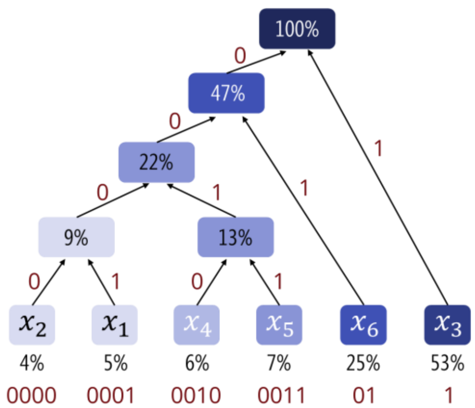
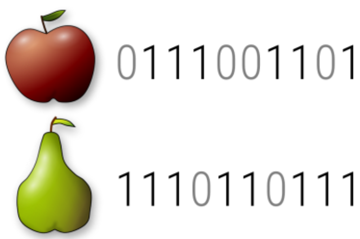
Machine Learning in Scientific Computing

CECAM/CSM/IRTG SCHOOL 2018





Max-Plank-Institut für Polymerforschung

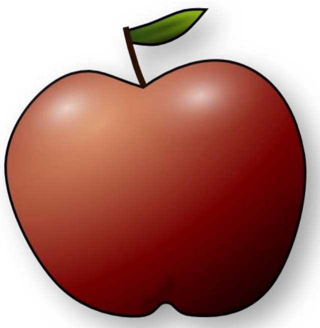
Computational Science Mainz

Lecture 3.1.1

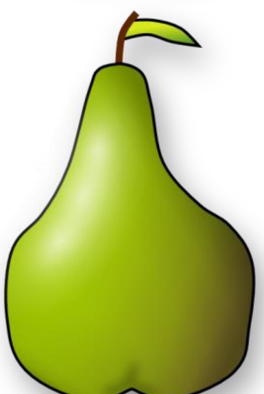
Information Theory

Michael Wand · Institut für Informatik, JGU Mainz · michael.wand@uni-mainz.de

Information



0111001101001010



1110110111001111

What is Information?

Defining Information

- Probability Theory
- Randomness = genuine new information

How much Information?

- Answer: “How random?”

LITERATURE:

Massimiliano Tomassoli: Information Theory for Machine Learning
May 2016, https://github.com/mtomassoli/papers/blob/master/inftheory.pdf

Axioms of Information

Random Information

- Random variable X
- Distribution $p(x)$

Information

- $I(x)$ – Information contained in observation of x

Axioms of Information

Axioms

- $I(x) = f(p(x))$ for some f
  - Information should only depend on distribution
- $p(x) < p(y) \Rightarrow f\big(p(x)\big) > f\big(p(y)\big)$
  - Strictly decreasing
  - Rarer events should carry more information
- $f(1) = 0$
  - Certain events carry no (new) information
- x, y independent $\Rightarrow I\big((\mathrm{x}, \mathrm{y})\big) = I(\mathrm{x}) + I(\mathrm{y})$
  - Information should add up
  - Independent experiments yield "totally new information"

Solution

Solution:

$$f(p) = -\log p = \log \frac{1}{p}$$

btw: the solution
is unique
(up to basis)

Proving the properties:

- $I(x) = \log \frac{1}{p(x)}$
- $p(x) < p(y) \Rightarrow \log \frac{1}{p(x)} > \log \frac{1}{p(y)}$
- $\log 1 = 0$
- x, y independent $\Rightarrow \log \frac{1}{p(x,y)} = \log \frac{1}{p(x)p(y)}$
  $= \log \frac{1}{p(x)} + \log \frac{1}{p(x)}$

