Visual Analytics
Lectorial week 7: Information Theory

Ian T. Nabney

University of BRISTOL

Learning Outcomes

- Reading: Section 1.6 of Bishop
- Able to compute the information content of a random variable
- Able to compute the Kullback–Leibler divergence between two random variables

Information theory for discrete variables

- How much infromation is received when we observe a specific value for a discrete random variable x?
- The amount of information can be viewed as the 'degree of surprise' on learning the value of x. If we are told that a highly improbable event has just occurred, we will have received more information than if we were told that some very likely event has just occurred.
- We are looking for a quantity $h(x)$ that is a function of $p(x)$.
- If we have two events x and y that are unrelated, then the information gain from observing both of them should be the sum of the information gained from each of them separately, so that $h(x, y) = h(x) + h(y)$.
- Two unrelated events will be statistically independent and so $p(x, y) = p(x)p(y)$.
- From these two relationships, it is easily shown that $h(x) = -\log_2 p(x)$.

University of BRISTOL

Entropy

- Now suppose that a sender wishes to transmit the value of a random variable to a receiver. The average amount of information that they transmit in the process is obtained by taking the expectation of $h(x)$ with respect to the distribution $p(x)$

$$\mathrm{H}[x] = -\sum_{x} p(x) \log_2 p(x). \tag{1}$$

- This is called the entropy of the random variable x.
- Note that $\lim_{p \to 0} p \ln p = 0$ and so we shall take $p(x) \ln p(x) = 0$ whenever we encounter a value for x such that $p(x) = 0$.

Worked example

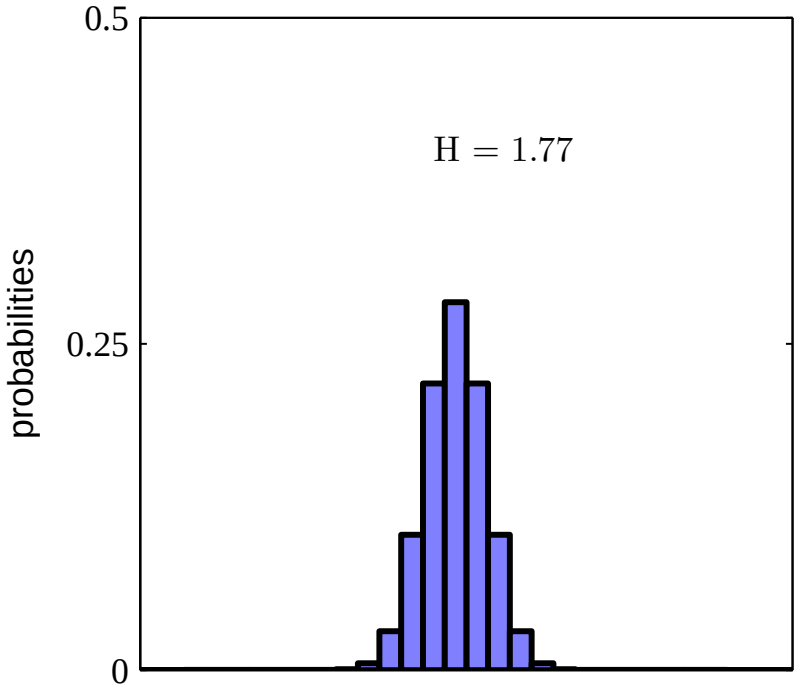
- Consider a random variable x having 8 possible states, each of which is equally likely. In order to communicate the value of x to a receiver, we would need to transmit a message of length 3 bits.
- The entropy of this variable is given by

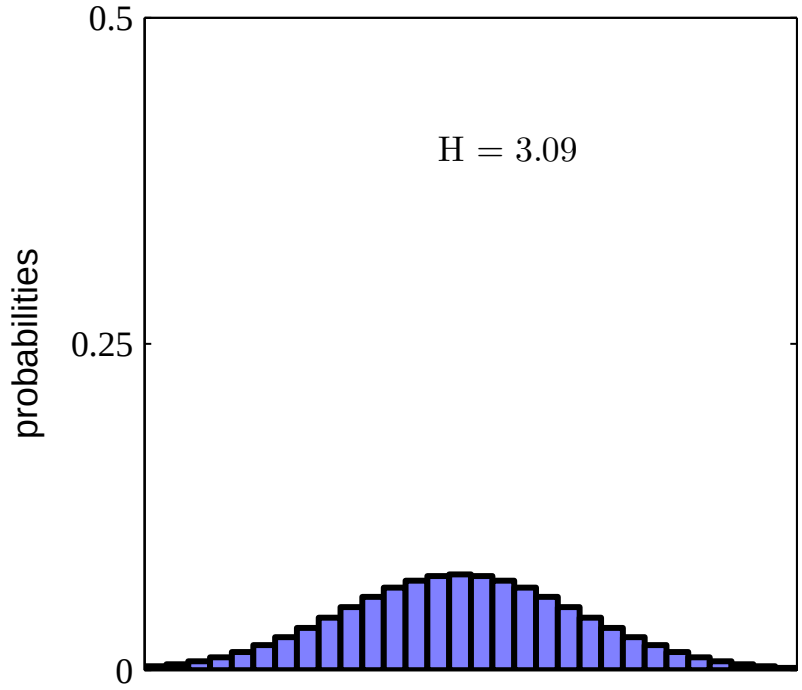
$$\mathrm{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits}.$$

- Consider a variable with 8 possible states $\{a, b, c, d, e, f, g, h\}$ for which the respective probabilities are given by $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$. The entropy in this case is given by

$$\mathrm{H}[x] = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{16} \log_2 \frac{1}{16} - \frac{4}{64} \log_2 \frac{1}{64} = 2 \text{ bits}.$$

- We see that the nonuniform distribution has a smaller entropy than the uniform one.





