Language Models

Natural Language Processing: Jordan Boyd-Graber

University of Maryland

Cross Entropy

How do you know what a good LM is?

Outline

- Math Review: Entropy
- The Shannon Game
- What Makes for a Good Language Model

Expectation

An *expectation* of a random variable is a weighted average:

$$\begin{aligned} \mathrm{E}[f(X)] &= \sum_{x=1}^{\infty} f(x)\, p(x) && \text{(discrete)} \\ &= \int_{-\infty}^{\infty} f(x)\, p(x)\, dx && \text{(continuous)} \end{aligned}$$

Expectation

Expectations of constants or known values:

- $\mathrm{E}[a] = a$
- $\mathrm{E}[Y \mid Y = y] = y$

Expectation

Example: Gaussian distribution $X \sim N(\mu, \sigma^2)$

$$
\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\
&=
\end{aligned}
$$

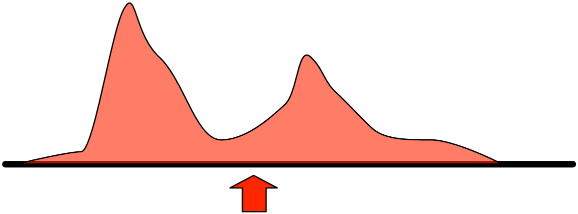
Expectation

Example: Gaussian distribution $X \sim N(\mu, \sigma^2)$

$$\begin{aligned} \mathrm{E}[X] &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx \\ &= \mu \end{aligned}$$

Expectation Intuition

- Average or outcome (might not be an event: 2.4 children)
- Center of mass



- "Fair Price" of a wager

Expectation of die / dice

What is the expectation of the roll of die?

Expectation of die / dice

What is the expectation of the roll of die?

One die

$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} =$

Expectation of die / dice

What is the expectation of the roll of die?

One die

$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

Expectation of die / dice

What is the expectation of the roll of die?

One die

$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

What is the expectation of the sum of two dice?

Expectation of die / dice

What is the expectation of the roll of die?

One die

$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

What is the expectation of the sum of two dice?

Two die

$2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} =$

Expectation of die / dice

What is the expectation of the roll of die?

One die

$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

What is the expectation of the sum of two dice?

Two die

$2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7$

Expectation of die / dice

What is the expectation of the roll of die?

One die

$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$

What is the expectation of the sum of two dice?

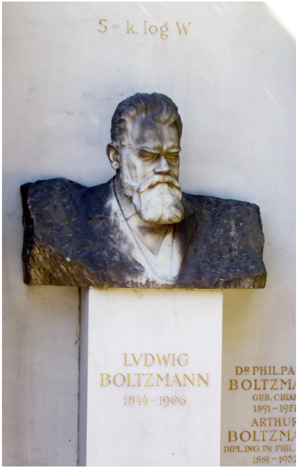
Two die

$2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + 5 \cdot \frac{4}{36} + 6 \cdot \frac{5}{36} + 7 \cdot \frac{6}{36} + 8 \cdot \frac{5}{36} + 9 \cdot \frac{4}{36} + 10 \cdot \frac{3}{36} + 11 \cdot \frac{2}{36} + 12 \cdot \frac{1}{36} = 7$

In general, the expected value of a sum of random variables in the sum of the expected values.

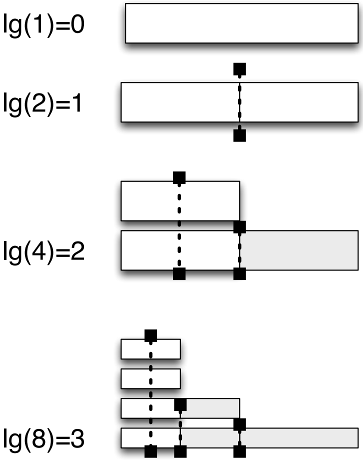
Entropy

- Measure of disorder in a system
- In the real world, entroy in a system tends to increase
- Can also be applied to probabilities:
  - Is one (or a few) outcomes certain (low entropy)
  - Are things equiprobable (high entropy)
- In data science
  - We look for features that allow us to <u>reduce</u> entropy (decision trees)
  - All else being equal, we seek models that have <u>maximum</u> entropy (Occam's razor)



