Probabilities, probably

Jon Dehdari

November 2, 2015

Good Morning!

Good Morning!

Fifty Shades of Nay

- Last week we discussed (formal) languages and grammars in terms of **set membership**

Fifty Shades of Nay

- Last week we discussed (formal) languages and grammars in terms of **set membership**
- That is, whether a particular sentence was **grammatical** or **ungrammatical**

Fifty Shades of Nay

- Last week we discussed (formal) languages and grammars in terms of **set membership**
- That is, whether a particular sentence was **grammatical** or **ungrammatical**
- This is, of course, an overly simplistic view of natural language

Fifty Shades of Nay

- