  and vice versa, so their mutual information is zero. At the other extreme, if  $X$  is a deterministic function of  $Y$  and  $Y$  is a deterministic function of  $X$  then all information conveyed by  $X$  is shared with  $Y$ : knowing  $X$  determines the value of  $Y$  and vice versa. As a result, in this case the mutual information is the same as the uncertainty contained in  $Y$  (or  $X$ ) alone, namely the entropy of  $Y$  (or  $X$ ). Moreover, this mutual information is the same as the entropy of  $X$  and as the entropy of  $Y$ . (A very special case of this is when  $X$  and  $Y$  are the same random variable.)

Mutual information is a measure of the inherent dependence expressed in the joint distribution of  $X$  and  $Y$  relative to the joint distribution of  $X$  and  $Y$  under the assumption of independence. Mutual information therefore measures dependence in the following sense:  $I(X; Y) = 0$  if and only if  $X$  and  $Y$  are independent random variables. This is easy to see in one direction: if  $X$  and  $Y$  are independent, then  $p(x, y) = p(x)p(y)$ , and therefore:

$$\log \left( \frac{p(x, y)}{p(x)p(y)} \right) = \log 1 = 0.$$

Moreover, mutual information is nonnegative (i.e.  $I(X; Y) \geq 0$ ; see below) and symmetric (i.e.  $I(X; Y) = I(Y; X)$ ).

## Relation to other quantities

Mutual information can be equivalently expressed as

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X, Y) \\ &= H(X, Y) - H(X|Y) - H(Y|X) \end{aligned}$$

where  $H(X)$  and  $H(Y)$  are the marginal entropies,  $H(X|Y)$  and  $H(Y|X)$  are the conditional entropies, and  $H(X, Y)$  is the joint entropy of  $X$  and  $Y$ . Note the analogy to the union, difference, and intersection of two sets, as illustrated in the Venn diagram.

Using Jensen's inequality on the definition of mutual information we can show that  $I(X; Y)$  is non-negative, consequently,  $H(X) \geq H(X|Y)$ . Here we give the detailed deduction of  $I(X; Y) = H(Y) - H(Y|X)$ :

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

The proofs of the other identities above are similar.

Intuitively, if entropy  $H(Y)$  is regarded as a measure of uncertainty about a random variable, then  $H(Y|X)$  is a measure of what  $X$  does *not* say about  $Y$ . This is "the amount of uncertainty remaining about  $Y$  after  $X$  is known", and thus the right side of the first of these equalities can be read as "the amount of uncertainty in  $Y$ , minus the amount of uncertainty in  $Y$  which remains after  $X$  is known", which is equivalent to "the amount of uncertainty in  $Y$  which is removed by knowing  $X$ ". This corroborates the intuitive meaning of mutual information as the amount of information (that is, reduction in uncertainty) that knowing either variable provides about the other.

Note that in the discrete case  $H(X|X) = 0$  and therefore  $H(X) = I(X;X)$ . Thus  $I(X;X) \geq I(X;Y)$ , and one can formulate the basic principle that a variable contains at least as much information about itself as any other variable can provide.

Mutual information can also be expressed as a Kullback–Leibler divergence, of the product  $p(x) \times p(y)$  of the marginal distributions of the two random variables  $X$  and  $Y$ , from  $p(x,y)$  the random variables' joint distribution:

$$I(X; Y) = D_{\text{KL}}(p(x, y) \| p(x)p(y)).$$

Furthermore, let  $p(x|y) = p(x, y) / p(y)$ . Then

$$\begin{aligned} I(X; Y) &= \sum_y p(y) \sum_x p(x|y) \log_2 \frac{p(x|y)}{p(x)} \\ &= \sum_y p(y) D_{\text{KL}}(p(x|y) \| p(x)) \\ &= \mathbb{E}_Y \{ D_{\text{KL}}(p(x|y) \| p(x)) \}. \end{aligned}$$

Note that here, the Kullback-Leibler divergence involves integration with respect to the random variable  $X$  only, and the expression  $D_{\text{KL}}(p(x|y) \| p(x))$  is now a random variable in  $Y$ . Thus mutual information can also be understood as the expectation of the Kullback–Leibler divergence of the univariate distribution  $p(x)$  of  $X$  from the conditional distribution  $p(x|y)$  of  $X$  given  $Y$ : the more different the distributions  $p(x|y)$  and  $p(x)$  are on average, the greater the information gain.

