ILLINOIS INSTITUTE OF TECHNOLOGY

Mathematics Review (2)

CS-585

Natural Language Processing

Sonjia Waxmonsky

Slides based in part on material from Derrick Higgins (IIT)

INFORMATION THEORY REVIEW

Information Theory

- Developed in the 1940s by Claude Shannon, mathematician and cryptographer
- Concerned with the optimal compression of information for communication over a channel with limited capacity
- Basic measure of information is *bits*—the number of binary 1/0 indicators used to encode a value

Encoding Random Variables

One coin → [0] → 1 bit

Three coins → [1] [0] [0] → 3 bits

A 6-sided die → 2 or 3 bits ~ $\log_2 6$ bits

1:100, 2:101, 3:110, 4:111, 5:00, 6:01

Three 6-sided dice → $3 \log_2 6$ bits

Five 20-sided dice → $5 \log_2 20$ bits

Information Content

- Generally, the information content or optimal code length of an event drawn from a distribution with N equiprobable outcomes is

$$-\log_2 \frac{1}{N} = \log_2 N \text{ bits}$$

- The information content of an event e drawn from a distribution P(X) over a discrete random variable X is

$$-\log_2 P(X = e) \text{ bits}$$

Bits and "Nats"

- In information theory, we generally use base-2 logs because it makes information values interpretable as the number of 0/1 bits of information we use to encode data for computers

$$-\log_2 P(X = e) \text{ bits}$$

- But an alternative unit using the natural logarithm is *nats*:

$$-\ln P(X = e) \text{ nats}$$

- To convert from bits to nats, divide by $\log_2 e$:

$$-\log_2 P = -\log_2 e^{\ln P}$$

$$-\log_2 P = -\ln P \times \log_2 e$$

$$-\frac{\log_2 P}{\log_2 e} = -\ln P$$

Entropy of a Random Variable

- Entropy (self-information) of a discrete random variable X is

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

- Optimal code length for $X = x$:

$$-\log_2 P(X = x) \text{ (bits)}$$
$$-\ln P(X = x) \text{ (nats)}$$

Entropy example

- $$\begin{aligned} H(\langle 0.5,0.5\rangle) &= -E[\log_2\langle 0.5,0.5\rangle] \\ &= -\frac{1}{2}\log_2(0.5)-\frac{1}{2}\log_2(0.5) \\ &= 1 \end{aligned}$$

$$\begin{aligned} H\left(\left\langle\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}\right\rangle\right) &= -E\left[\log_2\left\langle\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}\right\rangle\right] \\ &= -\sum_{1}^{6}\frac{1}{6}\log_2\left(\frac{1}{6}\right) \\ &= \log_2 6 \\ &= 2.58 \end{aligned}$$

Entropy example

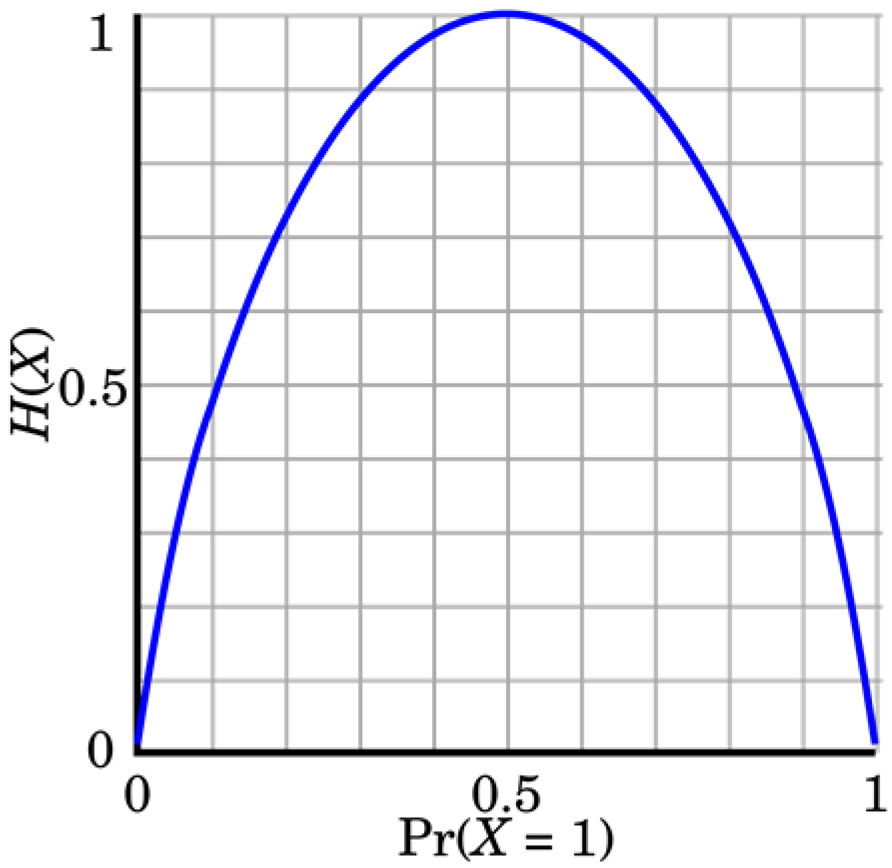
- $$H(\langle .1, .7, .15, .05 \rangle) = -E[\log_2 \langle .1, .7, .15, .05 \rangle]$$
  $$= \begin{array}{c} -.1 \log_2(.1) - .7 \log_2(.7) \\ -.15 \log_2(.15) - .05 \log_2(.05) \end{array}$$
  $$= .33 + .36 + .41 + .22 = 1.32$$

