Information theory

- A binary code for transmitting poker hands:

straight flush	000
four of a kind	001
full house	010
flush	100
straight	101
two pair	110
pair	011
high card	111

- This will require $C = 3$ bits per hand. In general, to encode k possibilities we will need $C = \lceil \log_2(k) \rceil$ bits.

Information theory

- An improved code, taking advantage of uneven probabilities:

straight flush	0.0000162	11111111
four of a kind	0.000252	11111110
full house	0.00151	1111110
flush	0.00206	11110
straight	0.00412	1110
two pair	0.0222	110
pair	0.444	10
high card	0.526	0

- Now the expected value of the message length is $E[C] = 1.52$ bits per hand. In general, an optimal code will use $\lceil -\log_2(P(x)) \rceil$ for hand x.

Entropy

- The *entropy* of a random variable $H(X)$ is a lower bound on the expected value of the message length $E_P[C]$:

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$$

The *Source Coding Theorem* tells us that there exists an encoding such that:

$$H(X) \leq C < H(X) + 1$$

- We can use $H(X)$ as a measure of how much uncertainty there is in the distribution of X.
- Shannon (1948) : `http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html`

Relative entropy

- How many bits do we waste on average by using the wrong encoding?

$$\begin{aligned} E_P[-\log_2 Q(x) - -\log_2 P(x)] &= \sum_{x \in X} P(x)(-\log_2 Q(x) - -\log_2 P(x)) \\ &= \sum_{x \in X} P(x)(\log_2 P(x) - \log_2 Q(x)) \\ D(P||Q) &= \sum_{x \in X} P(x) \log_2 \frac{P(x)}{Q(x)} \end{aligned}$$

- This *relative entropy* (or *Kullback-Leibler divergence*) can be used as a measure of how closely two probability distributions agree.
- The entropy and divergence is usually measured in *bits* ($\log_2$) or *nats* ($\log_e$).

Log likelihood

- Often, we are interested in comparing how well two models Q_1 and Q_2 match an empirical distribution P.
- In this case, we can equivalently look at the *log likelihood*:

$$\begin{aligned} D(P\|Q) &= \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \\ &= \sum_{x \in X} P(x) \log P(x) - \sum_{x \in X} P(x) \log Q(x) \end{aligned}$$

$$L_P(Q) = \sum_{x \in X} P(x) \log Q(x)$$

- Related to *cross entropy* $(-L_P(Q))$ and *perplexity* $(2^{-L_P(Q)})$.

Conditional entropy

- We can also define the joint and conditional entropy for combinations of random variables:

$$\begin{aligned} H(XY) &= -\sum_{x \in X}\sum_{y \in Y} P(x,y) \log P(x,y) \\ H(Y|X) &= \sum_{x \in X} P(x) H(Y|X=x) \\ &= \sum_{x \in X} P(x) \left(-\sum_{y \in Y} P(y|x) \log P(y|x) \right) \\ &= -\sum_{x \in X}\sum_{y \in Y} P(x,y) \log P(y|x) \end{aligned}$$

- Like joint and conditional probability, joint and conditional entropy are related:

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

Mutual information

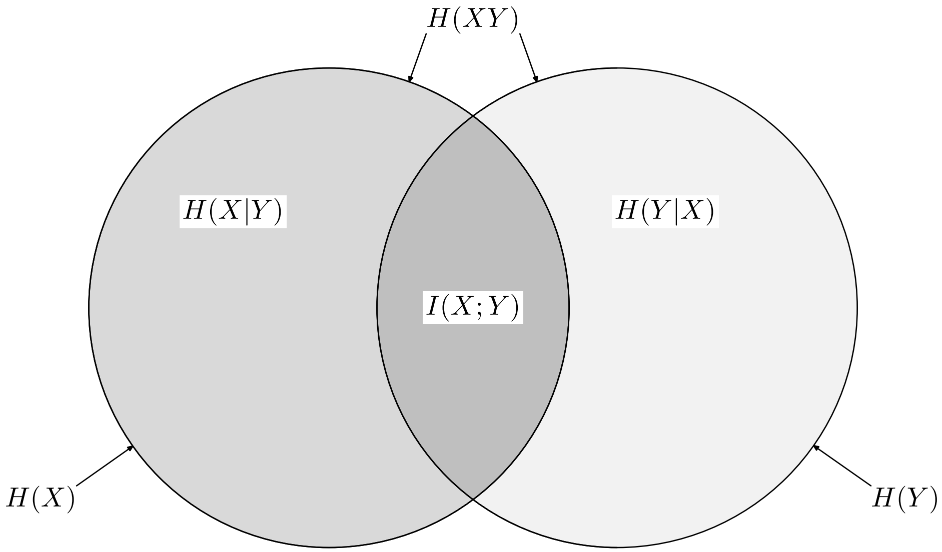
- The *average mutual information* $I(X;Y)$ measures how much knowing the value of one random variable reduces the uncertainty about another:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{x \in X}\sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\ &= I(Y;X) \\ &\geq 0 \end{aligned}$$

- $I(X;Y)$ is the expected value of:

$$I(x,y) = \log \frac{P(x,y)}{P(x)P(y)}$$

Mutual information



Mutual information

- If X and Y are independent, then $I(X;Y) = 0$.
- The average MI of dependent variables depenends on their entropy:

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

- Average MI is also related to relative entropy:

$$I(X;Y) = D(P(X,Y)||P(X)P(Y))$$

Recall that if X and Y are independent, then $P(X,Y) = P(X)P(Y)$.