- Histograms of two probability distributions over 30 bins illustrating the higher value of the entropy H for the broader distribution.
- The largest entropy would arise from a uniform distribution that would give $\mathrm{H} = -\ln(1/30) = 3.40$.
- Because $0 \leqslant p_i \leqslant 1$, the entropy is nonnegative, and it will equal its minimum value of 0 when one of the $p_i = 1$ and all other $p_{j \neq i} = 0$.

Differential entropy

- We can extend the definition of entropy to include distributions $p(x)$ over continuous variables x by quantising x into bins of width Δ.
- Consider the limit $\Delta \to 0$. The discrete approximation will approach the integral of $p(x) \ln p(x)$ in this limit so that

$$\lim_{\Delta \to 0} \left\{ \sum_i p(x_i) \Delta \ln p(x_i) \right\} = - \int p(x) \ln p(x) \, \mathrm{d}x \tag{2}$$

  where the quantity on the right-hand side is called the differential entropy.
- If we maximise the entropy subject to three constraints (normalisation, mean, and variance) we get the Gaussian again!
- The differential entropy of the Gaussian is

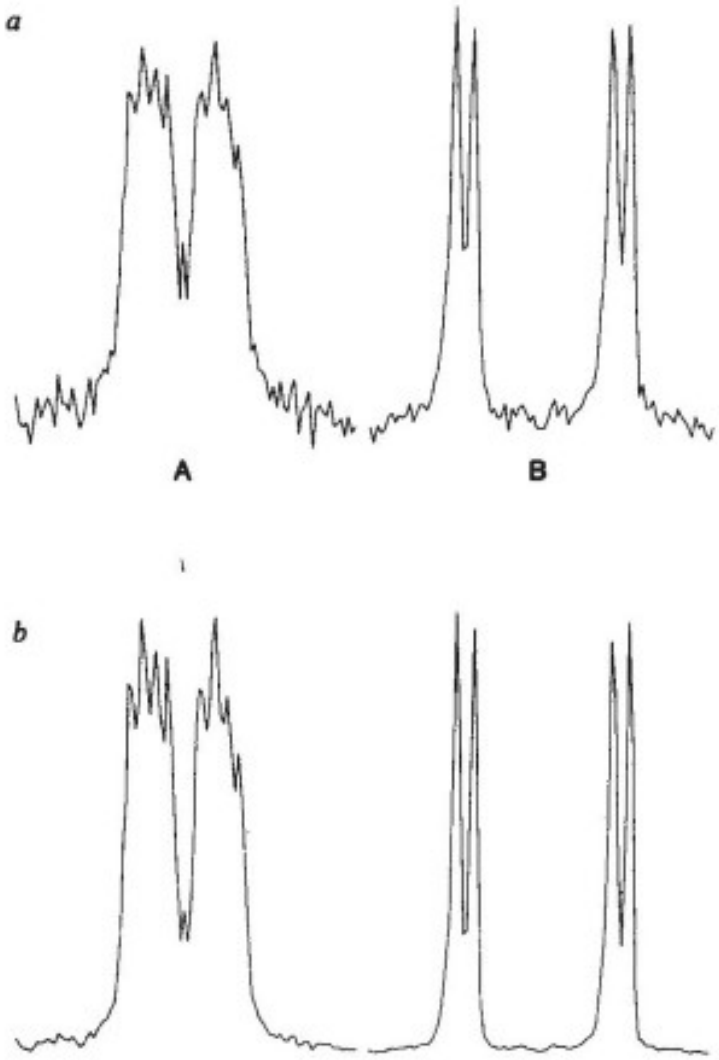
$$\mathrm{H}[x] = \frac{1}{2} \left\{ 1 + \ln(2\pi\sigma^2) \right\}. \tag{3}$$

  So the entropy increases as the distribution becomes broader, i.e., as σ^2 increases.
- This result also shows that the differential entropy can be negative, because $\mathrm{H}(x) < 0$ in (3) for $\sigma^2 < 1/(2\pi e)$.