## Variations of mutual information

Several variations on mutual information have been proposed to suit various needs. Among these are normalized variants and generalizations to more than two variables.

### Metric

Many applications require a metric, that is, a distance measure between pairs of points. The quantity

$$d(X, Y) = H(X, Y) - I(X; Y) = H(X) + H(Y) - 2I(X; Y) = H(X|Y) + H(Y|X)$$

satisfies the properties of a metric (triangle inequality, non-negativity, indiscernability and symmetry). This distance metric is also known as the Variation of information.

If  $X, Y$  are discrete random variables then all the entropy terms are non-negative, so  $0 \leq d(X, Y) \leq H(X, Y)$  and one can define a normalized distance

$$D(X, Y) = d(X, Y) / H(X, Y) \leq 1.$$

The metric  $D$  is a universal metric, in that if any other distance measure places  $X$  and  $Y$  close-by, then the  $D$  will also judge them close.<sup>[1]</sup>

A set-theoretic interpretation of information (see the figure for Conditional entropy) shows that

$$D(X, Y) = 1 - I(X; Y) / H(X, Y)$$

which is effectively the Jaccard distance between  $X$  and  $Y$ .

Finally,

$$D'(X, Y) = 1 - \frac{I(X; Y)}{\max(H(X), H(Y))}$$

is also a metric.

## Conditional mutual information

Sometimes it is useful to express the mutual information of two random variables conditioned on a third.

$$I(X; Y|Z) = \mathbb{E}_Z(I(X; Y)|Z) = \sum_{z \in Z} \sum_{y \in Y} \sum_{x \in X} p_Z(z) p_{X,Y|Z}(x, y|z) \log \frac{p_{X,Y|Z}(x, y|z)}{p_{X|Z}(x|z) p_{Y|Z}(y|z)},$$

which can be simplified as

$$I(X; Y|Z) = \sum_{z \in Z} \sum_{y \in Y} \sum_{x \in X} p_{X,Y,Z}(x, y, z) \log \frac{p_Z(z) p_{X,Y,Z}(x, y, z)}{p_{X,Z}(x, z) p_{Y,Z}(y, z)}.$$

Conditioning on a third random variable may either increase or decrease the mutual information, but it is always true that

$$I(X; Y|Z) \geq 0$$

for discrete, jointly distributed random variables  $X, Y, Z$ . This result has been used as a basic building block for proving other inequalities in information theory.

## Multivariate mutual information

Several generalizations of mutual information to more than two random variables have been proposed, such as total correlation and interaction information. If Shannon entropy is viewed as a signed measure in the context of information diagrams, as explained in the article *Information theory and measure theory*, then the only definition of multivariate mutual information that makes sense is as follows:

$$I(X_1; X_1) = H(X_1)$$

and for  $n > 1$ ,

$$I(X_1; \dots; X_n) = I(X_1; \dots; X_{n-1}) - I(X_1; \dots; X_{n-1}|X_n),$$

where (as above) we define

$$I(X_1; \dots; X_{n-1}|X_n) = \mathbb{E}_{X_n}(I(X_1; \dots; X_{n-1})|X_n).$$

(This definition of multivariate mutual information is identical to that of interaction information except for a change in sign when the number of random variables is odd.)

If  $A$  and  $B$  are two sets of variables, then the mutual information between them is:

$$I(A, B) = H(A \cup B) + H(A \cap B) - H(A) - H(B),$$

## Applications

Applying information diagrams blindly to derive the above definition has been criticised, and indeed it has found rather limited practical application, since it is difficult to visualize or grasp the significance of this quantity for a large number of random variables. It can be zero, positive, or negative for any  $n \geq 3$ .

