A refresher on information theory

Statistical Natural Language Processing

Çağrı Çöltekin

University of Tübingen
Seminar für Sprachwissenschaft

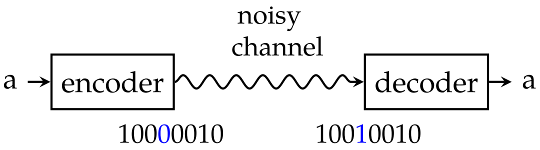
Summer Semester 2021

version: 2dc8350 @2021-04-30

Information theory

- 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

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

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

Coding example

binary coding of an eight-letter alphabet

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

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

Coding example

binary coding of an eight-letter alphabet

- 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?

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

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
- In most systems, linear increase in capacity increases possible outcomes exponentially
  - Number of possible word combinations in a two-word sentence is exponentially more than the number of possible words in a one-word sentence
  - But we expect information to increase linearly, not exponentially
- Working with logarithms is mathematically and computationally more suitable

Entropy

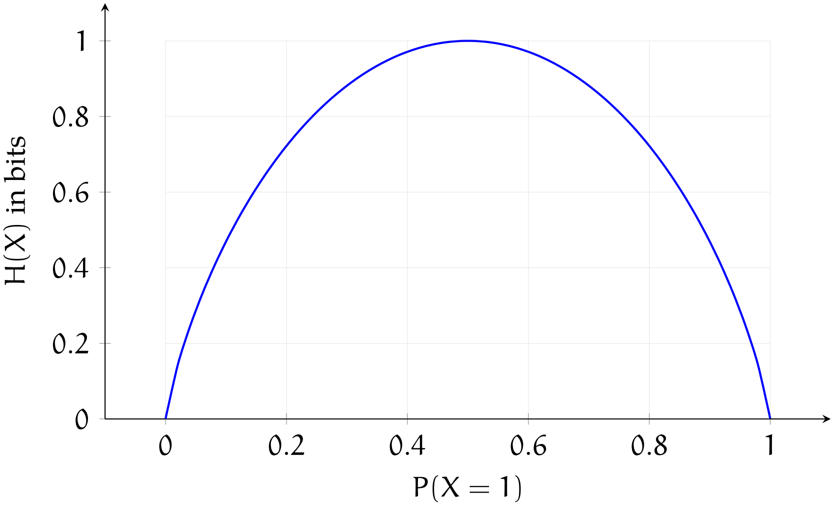
Entropy is a measure of the uncertainty of a random variable:

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

- 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)

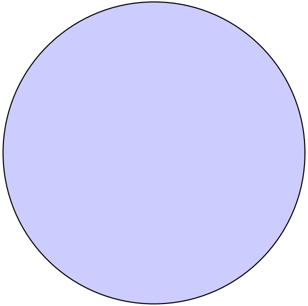
Entropy is about a distribution, while surprisal is about individual events

Example: entropy of a Bernoulli distribution



Entropy: demonstration

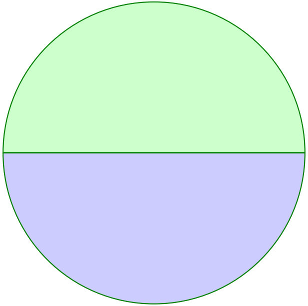
increasing number of outcomes increases entropy



$$H = -\log 1 = 0$$

Entropy: demonstration

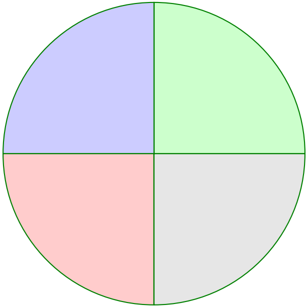
increasing number of outcomes increases entropy



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

Entropy: demonstration

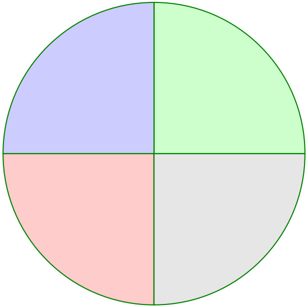
increasing number of outcomes increases entropy



?