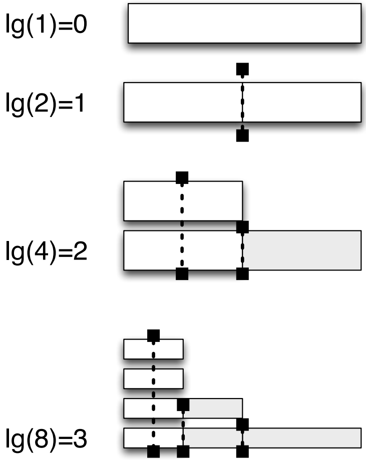
Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- Way to think about them: cutting a carrot



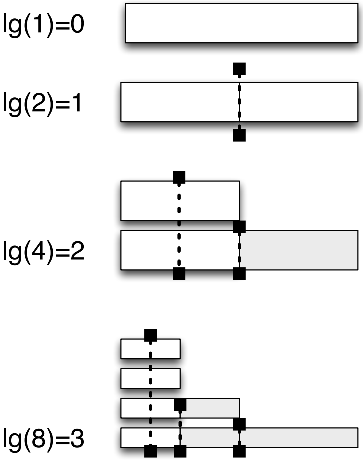
Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- Way to think about them: cutting a carrot
- **Fractions?**



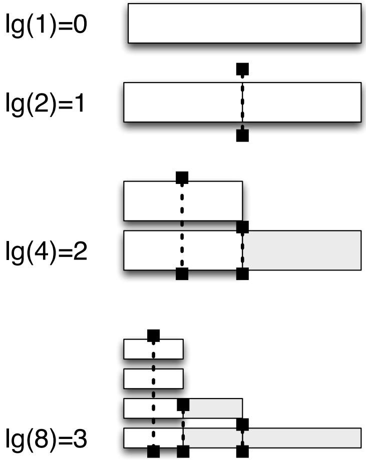
Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- Way to think about them: cutting a carrot
- **Fractions?** Makes them negative: you need to glue the carrot back together.



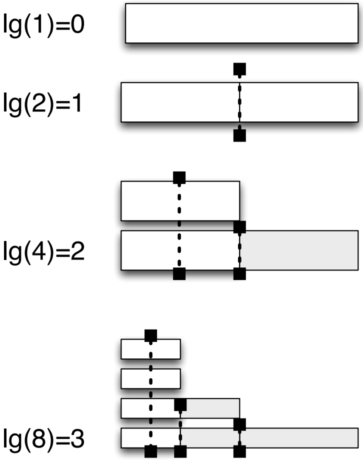
Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- Way to think about them: cutting a carrot
- **Fractions?** Makes them negative: you need to glue the carrot back together.
- **Non-integers?**



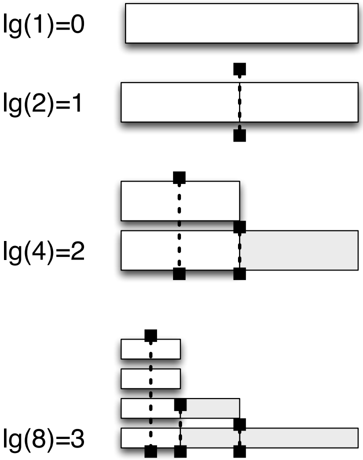
Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- Way to think about them: cutting a carrot
- **Fractions?** Makes them negative: you need to glue the carrot back together.
- **Non-integers?** Interpolates between integer values.



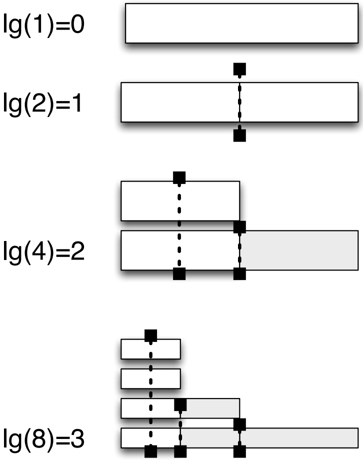
Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- Way to think about them: cutting a carrot
- **Fractions?** Makes them negative: you need to glue the carrot back together.
- **Non-integers?** Interpolates between integer values.
- **Negative numbers?**



Aside: Logarithms

- $\lg(x) = b \Leftrightarrow 2^b = x$
- Way to think about them: cutting a carrot
- **Fractions?** Makes them negative: you need to glue the carrot back together.
- **Non-integers?** Interpolates between integer values.
- **Negative numbers?** Not defined.