- Lower entropy than we would get for a uniform distribution $\langle 0.25, 0.25, 0.25, 0.25 \rangle$ (which would be 2 bits)

Entropy of a weighted coin

- Think of entropy as uncertainty

- For a Bernoulli distribution, the ***uncertainty is maximized*** when both outcomes are equiprobable



X: Coin is Heads

https://en.wikipedia.org/wiki/Entropy_(information_theory)#/media/File:Binary_entropy_plot.svg

Entropy of a weighted coin

[1] Fair coin (not weighted!)

P(X=0)=P(X=1)=50%

H(X) = -0.5 $\log_2$0.5 + -0.5 $\log_2$0.5

= 0.5 + 0.5 = 1.0

[2] A weighted coin

P(X=0)=25%

P(X=1)=75%

H(X) = -0.25 $\log_2$ 0.25 + -$\log_2$0.75

= 0.5 + 0.31 = 0.81

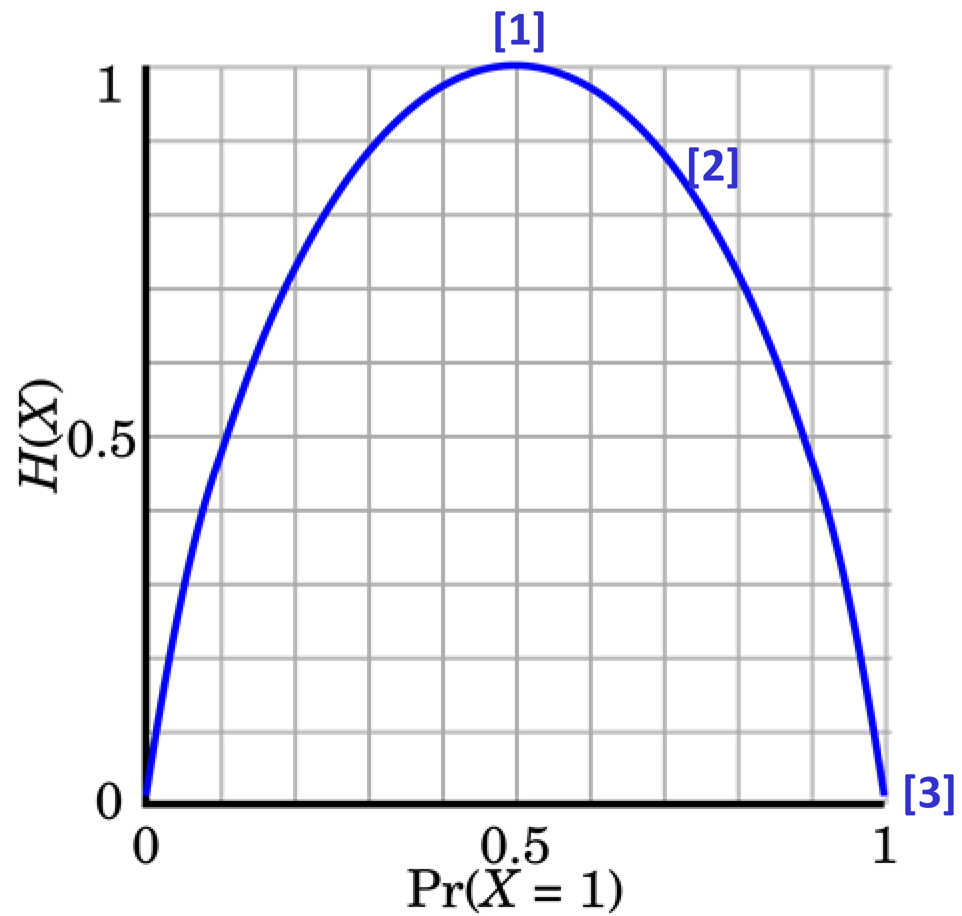
[3] Two heads? (not random!)