Summary so far…

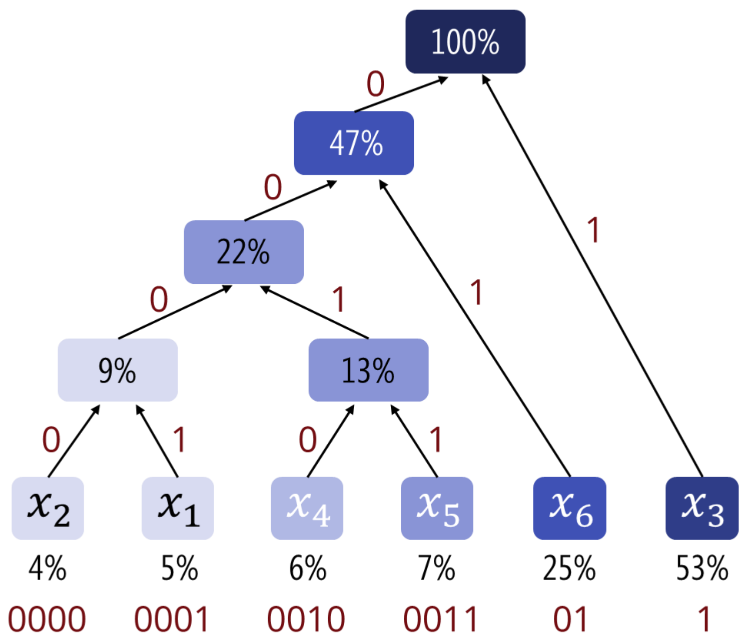
Probability

- Independent events: Product of probabilities
- Number between 0 and 1

Information

- Information is additive
  - More info: larger value
  - No information = 0
- Information of event = negative logarithm of prob.
  - $I(x) = -\log p(x) = \log \frac{1}{p(x)}$
  - Usually: base 2 (measured in bits)

Entropy



100%
0
47%
1
0
22%
1
0
1
9%
13%
0
1
0
1
x_2
x_1
x_4
x_5
x_6
x_3
4%
5%
6%
7%
25%
53%
0000
0001
0010
0011
01
1

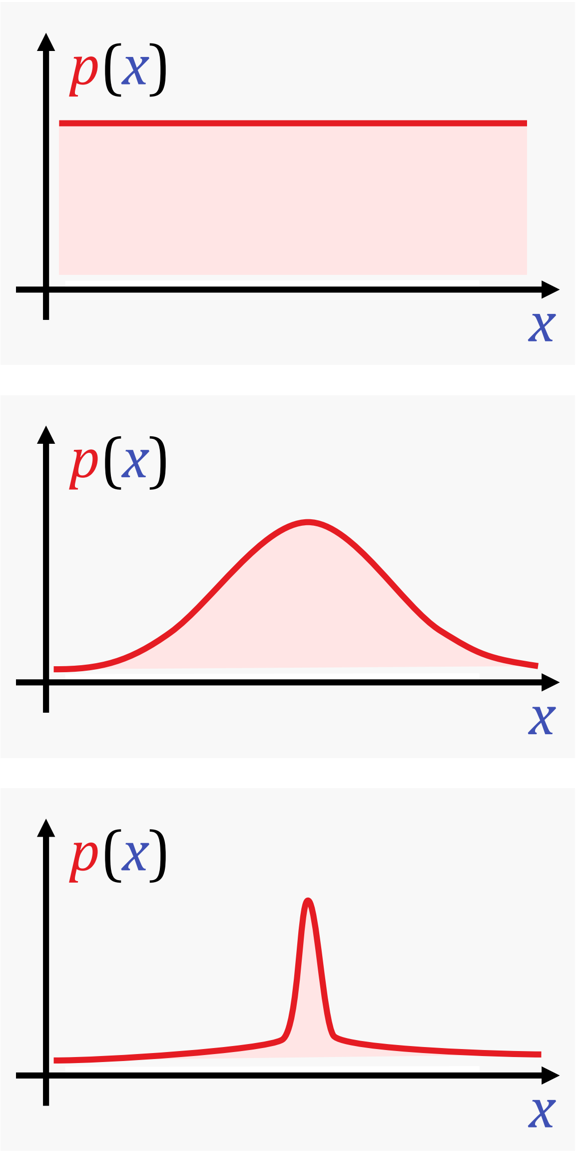
Entropy

Entropy: How random?