One high-dimensional generalization scheme which maximizes the mutual information between the joint distribution and other target variables is found to be useful in feature selection.<sup>[2]</sup>

Mutual information is also used in the area of signal processing as a measure of similarity between two signals. For example, FMI metric<sup>[3]</sup> is an image fusion performance measure that makes use of mutual information in order to measure the amount of information that the fused image contains about the source images. The Matlab code for this metric can be found at.<sup>[4]</sup>

## Normalized variants

Normalized variants of the mutual information are provided by the *coefficients of constraint*,<sup>[5]</sup> uncertainty coefficient<sup>[6]</sup> or proficiency:<sup>[7]</sup>

$$C_{XY} = \frac{I(X; Y)}{H(Y)} \quad \text{and} \quad C_{YX} = \frac{I(X; Y)}{H(X)}.$$

The two coefficients are not necessarily equal. In some cases a symmetric measure may be desired, such as the following *redundancy* measure:

$$R = \frac{I(X; Y)}{H(X) + H(Y)}$$

which attains a minimum of zero when the variables are independent and a maximum value of

$$R_{\max} = \frac{\min(H(X), H(Y))}{H(X) + H(Y)}$$

when one variable becomes completely redundant with the knowledge of the other. See also *Redundancy (information theory)*. Another symmetrical measure is the *symmetric uncertainty* (Witten & Frank 2005), given by

$$U(X, Y) = 2R = 2 \frac{I(X; Y)}{H(X) + H(Y)}$$

which represents a weighted average of the two uncertainty coefficients.<sup>[6]</sup>

If we consider mutual information as a special case of the total correlation or dual total correlation, the normalized version are respectively,

$$\frac{I(X; Y)}{\min[H(X), H(Y)]} \text{ and } \frac{I(X; Y)}{H(X, Y)}.$$

Finally there's a normalization<sup>[8]</sup> which derives from first thinking of mutual information as an analogue to covariance (thus Shannon entropy is analogous to variance). Then the normalized mutual information is calculated akin to the Pearson correlation coefficient,

$$\frac{I(X; Y)}{\sqrt{H(X)H(Y)}}.$$

## Weighted variants

In the traditional formulation of the mutual information,

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x) p(y)},$$

each *event* or *object* specified by  $(x, y)$  is weighted by the corresponding probability  $p(x, y)$ . This assumes that all objects or events are equivalent *apart from* their probability of occurrence. However, in some applications it may be the case that certain objects or events are more *significant* than others, or that certain patterns of association are more semantically important than others.

For example, the deterministic mapping  $\{(1, 1), (2, 2), (3, 3)\}$  may be viewed as stronger than the deterministic mapping  $\{(1, 3), (2, 1), (3, 2)\}$ , although these relationships would yield the same mutual information. This is because the mutual information is not sensitive at all to any inherent ordering in the variable values (Cronbach 1954, Coombs & Dawes 1970, Lockhead 1970), and is therefore not sensitive at all to the **form** of the relational mapping between the associated variables. If it is desired that the former relation — showing agreement on all variable values — be judged stronger than the later relation, then it is possible to use the following *weighted mutual information* (Guiasu 1977)

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} w(x, y) p(x, y) \log \frac{p(x, y)}{p(x) p(y)},$$

which places a weight  $w(x, y)$  on the probability of each variable value co-occurrence,  $p(x, y)$ . This allows that certain probabilities may carry more or less significance than others, thereby allowing the quantification of relevant *holistic* or *prägnanz* factors. In the above example, using larger relative weights for  $w(1, 1)$ ,  $w(2, 2)$ , and  $w(3, 3)$  would have the effect of assessing greater *informativeness* for the relation  $\{(1, 1), (2, 2), (3, 3)\}$  than for the relation  $\{(1, 3), (2, 1), (3, 2)\}$ , which may be desirable in some cases of pattern recognition, and the like. This weighted mutual information is a form of weighted KL-Divergence, which is known to take negative values for some inputs,<sup>[9]</sup> and there are examples where the weighted mutual information also takes negative values.<sup>[10]</sup>

## Adjusted mutual information

A probability distribution can be viewed as a partition of a set. One may then ask: if a set were partitioned randomly, what would the distribution of probabilities be? What would the expectation value of the mutual information be? The adjusted mutual information or AMI subtracts the expectation value of the MI, so that the AMI is zero when two different distributions are random, and one when two distributions are identical. The AMI is defined in analogy to the adjusted Rand index of two different partitions of a set.