P(X=0)=0%

P(X=1)=100%

H(X) = 0 $\log_2$ 0 + 1 $\log_2$1

= 0+0 = 0 ← no information



X: Coin is Heads

https://en.wikipedia.org/wiki/Entropy_(information_theory)#/media/File:Binary_entropy_plot.svg

The Entropy of English

- We can think of a language as an orthographic symbol generation process governed by some unknown probability distribution $P_{lang}(X)$
- What is $H(P_{lang}(X))$?
- How uncertain/unpredictable is the next symbol in a text from a given language?

The Entropy of English

Character-level entropy:

- Assume 27 equally likely symbols (a-z and space):

X: {a,b,c, …, x,y,z, <space>}

$$H(X) = \log_2 27 = 4.76 \text{ bits}$$

- BUT characters in English are not uniformly distributed
- Estimated entropy: 4.03 bits per letter (per symbol) based on observed **unigram** probabilities
- Additional gains: leverage "redundancy" of language and n-gram probabilities (e.g. "qu")

https://www.princeton.edu/~wbialek/rome/refs/shannon_51.pdf

Optimal Coding

- We know that the optimal code length for message m drawn from distribution X is $\log P(X = \mathrm{m})$, but how to construct code that approximates this bound?
- Multiple algorithms:
  - Huffman coding
  - Arithmetic coding
  - Hu-Tucker coding

Entropy Rate of a Message

We can compute the entropy **rate** of a message as the sum of the information content of its symbols.

For a message X of length n:

$$H_{\text{rate}} = \frac{1}{n} H(X_{1n}) = -\frac{1}{n} \sum_{x_{1n}} \mathrm{p}(x_{1n}) \log \mathrm{p}(x_{1n})$$

This allows us to normalize for message length.

Huffman Coding

Huffman Coding:

- Based on binary tree
- Builds prefix-code:
  - No code is a prefix of any other code
  - Allows for variable-length code

P(X) = 1/6

```
'1' →100
'2' →101
'3' →110
'4' →111
'5' →00
'6' →01
```

[Demo: Notebook]

Huffman Coding: Algorithm

1. Create a leaf node for each symbol and add it to the priority queue.
2. While there is more than one node in the queue:
   1. Remove the two nodes of highest priority (lowest probability) from the queue
   2. Create a new internal node with these two nodes as children and with probability equal to the sum of the two nodes' probabilities.
   3. Add the new node to the queue.
3. The remaining node is the root node, and the tree is complete.

https://en.wikipedia.org/wiki/Huffman_coding

Cross-entropy

- If *entropy* is the information in bits required to represent a message using an optimal encoding derived from the **true** distribution…
- *Cross-entropy* is the information in bits required to represent a message using an optimal encoding derived from a **different** distribution
- We encountered this already when bounding the entropy of English
- Cross-entropy is always an **upper bound** on the entropy

Cross-entropy

- The cross-entropy between two distributions *P* and *Q* (where Q is often a model of the true distribution *P*) is

$$H\big(P(X), Q(X)\big) = -E_{P(X)}[\log_2 Q(X)]$$
$$= -\sum_{x \in X} P(X = x) \log_2 Q(X = x)$$