Entropy

Entropy is a measure of uncertainty that is associated with the distribution of a random variable:

$$
\begin{aligned}
H(X) &= -\mathrm{E}\left[\lg(p(X))\right] \\
&= -\sum_{x} p(x)\,\lg(p(x)) && \text{(discrete)} \\
&= -\int_{-\infty}^{\infty} p(x)\,\lg(p(x))\,dx && \text{(continuous)}
\end{aligned}
$$

Entropy

Entropy is a measure of uncertainty that is associated with the distribution of a random variable:

$$\begin{aligned} H(X) &= -\mathrm{E}[\lg(p(X))] \\ &= -\sum_x p(x)\lg(p(x)) && \text{(discrete)} \\ &= -\int_{-\infty}^{\infty} p(x)\lg(p(x))\,dx && \text{(continuous)} \end{aligned}$$

Does not account for the values of the random variable, only the spread of the distribution.

- $H(X) \geq 0$
- uniform distribution $=$ highest entropy, point mass $=$ lowest
- suppose $P(X=1)=p,\ P(X=0)=1-p$ and $P(Y=100)=p,\ P(Y=0)=1-p$: X and Y have the same entropy

What does this have to do with language?

Prediction and Entropy of Printed English

By C. E. SHANNON

(*Manuscript Received Sept. 15, 1950*)

A new method of estimating the entropy and redundancy of a language is described. This method exploits the knowledge of the language statistics possessed by those who speak the language, and depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given, and some properties of an ideal predictor are developed.

Foundation of Information Theory

first line is the original text and the numbers in the second line indicate the guess at which the correct letter was obtained.

(1) T H E R E I S N O R E V E R S E O N A M O T O R C Y C L E A
(2) 1 1 1 5 1 1 2 1 1 2 1 1 15 1 1 7 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1 1 1 3 1
(1) F R I E N D O F M I N E F O U N D T H I S O U T
(2) 8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1
(1) R A T H E R D R A M A T I C A L L Y T H E O T H E R D A Y
(2) 4 1 1 1 1 1 1 1 1 5 1 1 1 1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 1 1 1 1 1 1 1 (9)

Reminder of Entropy

$$\mathbb{H}[X] \equiv -\sum_{x \in \mathcal{X}} p(x) \log p(x) = \mathbb{E}[-\log p(X)] \tag{1}$$

first line is the original text and the numbers in the second line indicate the guess at which the correct letter was obtained.

(1) T H E R E   I S   N O   R E V E R S E   O N   A   M O T O R C Y C L E   A
(2) 1 1 1 5 1 1 2 1 1 2 1 1 15 1 17 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1 1 1 3 1
(1) F R I E N D   O F   M I N E   F O U N D   T H I S   O U T
(2) 8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 2 1 1 1 1 1 1
(1) R A T H E R   D R A M A T I C A L L Y   T H E   O T H E R   D A Y
(2) 4 1 1 1 1 1 1 1 1 5 1 1 1 1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 1 1 1 1 1 1   (9)

Reminder of Entropy

$$\mathbb{H}[X] \equiv -\sum_{x \in \mathcal{X}} p(x) \log p(x) = \mathbb{E}[-\log p(X)] \tag{1}$$

Shannon's trick: Getting from the number probability of needing *i*

first line is the original text and the numbers in the second line indicate the guess at which the correct letter was obtained.

```
(1) T H E R E   I S   N O   R E V E R S E   O N   A   M O T O R C Y C L E   A
(2) 1 1 1 5 1 1 2 1 1 2 1 1 15 1 17 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1 1 1 3 1
(1) F R I E N D   O F   M I N E   F O U N D   T H I S   O U T
(2) 8 6 1 3 1 1 1 1 1 1 1 1 1 1 1 6 2 1 1 1 1 1 1 2 1 1 1 1 1 1
(1) R A T H E R   D R A M A T I C A L L Y   T H E   O T H E R   D A Y
(2) 4 1 1 1 1 1 1 1 1 5 1 1 1 1 1 1 1 1 1 1 1 6 1 1 1 1 1 1 1 1 1 1 1 1     (9)
```

Reminder of Entropy

$$\mathbb{H}[X] \equiv -\sum_{x \in \mathscr{X}} p(x) \log p(x) = \mathbb{E}[-\log p(X)] \tag{1}$$

Shannon's trick: Getting from the number probability of needing i

Relationship between guesses and entropy

Upper Bound

$$\mathbb{H}[X] \leq -\sum_{i=1}^{27} q_i \log q_i \tag{2}$$

Relationship between guesses and entropy

Upper Bound

$$\mathbb{H}[X] \leq -\sum_{i=1}^{27} q_i \log q_i \tag{2}$$

Max entropy for number of errors is the same as the underlying language: The sums involved will contain precisely the same terms although, perhaps, in a different order.