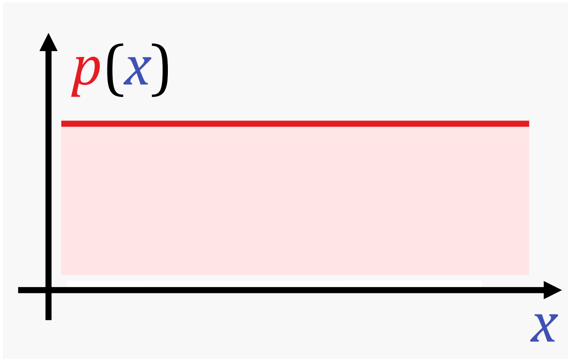
$$H(X) = \sum_{i=1}^{n} p(x_i)\log_2\frac{1}{p(x_i)}$$

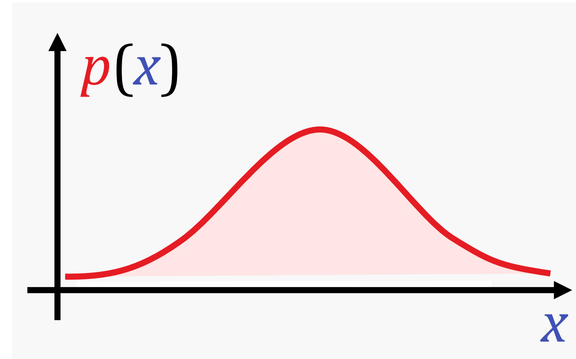
$$= -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

$$= \mathbb{E}_{x\sim p}\left[I_p(x)\right]$$

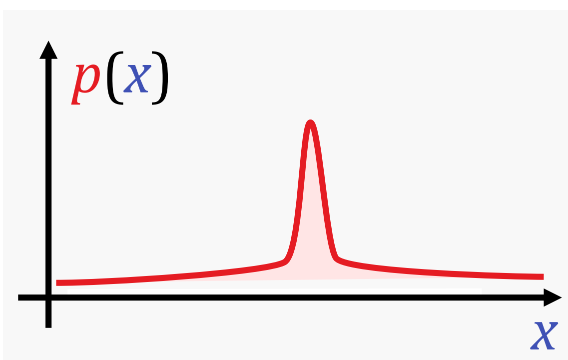


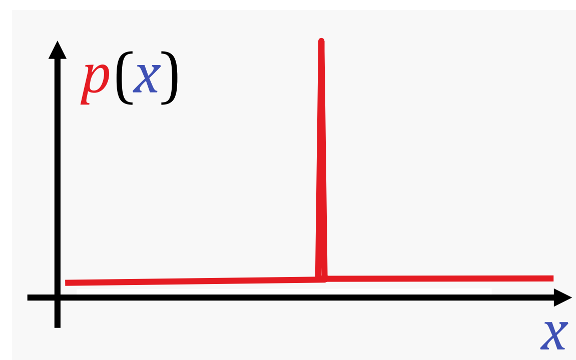
Examples





$$H = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n$$





$$H = 0$$

Entropy

Definition: Entropy

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

mean
neg log prob

$$= \sum_{i=1}^{n} p(x_i) I(x_i)$$

mean
information

$$= \mathbb{E}_{x \sim p(x)}\big(I(x)\big)$$

expected
information

Coding Theory

Entropy

- Minimum number of bits required to transmit information about event x
  - We draw events i.i.d.
  - We send each outcome separately
    - After being asked for the answer
    - (Certain outcomes: no answer required)
- Coding theorem:
  - $m(x)$ = message about x optimally encoded in bits
  - $H(X) \leq \mathbb{E}_{x \sim p(x)}\left(\text{length}\left(m(x)\right)\right) < H(X) + 1$

Random variable X distributed according to p(x)

Huffman Codes

Constructing a code

- Huffman algorithm
- Optimal for single events send in bits
  - Multiple symbols: Overhead up to one bit each
  - Optimality reached with “arithmetic coding”

