Lecture 7: Information theory

University of BRISTOL

EMAT31530/Feb 2018/Raul Santos-Rodriguez

Have a look at ...

... David MacKay, Information Theory, *Inference, and Learning Algorithms*. Cambridge University Press, 2003. (Ch. 2)

... T. Cover and J. Tomas, *Elements of Information Theory*. Wiley, 2006.

... Information Theoretical Estimators (ITE) Toolbox.

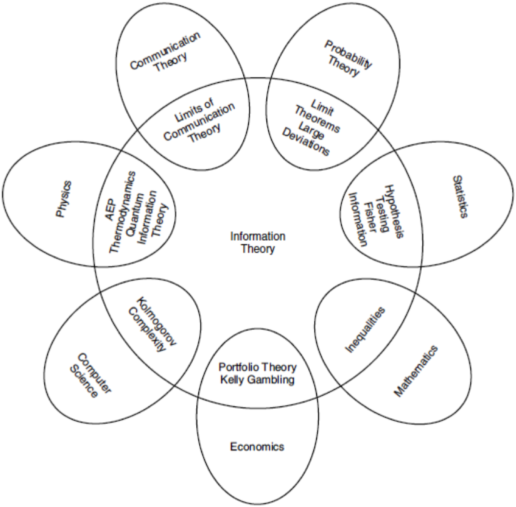
... 17 equations that changed the course of history!

Outline

Information theory is the use of probability theory to quantify and measure information. Basic concepts:

- Entropy and Information
- Mutual information
- Kullback-Leibler divergence

Information theory

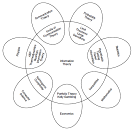


[Cover and Thomas, 2006]

2018-02-11

Information theory

Information theory



[Cover and Thomas, 2006]

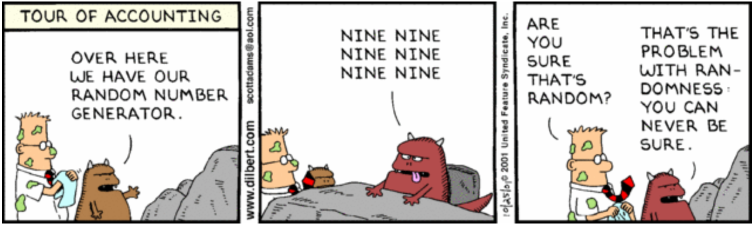
Information theory:

- Developed by Shannon in the 40s
- Idea: Maximising the amount of information that can be transmitted over an imperfect communication channel

Question

What is Information?

- "The sun will come up tomorrow"
- "It will rain tomorrow"
- "There was an earthquake this morning"



Probability theory revisited: probability distribution

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

Figure 2.1. Probability distribution over the 27 outcomes for a randomly selected letter in an English language document (estimated from *The Frequently Asked Questions Manual for Linux*). The picture shows the probabilities by the areas of white squares.

[Mackay, 2003]

Entropy

Definition

Entropy is a measure of the uncertainty associated with a distribution.

$$H(X) = -\sum_x p(x) \log p(x)$$

Lower bound on the number of *bits* that it takes to transmit messages!

Entropy

Definition

Entropy is a measure of the uncertainty associated with a distribution.

$$H(X) = -\sum_{x} p(x) \log p(x)$$

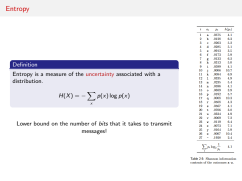
Lower bound on the number of *bits* that it takes to transmit messages!

i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
$\sum_i p_i \log_2 \frac{1}{p_i}$			4.1

Table 2.9. Shannon information contents of the outcomes a–z.

2018-02-11

Entropy



i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
$\sum_i p_i \log_2 \frac{1}{p_i}$			4.1

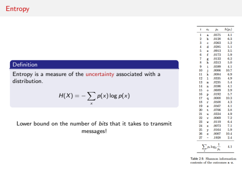
Table 2.9. Shannon information contents of the outcomes a–z.

Entropy measures the amount of information in a Random Variable; it can also be seen as the average length of the message needed to transmit an outcome of that variable using the optimal code.

The entropy of a randomly selected letter in an English document is about 4.11 bits, assuming its probability is as given in table 2.9. We obtain this number by averaging $\log 1/p_i$ (shown in the fourth column) under the probability distribution p_i (shown in the third column).