## Absolute mutual information

Using the ideas of Kolmogorov complexity, one can consider the mutual information of two sequences independent of any probability distribution:

$$I_K(X; Y) = K(X) - K(X|Y).$$

To establish that this quantity is symmetric up to a logarithmic factor ( $I_K(X; Y) \approx I_K(Y; X)$ ) requires the chain rule for Kolmogorov complexity (Li 1997). Approximations of this quantity via compression can be used to define a distance measure to perform a hierarchical clustering of sequences without having any domain knowledge of the sequences (Cilibiasi 2005).

## Mutual information for discrete data

When  $X$  and  $Y$  are limited to be in a discrete number of states, observation data is summarized in a contingency table, with row variable  $X$  (or  $i$ ) and column variable  $Y$  (or  $j$ ). Mutual information is one of the measures of association or correlation between the row and column variables. Other measures of association include Pearson's chi-squared test statistics, G-test statistics, etc. In fact, mutual information is equal to G-test statistics divided by  $2N$  where  $N$  is the sample size.

In the special case where the number of states for both row and column variables is 2 ( $i,j=1,2$ ), the degrees of freedom of the Pearson's chi-squared test is 1. Out of the four terms in the summation:

$$\sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_i p_j}$$

only one is independent. It is the reason that mutual information function has an exact relationship with the correlation function  $P_{X=1,Y=1} - P_{X=1}P_{Y=1}$  for binary sequences.<sup>[11]</sup>

## Applications of mutual information

In many applications, one wants to maximize mutual information (thus increasing dependencies), which is often equivalent to minimizing conditional entropy. Examples include:

- In search engine technology, mutual information between phrases and contexts is used as a feature for k-means clustering to discover semantic clusters (concepts).<sup>[12]</sup>
- In telecommunications, the channel capacity is equal to the mutual information, maximized over all input distributions.
- Discriminative training procedures for hidden Markov models have been proposed based on the maximum mutual information (MMI) criterion.
- RNA secondary structure prediction from a multiple sequence alignment.
- Phylogenetic profiling prediction from pairwise present and disappearance of functionally link genes.
- Mutual information has been used as a criterion for feature selection and feature transformations in machine learning. It can be used to characterize both the relevance and redundancy of variables, such as the minimum redundancy feature selection.
- Mutual information is used in determining the similarity of two different clusterings of a dataset. As such, it provides some advantages over the traditional Rand index.
- Mutual information of words is often used as a significance function for the computation of collocations in corpus linguistics. This has the added complexity that no word-instance is an instance to two different words; rather, one counts instances where 2 words occur adjacent or in close proximity; this slightly complicates the calculation, since the expected probability of one word occurring within N words of another, goes up with N.
- Mutual information is used in medical imaging for image registration. Given a reference image (for example, a brain scan), and a second image which needs to be put into the same coordinate system as the reference image, this image is deformed until the mutual information between it and the reference image is maximized.
- Detection of phase synchronization in time series analysis
- In the infomax method for neural-net and other machine learning, including the infomax-based Independent component analysis algorithm
- Average mutual information in delay embedding theorem is used for determining the *embedding delay* parameter.
- Mutual information between genes in expression microarray data is used by the ARACNE algorithm for reconstruction of gene networks.
- In statistical mechanics, Loschmidt's paradox may be expressed in terms of mutual information.<sup>[13][14]</sup> Loschmidt noted that it must be impossible to determine a physical law which lacks time reversal symmetry (e.g. the second law of thermodynamics) only from physical laws which have this symmetry. He pointed out that the H-theorem of Boltzmann made the assumption that the velocities of particles in a gas were permanently uncorrelated, which removed the time symmetry inherent in the H-theorem. It can be shown that if a system is described by a probability density in phase space, then Liouville's theorem implies that the joint information

(negative of the joint entropy) of the distribution remains constant in time. The joint information is equal to the mutual information plus the sum of all the marginal information (negative of the marginal entropies) for each particle coordinate. Boltzmann's assumption amounts to ignoring the mutual information in the calculation of entropy, which yields the thermodynamic entropy (divided by Boltzmann's constant).