University of BRISTOL

Maximum entropy

- The principle of maximum entropy states that the probability distribution which best represents the current state of knowledge is the one with largest entropy, in the context of precisely stated prior data.
- This has been used as a principle to develop signal processing algorithms by John Skilling.
- Compatibility with Bayes' theorem



University of BRISTOL

Conditional entropy

- Suppose we have a joint distribution $p(\mathbf{x}, \mathbf{y})$.
- If a value of $\mathbf{x}$ is already known, then the additional information needed to specify the corresponding value of $\mathbf{y}$ is given by $-\ln p(\mathbf{y}|\mathbf{x})$. Thus the average additional information needed to specify $\mathbf{y}$ cis

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x} \tag{4}$$

  which is called the conditional entropy of $\mathbf{y}$ given $\mathbf{x}$.
- The product rule leads to

$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}] \tag{5}$$

  Thus the information needed to describe $\mathbf{x}$ and $\mathbf{y}$ is given by the sum of the information needed to describe $\mathbf{x}$ alone plus the additional information required to specify $\mathbf{y}$ given $\mathbf{x}$.

Relative entropy and mutual information

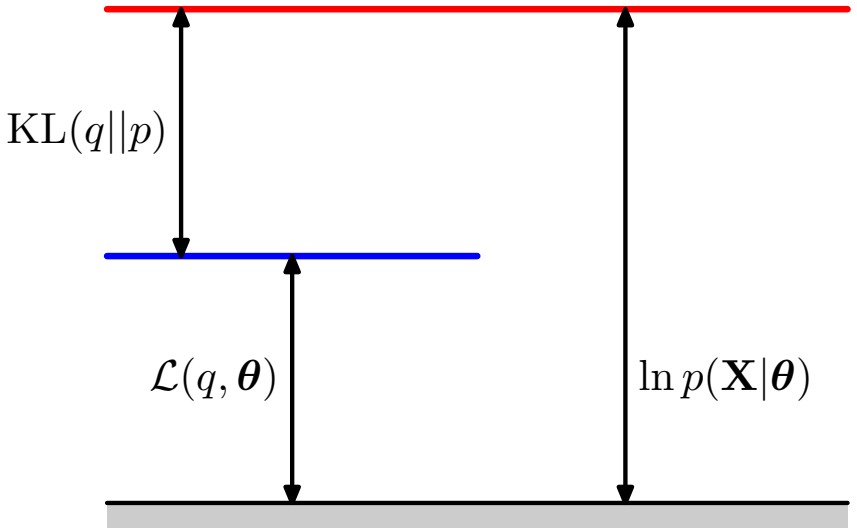
- The concept of information is useful when considering approximating one distribution $p(\mathbf{x})$ by another $q(\mathbf{x})$.
- The average additional amount of information (in nats) required to specify the value of $\mathbf{x}$ as a result of using $q(\mathbf{x})$ instead of the true distribution $p(\mathbf{x})$ is given by

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right) \\ &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x}. \end{aligned} \tag{6}$$

  This is known as the relative entropy or Kullback-Leibler divergence, or KL divergence, between the distributions $p(\mathbf{x})$ and $q(\mathbf{x})$.
- Note that it is not a symmetrical quantity: $\mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$.
- We can show that the Kullback-Leibler divergence satisfies $\mathrm{KL}(p\|q) \geqslant 0$ with equality if, and only if, $p(\mathbf{x}) = q(\mathbf{x})$.

KL divergence and machine learning

- There is an intimate relationship between data compression and density estimation.
- If we use a distribution that is different from the true one, then we must necessarily have a less efficient coding, and on average the additional information that must be transmitted is (at least) equal to the Kullback-Leibler divergence between the two distributions.



Decomposition of $\ln p(\mathbf{X}|\boldsymbol{\theta})$ for any choice of distribution $q(\mathbf{Z})$. Because the Kullback-Leibler divergence satisfies $\mathrm{KL}(q\|p) \geqslant 0$, we see that $\mathcal{L}(q,\boldsymbol{\theta})$ is a lower bound on the log likelihood function $\ln p(\mathbf{X}|\boldsymbol{\theta})$.

Mutual information

- Consider the joint distribution between two sets of variables $\mathbf{x}$ and $\mathbf{y}$ given by $p(\mathbf{x}, \mathbf{y})$. If the variables are independent, then their joint distribution will factorize into the product of their marginals $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$.
- We can measure how 'close' variables to being independent by considering the Kullback-Leibler divergence between the joint distribution and the product of the marginals, given by

$$\begin{aligned} \mathrm{I}[\mathbf{x}, \mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y} \end{aligned} \tag{7}$$

  which is called the mutual information between the variables $\mathbf{x}$ and $\mathbf{y}$.
- From the properties of the KL divergence, $I(\mathbf{x}, \mathbf{y}) \geqslant 0$ with equality if, and only if, $\mathbf{x}$ and $\mathbf{y}$ are independent.
- Using the sum and product rules of probability, the mutual information is related to the conditional entropy through

$$\mathrm{I}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]. \tag{8}$$

University of BRISTOL

Mutual information and machine learning

- Thus we can view the mutual information as the reduction in the uncertainty about **x** by virtue of being told the value of **y** (or vice versa).
- From a Bayesian perspective, we can view $p(\mathbf{x})$ as the prior distribution for **x** and $p(\mathbf{x}|\mathbf{y})$ as the posterior distribution after we have observed new data **y**.
- The mutual information therefore represents the reduction in uncertainty about **x** as a consequence of the new observation **y**.