Relationship between guesses and entropy

Upper Bound

$$\mathbb{H}[X] \leq -\sum_{i=1}^{27} q_i \log q_i \tag{2}$$

Lower Bound

$$\sum_{i=1}^{27} i(q_i - q_{i+1}) \log i \leq \mathbb{H}[X] \tag{3}$$

Relationship between guesses and entropy

Upper Bound

$$\mathbb{H}[X] \leq -\sum_{i=1}^{27} q_i \log q_i \tag{2}$$

Lower Bound

$$\sum_{i=1}^{27} i(q_i - q_{i+1}) \log i \leq \mathbb{H}[X] \tag{3}$$

Requires rectangular decomposition of a monotonic distribution

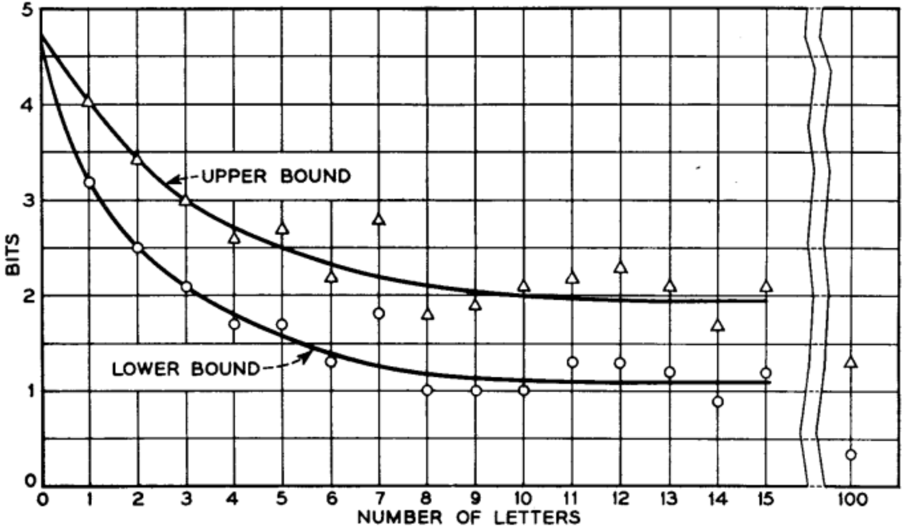


Fig. 4—Upper and lower experimental bounds for the entropy of 27-letter English.

Shannon's Experimental Bounds

How’d they do it?

- Word frequency lists
- World knowledge
- Bigram frequency lists

What Does a Language Model Do?

In the Shannon Game, I show you a bit of English, and you

What Does a Language Model Do?

In the Shannon Game, I show you a bit of English, and you n

What Does a Language Model Do?

In the Shannon Game, I show you a bit of English, and you ne

What Does a Language Model Do?

In the Shannon Game, I show you a bit of English, and you need to tell me what the

What Does a Language Model Do?

In the Shannon Game, I show you a bit of English, and you need to tell me what the n

What Does a Language Model Do?

In the Shannon Game, I show you a bit of English, and you need to tell me what the next character is going to

What Does a Language Model Do?

In the Shannon Game, I show you a bit of English, and you need to tell me what the next character is going to be

Most LMs today don’t predict characters

Most LMs today don’t predict characters

I have a sad story to tell you
It may hurt your feelings a _______

Most LMs today don’t predict characters

I have a sad story to tell you
It may hurt your feelings a bit

Most LMs today don't predict characters

I have a sad story to tell you
It may hurt your feelings a bit
Last night I walked into my _______

Most LMs today don’t predict characters

I have a sad story to tell you
It may hurt your feelings a bit
Last night I walked into my bathroom

Most LMs today don't predict characters

I have a sad story to tell you
It may hurt your feelings a bit
Last night I walked into my bathroom
And stepped in a pile of ________