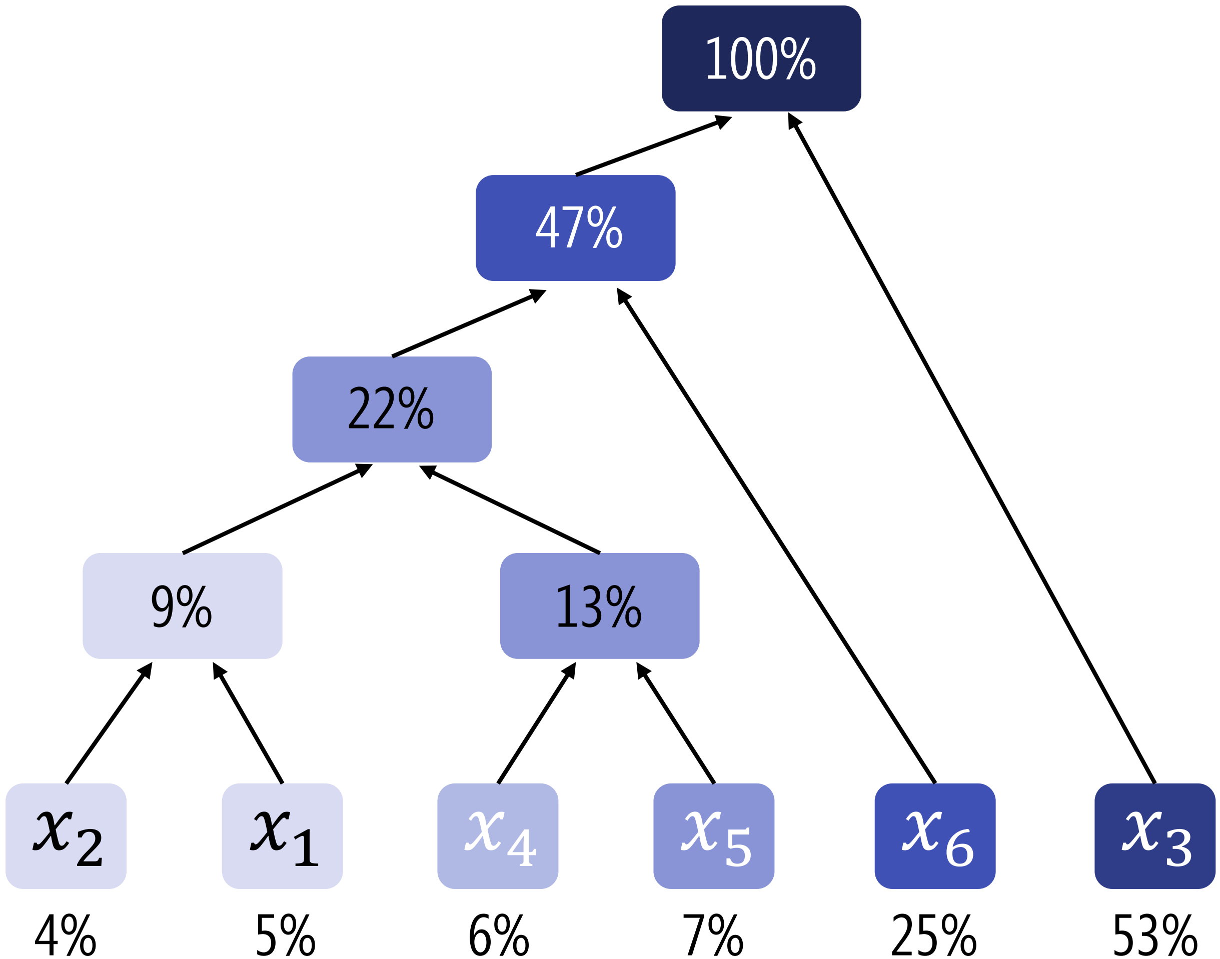
Huffman Codes

x_1	x_2	x_3	x_4	x_5	x_6
5%	4%	53%	6%	7%	25%

