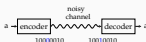


- Information theory is concerned with measurement, storage and transmission of information
- It has its roots in communication theory, but is applied to many different fields NLP
- We will revisit some of the major concepts

## Noisy channel model



- We want codes that are efficient: we do not want to waste the channel bandwidth
- We want codes that are resilient to errors: we want to be able to detect and correct errors
- This simple model has many applications in NLP, including in speech recognition and machine translation

## Coding example

binary coding of an eight-letter alphabet

letter	code
a	00000001
b	00000010
c	00000100
d	00001000
e	00010000
f	00100000
g	01000000
h	10000000

- We can encode an 8-letter alphabet with 8 bits using one-hot representation
- Can we do better than one-hot coding?

## Coding example

binary coding of an eight-letter alphabet

letter	code
a	00000000
b	00000001
c	00000010
d	00000011
e	00000100
f	00000101
g	00000110
h	00000111

- We can encode an 8-letter alphabet with 8 bits using one-hot representation
- Can we do better than one-hot coding?
- Can we do even better?

## Self information / surprisal

Self information (or surprisal) associated with an event  $x$  is

$$I(x) = \log \frac{1}{P(x)} = -\log P(x)$$

- If the event is certain, the information (or surprise) associated with it is 0
- Low probability (surprising) events have higher information content
- Base of the log determines the unit of information
  - 2 bits
  - e nats
  - 10 dit, ban, hartley

## Why log?

- Reminder: logarithms transform exponential relations to linear relations
 
$$\log ab = \log a + \log b \quad \log a^n = n \log a$$
- In most systems, linear increase in capacity increases possible outcomes exponentially
  - Number of possible  $n$ -word combinations is exponentially (twice) more than the number of possible  $(n-1)$ -word combinations
  - But we expect information to increase linearly, not exponentially
- Working with logarithms is more numerically stable

## Entropy

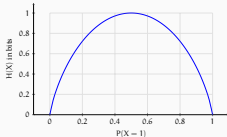
Entropy is a measure of the uncertainty of a random variable:

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

- Entropy is the lower bound on the best average code length, given the distribution  $P$  that generates the data
- Entropy is average surprisal:  $H(X) = E[-\log P(x)]$
- It generalizes to continuous distributions as well (replace sum with integral)

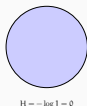
Entropy is about a distribution, while surprisal is about individual events

## Example: entropy of a Bernoulli distribution



## Entropy: demonstration

increasing number of outcomes increases entropy



## Entropy: demonstration

increasing number of outcomes increases entropy



$$H = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

## Entropy: demonstration

increasing number of outcomes increases entropy



$$H = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 2$$

## Entropy: demonstration

the distribution matters



$$H = -\frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} = 1.792481250360578$$

## Entropy: demonstration

the distribution matters



$$H = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{12} \log_2 \frac{1}{12} - \frac{1}{12} \log_2 \frac{1}{12} - \frac{1}{12} \log_2 \frac{1}{12} - \frac{1}{12} \log_2 \frac{1}{12} = 1.207518749639422$$

## Back to coding letters



- Can we do better?
- No.  $H = 3$  bits, we need 3 bits on average

letter	prob	code
a	$\frac{1}{8}$	000
b	$\frac{1}{8}$	001
c	$\frac{1}{8}$	010
d	$\frac{1}{8}$	011
e	$\frac{1}{8}$	100
f	$\frac{1}{8}$	101
g	$\frac{1}{8}$	110
h	$\frac{1}{8}$	111

Uniform distribution has the maximum uncertainty, hence the maximum entropy.

## Back to coding letters



- Can we do better?
- No.  $H = 3$  bits, we need 3 bits on average
- If the probabilities were different, could we do better?
- Yes. Now  $H = 2$  bits, we need 2 bits on average

letter	prob	code
a	$\frac{1}{4}$	0
b	$\frac{1}{4}$	10
c	$\frac{1}{4}$	110
d	$\frac{1}{4}$	1110
e	$\frac{1}{16}$	111100
f	$\frac{1}{16}$	111101
g	$\frac{1}{16}$	111110
h	$\frac{1}{16}$	111111

Uniform distribution has the maximum uncertainty, hence the maximum entropy.