Hints

- Something you can have a "pile" of
- Something in a bathroom
- Rhymes with bit

Most LMs today don’t predict characters

I have a sad story to tell you
It may hurt your feelings a bit
Last night I walked into my bathroom
And stepped in a pile of ________

Hints

- Something you can have a “pile” of
- Something in a bathroom
- Rhymes with bit

Most LMs today don't predict characters

I have a sad story to tell you
It may hurt your feelings a bit
Last night I walked into my bathroom
And stepped in a pile of ________

Hints

- Something you can have a "pile" of
- Something in a bathroom
- Rhymes with bit

Most LMs today don't predict characters

I have a sad story to tell you
It may hurt your feelings a bit
Last night I walked into my bathroom
And stepped in a pile of ________

Hints

- Something you can have a "pile" of
- Something in a bathroom
- Rhymes with bit

Most LMs today don't predict characters

I have a sad story to tell you
It may hurt your feelings a bit
Last night I walked into my bathroom
And stepped in a pile of ________

Hints

- Something you can have a "pile" of
- Something in a bathroom
- Rhymes with bit

Most LMs today don't predict characters

I have a sad story to tell you
It may hurt your feelings a bit
Last night I walked into my bathroom
And stepped in a pile of shaving cream

Hints

- Something you can have a "pile" of
- Something in a bathroom
- Rhymes with bit

Can’t Use Accuracy

- Take a sequence: a b c d e f g

Can't Use Accuracy

- Take a sequence: a b c d e f g
- Get predictions: a b b d e g g

Can't Use Accuracy

- Take a sequence: a b c d e f g
- Get predictions: a b b d e g g
- Accuracy: $\frac{5}{7}$

We're really comparing two distributions!

- If language model only predicts articles, pronouns, and prepositions accuracy is going to be pretty good
- We care about the whole distribution, and predicting rare things well should be rewarded
- There isn't really one real answer: there's a distribution over answers

Intuition

What comes next?

I came home from a long day of work and sat on my...

Intuition

What comes next?

I came home from a long day of work and sat on my...

Intuition

What comes next?

I came home from a long day of work and sat on my...

Good guesses

- Sofa
- Butt
- Recliner
- Couch

Bad guesses

- Excavator
- Workstation
- Spaceship
- Jumpseat

Thus, we need tools to compare distributions

- Cross-Entropy
- KL-Divergence
- Perplexity

Thus, we need tools to compare distributions

- Cross-Entropy
- KL-Divergence
- Perplexity

We'll use this the most, but others will pop up from time to time

Entropy vs. Cross-Entropy

Entropy

$$\mathbb{H}_p[X] = -\sum_x p(x) \log p(x) \quad (4)$$

Number of bits you need to encode the distribution's complexity

Cross Entropy

$$\mathbb{H}_{(p,q)}[X] \equiv -\sum_x p(x) \log q(x) \quad (5)$$

How many bits you need to encode the distribution if you use q instead

Entropy vs. Cross-Entropy

Entropy

$$\mathbb{H}_p[X] = -\sum_x p(x) \log p(x) \quad (4)$$

Number of bits you need to encode the distribution's complexity

Cross Entropy

$$\mathbb{H}_{(p,q)}[X] \equiv -\sum_x p(x) \log q(x) \quad (5)$$

How many bits you need to encode the distribution if you use q instead

Cross-entropy is lower bounded by underlying entropy: $\mathbb{H}_{(p,q)}[\cdot] \geq \mathbb{H}_p[\cdot]$

Kullback-Leibler (KL) Divergence

$$\mathrm{KL}(p||q) \equiv \mathbb{H}_q[X] - \mathbb{H}[X] \tag{6}$$

$$\tag{7}$$

Kullback-Leibler (KL) Divergence

$$\mathrm{KL}(p||q) \equiv \mathbb{H}_q[X] - \mathbb{H}[X] \tag{6}$$

$$= -p(x)\log q(x) - \left(-\sum_x p(x)\log p(x)\right) \tag{7}$$

$$\tag{8}$$

Expand definitions

Kullback-Leibler (KL) Divergence

$$
\begin{aligned}
\mathrm{KL}(p||q) &\equiv \mathbb{H}_q[X] - \mathbb{H}[X] &(6)\\
&= -p(x)\log q(x) - \left(-\sum_x p(x)\log p(x)\right) &(7)\\
&= p(x)\left[\log p(x) - \log q(x)\right] &(8)\\
& &(9)
\end{aligned}
$$