Mutual information

- Rate of information transmission
- Infomax principle
- Corpus linguistics, lexicography
  - $I(w_1, w_2) >> 0$ means w_1 occurs together with w_2 *more* often than you would expect by chance
  - $I(w_1, w_2) << 0$ means w_1 occurs together with w_2 *less* often than you would expect by chance
  - $I(w_1, w_2) \approx 0$ means there is no evidence for a relationship between w_1 and w_2
- For word counts:

$$I(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)} = \log_2 \frac{f(w_1, w_2) \times N}{f(w_1) \times f(w_2)}$$

Mutual information

$f(ik)$	4094
$f(mijn)$	951
$f(ik, mijn)$	528
$I(ik, mijn)$	2.61

$f(ik)$	4094
$f(voor)$	4491
$f(ik, voor)$	449
$I(ik, voor)$	0.137

$f(ik)$	4094
$f(zij)$	1321
$f(ik, zij)$	60
$I(ik, zij)$	-1.00

Mutual information

$f(regering)$	150
$f(partij)$	133
$f(regering, partij)$	4
$I(minister, regering)$	3.17

$f(jongen)$	137
$f(meisje)$	112
$f(jongen, meisje)$	10
$I(jongen, meisje)$	4.88

$f(jong)$	95
$f(oud)$	118
$f(jong, oud)$	5
$I(jong, oud)$	4.32

Mutual information

$f(tweede)$	296
$f(kamer)$	194
$f(tweede, kamer)$	31
$I(tweede, kamer)$	4.60

$f(verenigde)$	66
$f(naties)$	16
$f(verenigde, naties)$	13
$I(verenigde, naties)$	9.11

Mutual information

Main verb	Object noun	MI	Joint freq
drink	martinis	12.6	3
drink	cup of water	11.6	3
drink	champagne	10.9	3
drink	beverage	10.8	8
drink	cup of coffee	10.6	2
drink	cognac	10.6	2
drink	beer	9.9	29
drink	cup	9.7	6
drink	coffee	9.7	12
drink	toast	9.6	4
drink	alcohol	9.4	20
drink	wine	9.3	10
drink	fluid	9.0	5
drink	liquor	8.9	4
drink	tea	8.9	5
drink	milk	8.7	8
drink	juice	8.3	4
drink	water	7.2	43
drink	quantity	7.1	4

Association ratio

- The *association ratio* (Church and Hanks 1991) is an order dependent version of MI better suited for identifying collocations.
- For a specified window size k, we count the number of occurences of w_2 inside k grams which start with w_1:

Trading	,	prevents	banks	from	processing	card	transactions
		w_1		w_2			

- In calculating AR, we need to account for the size of the window:

$$AR(w_1, w_2, k) = \log_2 \frac{f(w_1, w_2) \times N}{f(w_1) \times f(w_2) \times (k-1)}$$

Association ratio

nez	perce	16.31922
saudi	arabia	16.29303
los	angeles	16.11955
saddam	hussein	16.05913
latent	inhibition	15.36061
frau	nordern	15.28476
san	francisco	15.05446
outer	hebrides	14.96902
herr	nordern	14.95295
el	salvador	14.91480

british	telecom	10.71992
lake	district	10.68750
grammar	schools	10.62924
rugby	league	10.61507
football	league	10.61286
rugby	union	10.60777
de	la	10.57064
eastern	europe	10.54401
governing	body	10.53793
vast	majority	10.53119

Association ratio

world	and	0.0015891377
)	as	0.0013817724
away	the	0.0011113275
,	through	0.0005632977
system	to	0.0003462333
met	,	0.0001937076
,	poor	-0.0007452413
it	very	-0.0015464439
the	hour	-0.0028314710
.	price	-0.0041136840

to	and	-5.429041
a	and	-5.791279
a	of	-5.809393
to	in	-5.917458
of	.	-6.124886
a	,	-6.429715
a	.	-6.674252
.	.	-6.991823
of	and	-7.060864
.	,	-8.134681

Information theory

"Information theory has perhaps ballooned to an importance beyond its actual accomplishments." —Claude Shannon, 1956

- Surprise, optimal codes, entropy
- Kullback-Keibler divergence, cross-entropy, perplexity
- Mutual information, association ratio