Entropy: demonstration

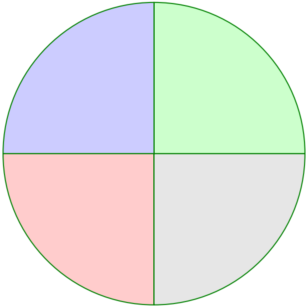
increasing number of outcomes increases entropy



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

Entropy: demonstration

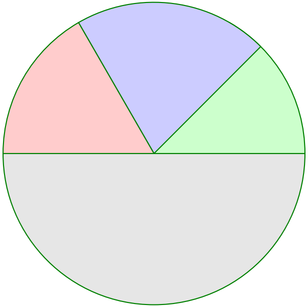
the distribution matters



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

Entropy: demonstration

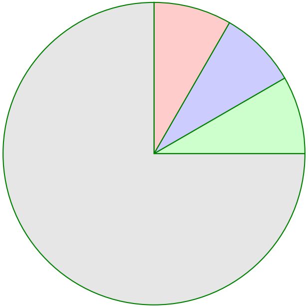
the distribution matters



$$H = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{6}\log_2\frac{1}{6} - \frac{1}{6}\log_2\frac{1}{6} - \frac{1}{6}\log_2\frac{1}{6} = 1.79$$

Entropy: demonstration

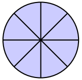
the distribution matters



$$H = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{12}\log_2\frac{1}{12} - \frac{1}{12}\log_2\frac{1}{12} - \frac{1}{12}\log_2\frac{1}{12} = 1.21$$

Back to coding letters

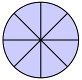
- Can we do better?



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

Back to coding letters

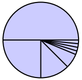
- Can we do better?
- No. $H = 3$ bits, we need 3 bits on average



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

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?

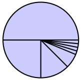


letter	prob	code
a	$\frac{1}{2}$	
b	$\frac{1}{4}$	
c	$\frac{1}{8}$	
d	$\frac{1}{16}$	
e	$\frac{1}{64}$	
f	$\frac{1}{64}$	
g	$\frac{1}{64}$	
h	$\frac{1}{64}$	

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

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



letter	prob	code
a	$\frac{1}{2}$	0
b	$\frac{1}{4}$	10
c	$\frac{1}{8}$	110
d	$\frac{1}{16}$	1110
e	$\frac{1}{64}$	111100
f	$\frac{1}{64}$	111101
g	$\frac{1}{64}$	111110
h	$\frac{1}{64}$	111111

Differential entropy

- Information entropy 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*

Pointwise mutual information

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

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

- Reminder: $P(x, y) = P(x)P(y)$ if two events are independent

Pointwise mutual information

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

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

- Reminder: $P(x, y) = P(x)P(y)$ if two events are independent PMI
  - 0 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 $\mathrm{PMI}(X, Y) = \mathrm{PMI}(Y, X)$
- 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 the non-linear dependence

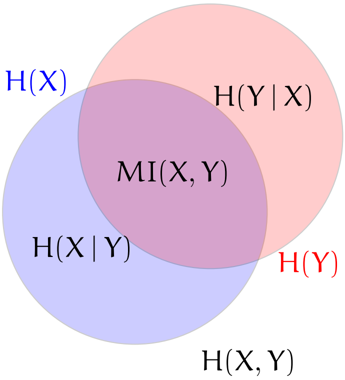
Conditional entropy

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

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

- $H(X \mid Y) = H(X)$ 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*.

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)

Short divergence: distance measure

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

- $d(x, y) \geqslant 0$
- $d(x, y) = d(y, x)$
- $d(x, y) = 0 \iff x = y$
- $d(x, y) \leqslant d(x, z) + d(z, y)$

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

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:

Mon	ML intro / regression
Wed	Classification
Fri	Classification

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. ISBN: 978-05-2164-298-9. URL: http://www.inference.phy.cam.ac.uk/itprnn/book.html.

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