2018-02-11

Entropy



i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	-	.1928	2.4
$\sum_i p_i \log_2 \frac{1}{p_i}$			4.1

Table 2.9. Shannon information contents of the outcomes a–z.

The quantity of information is the entropy of the associated probability distribution. Suppose we have a single event A with possible outcomes $\{a_i\}$:

- If one, a_0, is certain to occur, $p(a_0) = 1$, then we acquire no information by observing A.
- If $A = a_0$ is very likely then we might have confidently expected it and so learn very little.
- If $A = a_0$ is highly unlikely then we might need to drastically change our plans.

Learning the value of A provides a quantity of information that increases as the corresponding probability decreases ($\sim -\log(\cdot)$).
We think of learning about something new as adding to the available information: the entropy is a weighted sum of the information we get from each event.

Letters

Consider a message formed from the alphabet A, B, C and D, with probabilities:

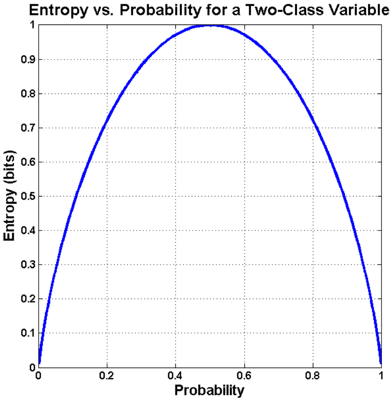
$$p(A) = \frac{1}{2}, \quad p(B) = \frac{1}{4}, \quad p(C) = \frac{1}{8} = p(D)$$

The information for the letter probabilities is

$$H(X) = -\left(\tfrac{1}{2}\log\tfrac{1}{2} + \tfrac{1}{4}\log\tfrac{1}{4} + 2\times\tfrac{1}{8}\log\tfrac{1}{8}\right) = \frac{7}{4}\ \text{bits}$$

Entropy: examples

What's the entropy of a uniform discrete random variable taking on 2 values?



Entropy

The entropy $H(X) = -\sum_x p(x) \log p(x)$ has the following properties:

Properties

- $H(X) \geq 0$. Entropy is always non-negative. $H(X) = 0$ iff X is deterministic.
- Upper-bound: $H(X) \leq \log(|\mathcal{X}|)$. $H(X) = \log(|\mathcal{X}|)$ iff X has a uniform distribution over $\mathcal{X}$.
- Since $H_b(X) = \log_b(a) H_a(X)$, we don't need to specify the base of the logarithm ($\log_2 \to$ *bits*, $\log_e \to$ *nats*).

Probability theory revisited: joint probability

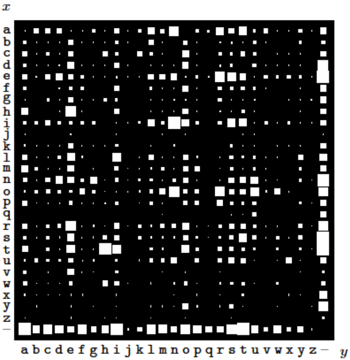


Figure 2.2. The probability distribution over the 27×27 possible bigrams xy in an English language document, *The Frequently Asked Questions Manual for Linux.*

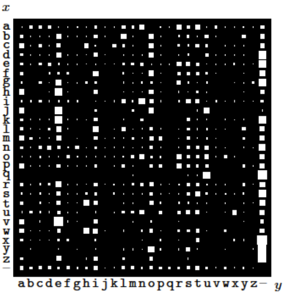
[Mackay, 2003]

Joint entropy

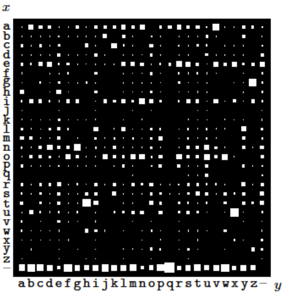
The joint entropy of a pair of two discrete random variables X and Y is:

$$H(X, Y) = -\sum_{x}\sum_{y} p(x, y) \log p(x, y)$$

Probability theory revisited: conditional probability



(a) $P(y \mid x)$



(b) $P(x \mid y)$

Figure 2.3. Conditional probability distributions. (a) $P(y \mid x)$: Each *row* shows the conditional distribution of the second letter, y, given the first letter, x, in a bigram xy. (b) $P(x \mid y)$: Each *column* shows the conditional distribution of the first letter, x, given the second letter, y.