Huffman Codes

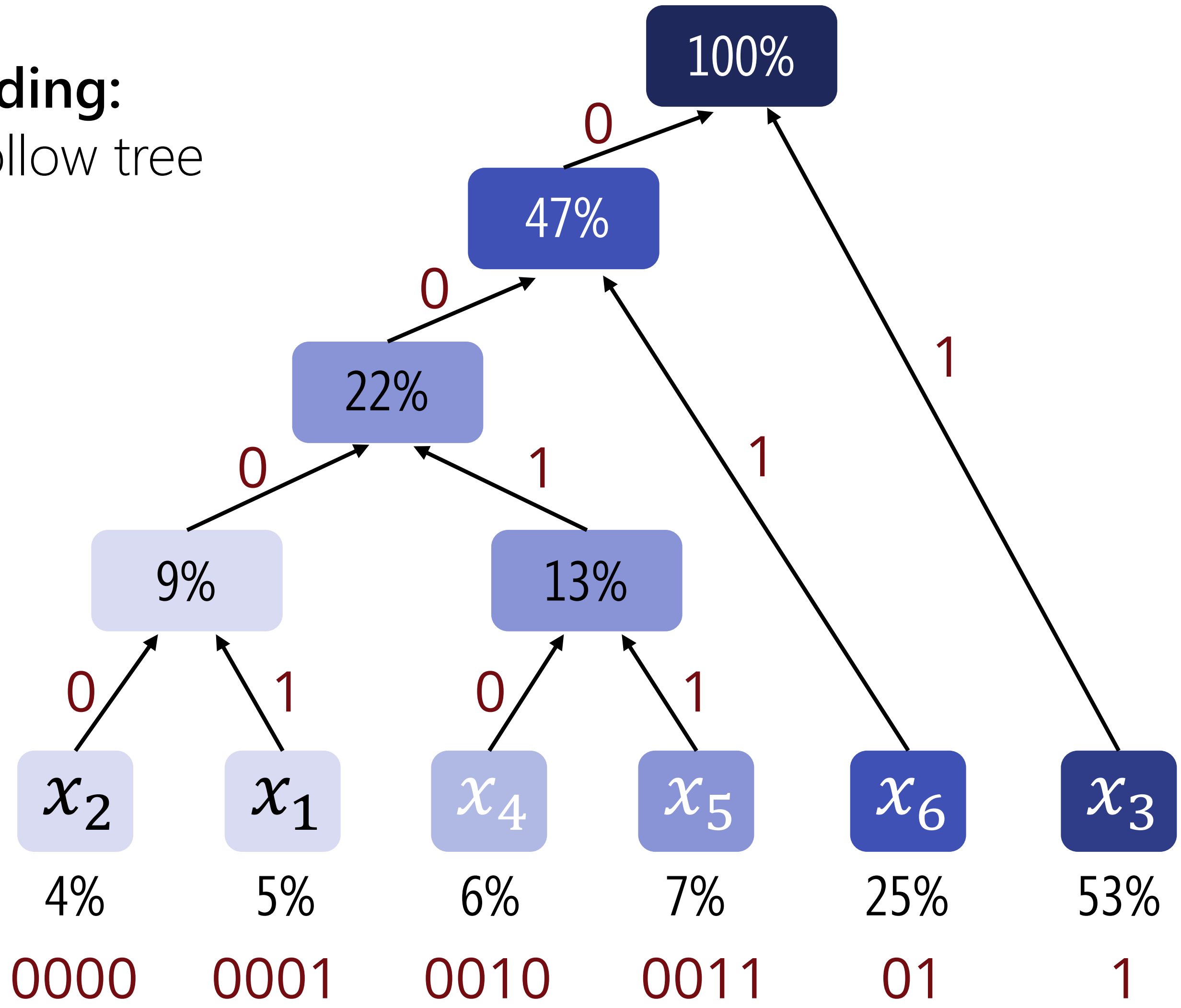
x_2	x_1	x_4	x_5	x_6	x_3
4%	5%	6%	7%	25%	53%