- This is the expected number of bits required to encode messages from P using an encoding system from Q, and is <u>not</u> symmetric

$$H(P, Q) \neq H(Q, P)$$
$$H(P, Q) \geq H(P)$$

Cross-entropy and perplexity

- Many speech recognition and language modeling tasks use *perplexity*, rather than cross-entropy as an evaluation measure

$$Perplexity\big(P(X), Q(X)\big) = 2^{H(P(X),Q(X))}$$

- For a sequence of observations (words, characters), the perplexity is just

$$\prod_{i=1..n} Q(X_i = x_i)^{-1}$$

Where *n* is the length of the sequence

- Perplexity is the inverse of the probability of the sequence under the model

Conditional Entropy

- How related are two random variables X and Y to one another?
- In information-theoretic terms, how efficiently can you encode X given the value of Y?
- This is the conditional entropy:

$$H(X|Y) = \sum_{y \in Y} P(Y = y) H(X|Y = y)$$

Mutual Information

- The difference between the entropy *H(X)* and the conditional entropy *H(X|Y)* is called the mutual information between the two random variables:

$$I(X;Y) = H(X) - H(X|Y)$$

- When *Y* provides no information about *X*, $I(X;Y) = 0$
- When *Y* provides complete information about *X*, $I(X;Y) = H(X)$
- Mutual information is symmetric:

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

Distributional Similarity Measures

- How different are two distributions $P(X)$ and $Q(X)$?
  - We looked at cross-entropy, which tells us how efficient a coding system designed for one distribution is for encoding a different distribution
  - But cross-entropy depends on the entropy of the distribution to be encoded:

$$H(P,Q) \geq H(P)$$

Distributional Similarity Measures: KL Divergence

- Solution: measure the incremental encoding length, rather then the encoding length directly.
- This measure is the Kullback-Leibler (KL) Divergence
- It is defined as the cross-entropy minus the entropy of the distribution to be encoded:

$$D_{KL}\big(P(X) \parallel Q(X)\big) = H(P(X), Q(X)) - H\big(P(X)\big)$$

$$= -\textstyle\sum_{x \in X} P(X=x) \log_2 Q(X=x) + \textstyle\sum_{x \in X} P(X=x) \log_2 P(X=x)$$

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

$$= -\sum_{x \in X} P(X=x)\left(\log_2 \frac{Q(X=x)}{P(X=x)}\right)$$

KL Divergence example

$$P(X) = \langle .1, .5, .4 \rangle$$

$$Q(X) = \langle .2, .2, .6 \rangle$$

$$\begin{aligned}
D_{KL}(P \parallel Q) &= -\sum_{x \in X} P(X = x) \left(\log_2 \frac{Q(X = x)}{P(X = x)} \right) \\
&= -.1 \log_2 (\frac{.2}{.1}) - .5 \log_2 (\frac{.2}{.5}) - .4 \log_2 (\frac{.6}{.4}) \\
&= -0.1 + 0.66 - .23 = \mathbf{0.33}
\end{aligned}$$

KL Divergence example

$$P(X) = \langle .1, .5, .4 \rangle$$

$$Q(X) = \langle .2, .2, .6 \rangle$$

$$\begin{aligned} D_{KL}(Q \parallel P) &= -\sum_{x \in X} Q(X = x) \left(\log_2 \frac{P(X = x)}{Q(X = x)} \right) \\ &= -.2 \log_2 (\frac{.1}{.2}) - .2 \log_2 (\frac{.5}{.2}) - .6 \log_2 (\frac{.4}{.6}) \\ &= 0.2 - 0.26 + 0.35 = \mathbf{0.29} \end{aligned}$$

Distributional Similarity Measures: JS Divergence

- KL Divergence is not symmetric:

$$D_{KL}\big(P(X) \parallel Q(X)\big) \neq D_{KL}\big(Q(X) \parallel P(X)\big)$$

- A commonly-used symmetric measure of distributional distance is the Jensen-Shannon (JS) Divergence:

$$M(X) \stackrel{\text{def}}{=} \frac{(P(X) + Q(X))}{2} \quad \leftarrow \text{Mixture Distribution}$$

$$D_{JS}(P(X) \parallel Q(X)) = \frac{D_{KL}\big(P(X) \parallel M(X)\big) + D_{KL}(Q(X) \parallel M(X))}{2}$$

- Why not $\frac{D_{KL}(P\parallel Q)+D_{KL}(Q\parallel P)}{2}$?