[Mackay, 2003]

Conditional entropy

Conditional entropy

$$\begin{aligned}
H(Y|X) &= \sum_x p(x) H(Y|X=x) \\
&= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\
&= -\sum_x \sum_y p(x,y) \log p(y|x) \\
&= H(X,Y) - H(X)
\end{aligned}$$

Careful!

$$H(X|Y) \neq H(Y|X)$$

Mutual information

Mutual information

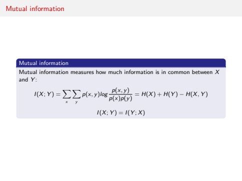
Mutual information measures how much information is in common between X and Y:

$$I(X;Y) = \sum_x \sum_y p(x,y) log \frac{p(x,y)}{p(x)p(y)} = H(X) + H(Y) - H(X,Y)$$

$$I(X;Y) = I(Y;X)$$

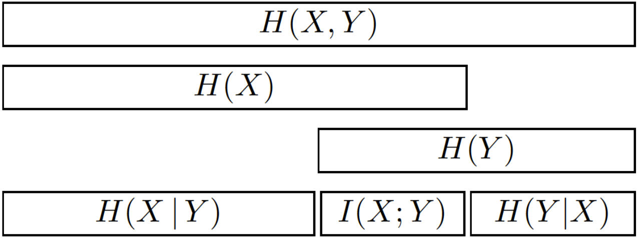
2018-02-11

Mutual information



$I(X;Y)$ is the mutual information between X and Y. It is the reduction of uncertainty of one Random Variable due to knowing about the other, or the amount of information one Random Variable contains about the other. Two interesting facts:

- I is 0 only when X,Y are independent: $H(X|Y) = H(X)$
- $H(X) = H(X) - H(X|X) = I(X,X)$. Entropy is the self-information.


$H(X,Y)$
$H(X)$
$H(Y)$
$H(X|Y)$
$I(X;Y)$
$H(Y|X)$

Kullback-Leibler divergence

Kullback-Leibler (KL) divergence

The Kullback-Leibler divergence between two distributions on the same alphabet is

$$KL(p||q) = \sum p(x) log \frac{p(x)}{q(x)}$$

- KL is a 'distance' measure between probability functions p and q.
- KL divergence is asymmetric (not a true distance):

$$KL(p||q) \neq KL(q||p)$$

- KL and Mutual information:

$$I(X;Y) = KL(p(x,y)||p(x)p(y))$$

Nonlinear Dimensionality Reduction

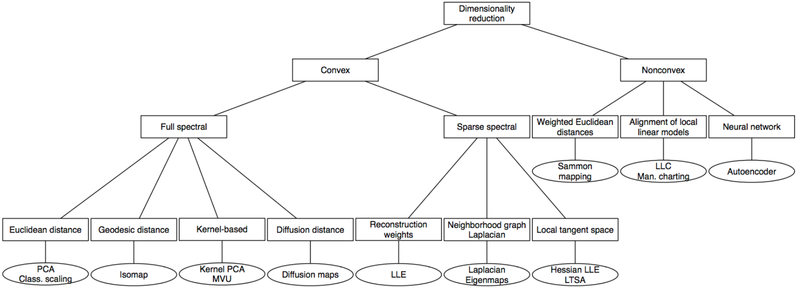


Figure: Taxonomy of dimensionality reduction techniques

t-Distributed stochastic neighbour embedding (t-SNE)

t-SNE minimises divergence of two distributions over pairwise similarities of

P input objects

Q low-dimensional points in the embedding

t-SNE algorithm

- Distance between a pair of objects, e.g., $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|$
- Joint probabilities p_{ij} that measure the pairwise similarity between objects $\mathbf{x}_i$ and $\mathbf{x}_j$

$$p(j|i) = \frac{\exp(-d(\mathbf{x}_i, \mathbf{x}_j)^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-d(\mathbf{x}_i, \mathbf{x}_k)^2/2\sigma_i^2)}, \quad p(i|i) = 0$$

$$p_{ij} = \frac{p(j|i) + p(i|j)}{2N}$$

$$q_{ij} = \frac{(1 + d(\mathbf{y}_i - \mathbf{y}_j)^2)^{-1}}{\sum_{k \neq i}(1 + d(\mathbf{y}_i - \mathbf{y}_k)^2)^{-1}}$$