Huffman Codes



Huffman Codes

Decoding:
just follow tree



Bit-Coding

Coding of Symbols

- Number of bits $\leq \log \frac{1}{p(x)} + 1$
- Information = code length (up to one bit)
- Entropy = expected code length (up to one bit)

More Equations for Entropy

ADDITIONAL LITERATURE:

David McKay: Information Theory, Inference, and Learning Algorithms
Cambridge University Press, 2003. http://www.inference.org.uk/itprnn/book.pdf

Joint Entropy

Joint Entropy

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

- Simply the entropy of the joint distribution $p(x,y)$

Theorem

$$H(X,Y) = H(X) + H(Y)$$
$$\Leftrightarrow \; p(x,y) = p(x)p(y)$$

- Additive iff independent

Attention: Do not mix up with $H(p_1,p_2)$ for cross-entropy

Conditional Entropy

Conditional Entropy

$$H(X|Y) = -\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i|y_j)\log_2 p(x_i|y_j)$$

- Simply the entropy of the conditional distribution $p(x|y)$

Conditional Entropy

Marginal Entropy

$$H(X) = -\sum_{i=1}^{n_x} p(x_i) \log_2 p(x_i)$$
$$= -\sum_{i=1}^{n_x} \left(\sum_{j=1}^{n_y} p(x_i, y_j) \right) \left(\log_2 \sum_{j=1}^{n_y} p(x_i, y_j) \right)$$

- Simply the entropy of the marginal distribution $p(x)$

Conditional Entropy

Theorem: Chain Rule

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

“Divergences”:

Comparing Probability Distributions

Cross Entropy

Situation

- Two different distributions p_1, p_2 on the same probability space

Definition: Cross Entropy

$$H(p_1, p_2) = -\sum_{i=1}^{n} p_1(x) \log_2 p_2(x)$$

$$= \mathbb{E}_{x \sim p_1}\left[I_{p_2}(x)\right]$$

Idea

- Coding events $x \sim p_1$ with codes optimized for p_2

Kullback-Leibler Divergence

Kullback-Leibler Divergence

$$KL(p_1 \parallel p_2) = \sum_{i=1}^{n} p_1(x) \log_2 \frac{p_1(x)}{p_2(x)}$$

$$= H(p_1, p_2) - H(p_1, p_1)$$

$$= H(p_1, p_2) - H(p_1)$$

Idea

- Measure coding efficiency p_1 using p_2-codes
- Compare with opimum for p_1
- Price to pay for coding in p_2 rather than p_1
- Measures how far distribution p_2 is from p_1

KL and JS Divergences

Kullback-Leibler Divergence

- Distance ≥ 0
- Zero distance means same distribution
- Not symmetric:

$$KL(p_1 \parallel p_2) \text{ different from } KL(p_2 \parallel p_1)$$

- "Almost a metric"

Jensen–Shannon Divergence

- Symmetrized version
- $JSD(p_1 \parallel p_2) \coloneqq \frac{1}{2} KL(p_1 \parallel p_2) + \frac{1}{2} KL(p_2 \parallel p_1)$