Factor out p(x), swap log terms

Kullback-Leibler (KL) Divergence

$$
\begin{aligned}
\mathrm{KL}(p||q) &\equiv \mathbb{H}_q[X] - \mathbb{H}[X] && (6)\\
&= -p(x)\log q(x) - \left(-\sum_x p(x)\log p(x)\right) && (7)\\
&= p(x)\left[\log p(x) - \log q(x)\right] && (8)\\
&= p(x)\log\left(\frac{p(x)}{q(x)}\right) && (9)\\
& && (10)
\end{aligned}
$$

Difference of logs in log of quotient

Kullback-Leibler (KL) Divergence

$$
\begin{aligned}
\mathrm{KL}(p||q) &\equiv \mathbb{H}_q[X] - \mathbb{H}[X] && (6)\\
&= -p(x)\log q(x) - \left(-\sum_x p(x)\log p(x)\right) && (7)\\
&= p(x)\left[\log p(x) - \log q(x)\right] && (8)\\
&= p(x)\log\left(\frac{p(x)}{q(x)}\right) && (9)\\
& && (10)
\end{aligned}
$$

When equations are equal, quotient is 1, so log is 0, and thus KL divergence is zero

Perplexity: Fewer numbers, shouldn’t be a standard

Perplexity

$$\text{perplexity}(x_{1:N}, q) = 2^{\mathbb{H}_q[X]} \tag{11}$$

$$\tag{12}$$

Perplexity

$$\text{perplexity}(x_{1:N}, q) = 2^{\mathbb{H}_q[X]} \tag{11}$$

$$\tag{12}$$

q: Your model, assigns a probability to each observation (should use context!)

Perplexity

$$\text{perplexity}(x_{1:N}, q) = 2^{\mathbb{H}_q[X]} \tag{11}$$

$$2^{\sum_{x_i \in X} -\frac{1}{N} \log q(x_i)} \tag{12}$$

$$\tag{13}$$

(Not always an accurate assumption): Assume we observe each sample once, our estimate of $p(x)$

Perplexity

$$\text{perplexity}(x_{1:N}, q) = 2^{\mathbb{H}_q[X]} \tag{11}$$

$$2^{\sum_{x_i \in X} -\frac{1}{N} \log q(x_i)} \tag{12}$$

$$\tag{13}$$

q in our cross-entropy

Perplexity

$$\text{perplexity}(x_{1:N}, q) = 2^{\mathbb{H}_q[X]} \tag{11}$$

$$2^{\sum_{x_i \in X} -\frac{1}{N} \log q(x_i)} \tag{12}$$

$$2^{-\frac{1}{N} \sum_{x_i} \log q(x_i)} \tag{13}$$

$$\tag{14}$$

Move $\frac{1}{N}$ out

Perplexity

$$\begin{aligned}
\text{perplexity}(x_{1:N}, q) = 2^{\mathbb{H}_q[X]} && (11)\\
2^{\sum_{x_i \in X} -\frac{1}{N} \log q(x_i)} && (12)\\
2^{-\frac{1}{N} \sum_{x_i} \log q(x_i)} && (13)\\
&& (14)
\end{aligned}$$

Bases much match

Perplexity

$$\text{perplexity}(x_{1:N}, q) = 2^{\mathbb{H}_q[X]} \tag{11}$$

$$2^{\sum_{x_i \in X} -\frac{1}{N} \log q(x_i)} \tag{12}$$

$$2^{-\frac{1}{N} \sum_{x_i} \log q(x_i)} \tag{13}$$

$$\sqrt[N]{\prod_{i}^{N} \frac{1}{q(x_i)}} \tag{14}$$

Geometric mean of inverse token probabilities, leads to nice numbers

Devil in the Details

- What's the vocabulary? How to handle unknown tokens?
  - Throw out?
  - Use suffix?
  - Special token?
- What's the dataset size?
- How similar is the test set?
- Base of log / exp
- Restart each sentence? (start token)
- Long sentences?

Recap

- Language models predict the next word
- We need metrics to measure how good they are
- We'll mostly use cross-entropy, but you'll see KL divergence (variational inference) and perplexity (papers) quite a bit