- Minimise cost function (KL-divergence)

$$KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

t-Distributed stochastic neighbour embedding (t-SNE)

t-SNE minimises divergence of two distributions over pairwise similarities of

P input objects

Q low-dimensional points in the embedding

t-SNE algorithm

- Distance between a pair of objects, e.g., $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|$
- Joint probabilities p_{ij} that measure the pairwise similarity between objects $\mathbf{x}_i$ and $\mathbf{x}_j$

$$p(j|i) = \frac{\exp(-d(\mathbf{x}_i, \mathbf{x}_j)^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-d(\mathbf{x}_i, \mathbf{x}_k)^2/2\sigma_i^2)}, \quad p(i|i) = 0$$

$$p_{ij} = \frac{p(j|i) + p(i|j)}{2N}$$

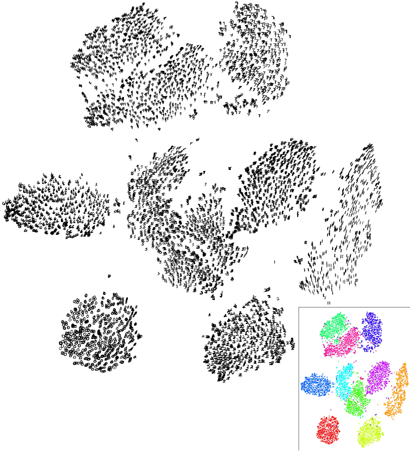
$$q_{ij} = \frac{(1 + d(\mathbf{y}_i - \mathbf{y}_j)^2)^{-1}}{\sum_{k \neq i}(1 + d(\mathbf{y}_i - \mathbf{y}_k)^2)^{-1}}$$

- Minimise cost function (KL-divergence)

$$KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

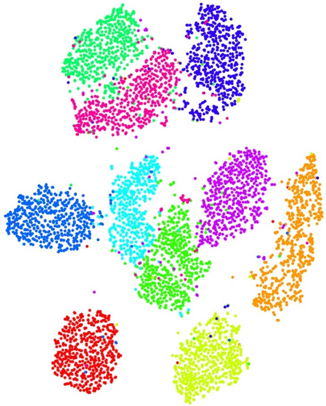
+ t-SNE compares favourably to other techniques for data visualisation
- unclear how t-SNE performs on general dimensionality reduction tasks
- relatively local nature of t-SNE makes it sensitive to the curse of the intrinsic dimensionality of the data
- not guaranteed to converge to a global optimum of its cost function

t-SNE on MNIST



http://scikit-learn.org/stable/auto_examples/manifold/plot_t_sne_perplexity.html

t-SNE on MNIST



http://scikit-learn.org/stable/auto_examples/manifold/plot_t_sne_perplexity.html

Summary

- Entropy is the measure of average *uncertainty* in the random variable.
- Entropy is the average number of *bits needed* to describe the random variable.
- Entropy is a lower bound on the average length of the *shortest description* of the random variable.
- Conditional entropy $H(X|Y)$ is the entropy of one random variable conditional *upon knowledge* of another.
- The average amount of decrease of the randomness of X by observing Y is the average information that Y gives us about X.
- Mutual information measures how much information is in *common* between X and Y.
- KL divergence measures the *'distance'* between probability distributions.

Summary

Summary

- Entropy is the measure of average *uncertainty* in the random variable.
- Entropy is the average number of *bits needed* to describe the random variable.
- Entropy is a lower bound on the average length of the *shortest description* of the random variable.
- Conditional entropy $H(X|Y)$ is the entropy of one random variable conditional *upon knowledge* of another.
- The average amount of decrease of the randomness of X by observing Y is the average information that Y gives us about X.
- Mutual information measures how much information is in *common* between X and Y.
- KL divergence measures the *'distance'* between probability distributions.

Entropy is measure of uncertainty: the more we know about something, the lower the entropy. If a model captures more of the structure of the data, then the entropy should be lower. We can use entropy as a measure of the quality of our models. The KL divergence measures of how different two probability distributions are and can be interpreted as the average number of bits that are wasted by encoding events from a distribution p with a code based on a not-quite right distribution q:

Goal: minimise $D(p\|\|q)$ to have a probabilistic model as accurate as possible

Next lecture

We will discuss decision trees and random forests.