Mutual Information

Mutual Information

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

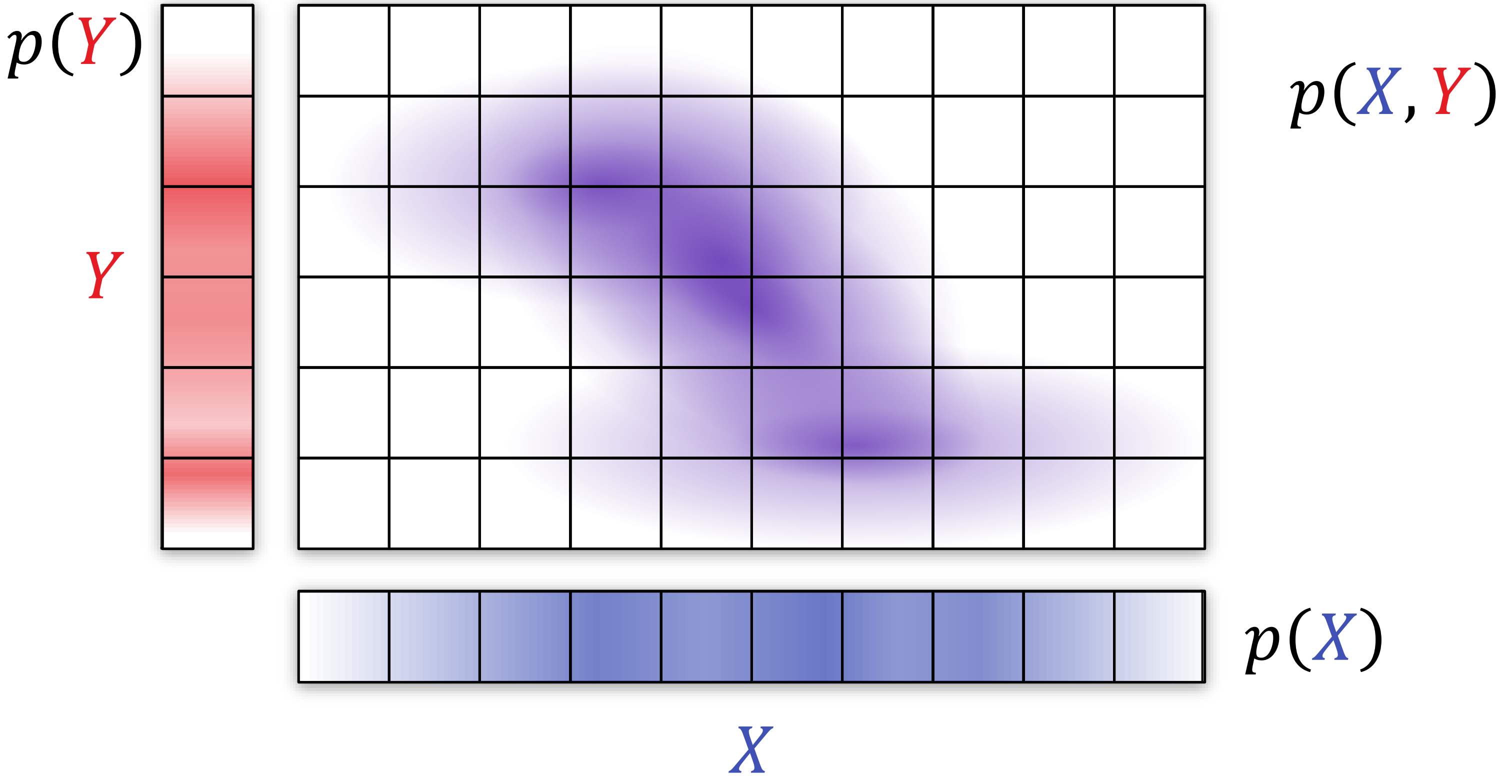
- Entropy of the marginal distributions minus that of the joint distribution

Mutual Information

Alternative Formulas

$$
\begin{aligned}
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= -\sum_{i=1}^{n_x} \sum_{j=1}^{n_y} p(x_i, y_j) \log_2 \left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)} \right) \\
&= KL\left(p(x_i, y_j) \parallel p(x_i)p(y_j) \right)
\end{aligned}
$$

Computing Mutual Information



Joint Histogram

- Compute $H(X)$, $H(Y)$, $H(X,Y)$
- Costly: $O(|\Omega_X| \times |\Omega_Y|)$ (exponential in $\dim(\Omega)$)

Alternatives

Parametric Distributions

- Closed-Form Expressions for Gaussians etc.
- $H\left(\mathcal{N}_{\mu,\Sigma}\right) = \frac{1}{2}\ln\left((2\pi e)^d \det(\Sigma)\right)$

Approximations

- Nearest-neighbors-methods
- Lower-bounds by “variational Bayes”
  - Build a neural network that predicts X from Y or vice versa
  - Least-squares fit
  - Entropy of Gaussian error (Covariance of errors) gives an upper bound of $H(X,Y)$ (joint Histogram, negative contrib.)