- The mutual information is used to learn the structure of Bayesian networks/dynamic Bayesian networks, which explain the causal relationship between random variables, as exemplified by the GlobalMIT toolkit [2] (<http://code.google.com/p/globalmit/>): learning the globally optimal dynamic Bayesian network with the Mutual Information Test criterion.
- Popular cost function in Decision tree learning.

## See also

- Pointwise mutual information
- Quantum mutual information

## Notes

1. Alexander Kraskov, Harald Stögbauer, Ralph G. Andrzejak, and Peter Grassberger, "Hierarchical Clustering Based on Mutual Information", (2003) *ArXiv q-bio/0311039* (<http://arxiv.org/abs/q-bio/0311039>)
2. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze (2008). *An Introduction to Information Retrieval*. Cambridge University Press. ISBN 0-521-86571-9.
3. Haghighat, M. B. A.; Aghagolzadeh, A.; Seyedarabi, H. (2011). "A non-reference image fusion metric based on mutual information of image features". *Computers & Electrical Engineering* **37** (5): 744–756.  
doi:10.1016/j.compeleceng.2011.07.012 (<https://dx.doi.org/10.1016%2Fj.compeleceng.2011.07.012>).
4. <http://www.mathworks.com/matlabcentral/fileexchange/45926-feature-mutual-information-fmi-image-fusion-metric>
5. Coombs, C. H., Dawes, R. M. & Tversky, A. (1970). *Mathematical Psychology: An Elementary Introduction*. Englewood Cliffs, New Jersey: Prentice-Hall.
6. Press, WH; Teukolsky, SA; Vetterling, WT; Flannery, BP (2007). "Section 14.7.3. Conditional Entropy and Mutual Information" (<http://apps.nrbook.com/empanel/index.html#pg=758>). *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). New York: Cambridge University Press. ISBN 978-0-521-88068-8
7. White, Jim; Steingold, Sam; Fournelle, Connie. "Performance Metrics for Group-Detection Algorithms" (<http://www.interfacesymposia.org/I04/I2004Proceedings/WhiteJim/WhiteJim.paper.pdf>).
8. Strehl, Alexander; Ghosh, Joydeep (2002), "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions" (<http://www.jmlr.org/papers/volume3/strehl02a/strehl02a.pdf>), *The Journal of Machine Learning Research* **3** (Dec): 583–617
9. Kvålseth, T. O. (1991). "The relative useful information measure: some comments". *Information sciences* **56** (1): 35–38.  
doi:10.1016/0020-0255(91)90022-m (<https://dx.doi.org/10.1016%2F0020-0255%2891%2990022-m>).
10. Pocock, A. (2012). *Feature Selection Via Joint Likelihood* ([http://academy.bcs.org/sites/academy.bcs.org/files/DD\\_Adam%20Pocock.pdf](http://academy.bcs.org/sites/academy.bcs.org/files/DD_Adam%20Pocock.pdf)) (Thesis).
11. Wentian Li (1990). "Mutual information functions versus correlation functions". *J. Stat. Phys.* **60** (5-6): 823–837.  
doi:10.1007/BF01025996 (<https://dx.doi.org/10.1007%2FBF01025996>).
12. [1] (<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.78.4178&rep=rep1&type=pdf>) Parsing a Natural Language Using Mutual Information Statistics by David M. Magerman and Mitchell P. Marcus
13. Hugh Everett Theory of the Universal Wavefunction (<http://www.pbs.org/wgbh/nova/manyworlds/pdf/dissertation.pdf>), Thesis, Princeton University, (1956, 1973), pp 1–140 (page 30)
14. Everett, Hugh (1957). "Relative State Formulation of Quantum Mechanics" (<http://www.univer.omsk.su/omsk/Sci/Everett/paper1957.html>). *Reviews of Modern Physics* **29**: 454–462.  
doi:10.1103/revmodphys.29.454 (<https://dx.doi.org/10.1103%2Frevmodphys.29.454>).