JS Divergence example

$$P(X) = \langle .1, .5, .4 \rangle$$

$$Q(X) = \langle .2, .2, .6 \rangle$$

$$M(X) = \frac{P(X) + Q(X)}{2} = \langle .15, .35, .5 \rangle$$

$$\begin{aligned}
D_{KL}(P \parallel M) &= -\sum_{x \in X} P(X = x) \left(\log_2 \frac{M(X = x)}{P(X = x)} \right) \\
&= -.1 \log_2\left(\frac{.15}{.1}\right) - .5 \log_2\left(\frac{.35}{.5}\right) - .4 \log_2\left(\frac{.5}{.4}\right) \\
&= -0.058 + 0.257 - 0.129 = 0.070
\end{aligned}$$

JS Divergence example

$$P(X) = \langle .1, .5, .4 \rangle$$

$$Q(X) = \langle .2, .2, .6 \rangle$$

$$M(X) = \frac{P(X) + Q(X)}{2} = \langle .15, .35, .5 \rangle$$

$$\begin{aligned} D_{KL}(Q \parallel M) &= -\sum_{x \in X} Q(X = x) \left(\log_2 \frac{M(X = x)}{Q(X = x)} \right) \\ &= -.2 \log_2(\frac{.15}{.2}) - .2 \log_2(\frac{.35}{.2}) - .6 \log_2(\frac{.5}{.6}) \\ &= 0.083 - 0.161 + 0.158 = 0.079 \end{aligned}$$

JS Divergence example

$$\begin{aligned} D_{JS}(P \parallel Q) &= \frac{D_{KL}(P \parallel M) + D_{KL}(Q \parallel M)}{2} \\ &= \frac{0.70 + 0.79}{2} \\ &\approx 0.75 \end{aligned}$$

$$\begin{aligned} D_{JS}(Q \parallel P) &= \frac{D_{KL}(Q \parallel M) + D_{KL}(P \parallel M)}{2} \\ &= \frac{0.79 + 0.70}{2} \\ &\approx 0.75 \quad \leftarrow \text{Results Match} \end{aligned}$$

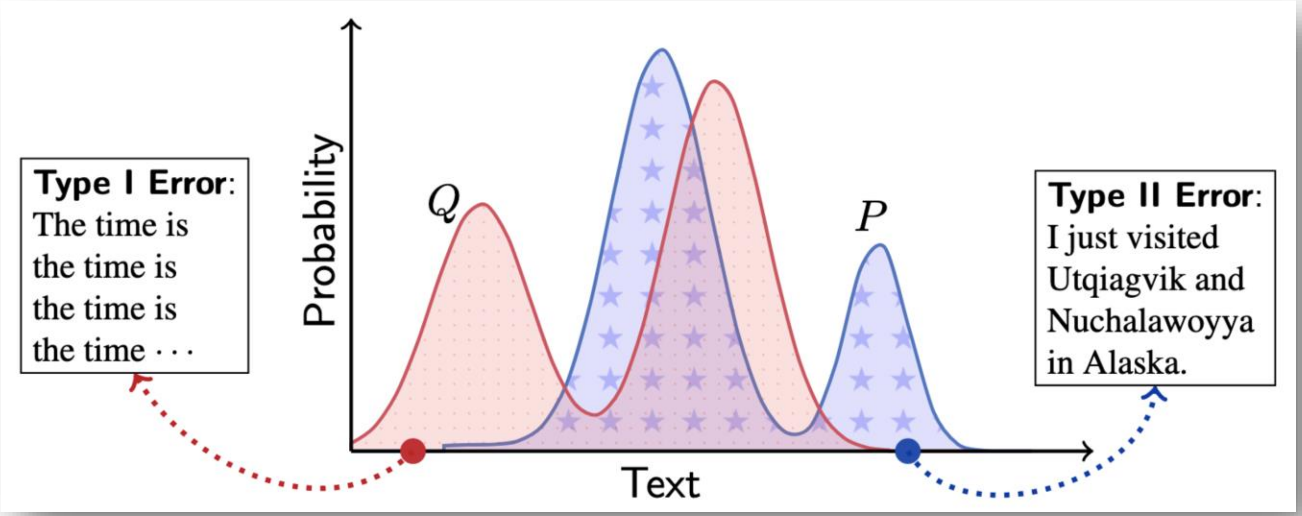
Application of KL Divergence

MAUVE: Measuring the Gap Between Neural Text and Human Text (Pillutla et al, 2021)

P: Human text
Q: Machine text

Type I Error:
Machine text is not plausible human language

Type II Error:
Machine is unable to generate plausible human text



https://krishnap25.github.io/mauve/