## Pointwise mutual information

Pointwise mutual information (PMI) between two events is defined as

$$PMI(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- Reminder:  $P(x, y) = P(x)P(y)$  if two events are independent PMI
  - if the events are independent
  - if events cooccur more than they would occur by chance
  - if events cooccur less than they would occur by chance
- Pointwise mutual information is symmetric  $PMI(X, Y) = PMI(Y, X)$
- PMI is often used as a measure of association (e.g., between words) in computational/corpus linguistics

## Mutual information

Mutual information measures mutual dependence between two random variables

$$MI(X, Y) = \sum_x \sum_y P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)}$$

- MI is the average (expected value of) PMI
- PMI is defined on events, MI is defined on distributions
- Note the similarity with the covariance (or correlation)
- Unlike correlation, mutual information is
  - also defined for discrete variables
  - also sensitive to the non-linear dependence

## Conditional entropy

Conditional entropy is the entropy of a random variable conditioned on another random variable.

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

- $H(X|Y) = H(X)$  if random variables are independent
- Conditional entropy is lower if random variables are dependent

## Entropy, mutual information and conditional entropy



## Cross entropy

Cross entropy measures entropy of a distribution  $P$ , under another distribution  $Q$ .

$$H(P, Q) = - \sum_x P(x) \log Q(x)$$

- It often arises in the context of approximation:
  - if we approximate the true distribution  $P$  with  $Q$
- It is always larger than  $H(P)$ : it is the (non-optimum) average code-length of  $P$  coded using  $Q$
- It is a common *error function* in ML for categorical distributions

Note: the notation  $H(X, Y)$  is also used for *joint entropy*.

## Perplexity

Perplexity is the exponential version of (cross) entropy:

$$PP(X) = 2^{H(X)}$$

- Perplexity 'undoes' the logarithmic scaling
- Perplexity easier to interpret in some contexts
- Especially for language models, its interpretation is the average 'branching factor'

Predict the next word: (S) The perplexity of a random variable ( $/S$ )

## KL-divergence / relative entropy

For two distribution  $P$  and  $Q$  with same support, Kullback-Leibler divergence of  $Q$  from  $P$  (or relative entropy of  $P$  given  $Q$ ) is defined as

$$D_{KL}(P||Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}$$

- $D_{KL}$  measures the amount of extra bits needed when  $Q$  is used instead of  $P$
- $D_{KL}(P||Q) = H(P, Q) - H(P)$
- Used for measuring the difference between two distributions
- Note: it is not symmetric (not a distance measure)

## Continuous random variables and differential entropy

- For continuous random variables, we cannot sum over all possible outcomes, but we can integrate over the ranges of outcomes
- Information entropy (and all relevant measures) generalizes to the continuous distributions

$$h(X) = - \int_x p(x) \log p(x)$$

- The entropy of continuous variables is called *differential entropy*
- Differential entropy is typically measures in *nats*

## Short divergence: distance measure (again)

A *distance function*, or a *metric*, satisfies:

- $d(x, y) \geq 0$
- $d(x, y) = d(y, x)$
- $d(x, y) = 0 \iff x = y$
- $d(x, y) \leq d(x, z) + d(z, y)$

We will encounter measures/metrics frequently in this course.

## Summary

- Information theory has many applications in NLP and ML
- We reviewed a number of important concepts from the information theory
  - Self information
  - Pointwise MI
  - Cross entropy
  - Entropy
  - Mutual information
  - KL-divergence

Next:

- Statistical estimation and regression (again)

## Further reading

- The original article from Shannon (1948), which started the field, is also quite easy to read
- MacKay (2003) covers most of the topics discussed, in a way quite relevant to machine learning. The complete book is available freely online (see the link below)

 MacKay, David J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511526188> <http://www.inference.gy/~cam.ac.uk/courses/mack3.html>.

 Shannon, Claude E. (1948). "A mathematical theory of communication". In: *Bell System Technical Journal* 27, pp. 379-423, 623-656.