## References

- Cilibrasi, R.; Paul Vitányi (2005). "Clustering by compression" (<http://www.cwi.nl/~paulv/papers/cluster.pdf>) (PDF). *IEEE Transactions on Information Theory* **51** (4): 1523–1545. doi:10.1109/TIT.2005.844059 (<https://dx.doi.org/10.1109%2FTIT.2005.844059>).
- Cronbach L. J. (1954). On the non-rational application of information measures in psychology, in Henry Quastler, ed., *Information Theory in Psychology: Problems and Methods*, Free Press, Glencoe, Illinois, pp. 14–30.
- Church, Kenneth Ward; Hanks, Patrick (1989). "Word association norms, mutual information, and lexicography". *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. doi:10.1145/90000/89095 (<https://dx.doi.org/10.1145%2F90000%2F89095>).
- Guiasu, Silviu (1977). *Information Theory with Applications*. McGraw-Hill, New York. ISBN 978-0070251090.
- Li, Ming; Paul Vitányi (February 1997). *An introduction to Kolmogorov complexity and its applications*. New York: Springer-Verlag. ISBN 0-387-94868-6.
- Lockhead, G. R. (1970). "Identification and the form of multidimensional discrimination space". *Journal of Experimental Psychology* **85** (1): 1–10. doi:10.1037/h0029508 (<https://dx.doi.org/10.1037%2Fh0029508>).
- David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms* (<http://www.inference.phy.cam.ac.uk/mackay/itila/book.html>) Cambridge: Cambridge University Press, 2003. ISBN 0-521-64298-1 (available free online)
- Haghighat, M. B. A.; Aghagolzadeh, A.; Seyedarabi, H. (2011). "A non-reference image fusion metric based on mutual information of image features". *Computers & Electrical Engineering* **37** (5): 744–756. doi:10.1016/j.compeleceng.2011.07.012 (<https://dx.doi.org/10.1016%2Fj.compeleceng.2011.07.012>).
- Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*, second edition. New York: McGraw-Hill, 1984. (See Chapter 15.)
- Witten, Ian H. & Frank, Eibe. *Data Mining: Practical Machine Learning Tools and Techniques* (<http://www.cs.waikato.ac.nz/~ml/weka/book.html>). Morgan Kaufmann, Amsterdam. ISBN 978-0-12-374856-0.
- Peng, H.C., Long, F., and Ding, C. (2005). "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy" (<http://research.janelia.org/peng/proj/mRMR/index.htm>). *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27** (8): 1226–1238. doi:10.1109/tpami.2005.159 (<https://dx.doi.org/10.1109%2Ftpami.2005.159>).
- Andre S. Ribeiro, Stuart A. Kauffman, Jason Lloyd-Price, Bjorn Samuelsson, and Joshua Socolar (2008). "Mutual Information in Random Boolean models of regulatory networks". *Physical Review E* **77** (1). arXiv:0707.3642 (<https://arxiv.org/abs/0707.3642>).
- Wells, W.M. III; Viola, P., Atsumi, H., Nakajima, S., Kikinis, R. (1996). "Multi-modal volume registration by maximization of mutual information" (<http://www.ai.mit.edu/people/sw/papers/mia.pdf>) (PDF). *Medical Image Analysis* **1** (1): 35–51. doi:10.1016/S1361-8415(01)80004-9 (<https://dx.doi.org/10.1016%2FS1361-8415%2801%2980004-9>). PMID 9873920 (<https://www.ncbi.nlm.nih.gov/pubmed/9873920>).

Retrieved from "[http://en.wikipedia.org/w/index.php?title=Mutual\\_information&oldid=656383189](http://en.wikipedia.org/w/index.php?title=Mutual_information&oldid=656383189)"

Categories: Information theory | Entropy and information

- 
- This page was last modified on 14 April 2015, at 04:06.
  - Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the

Wikimedia Foundation, Inc., a non-profit organization.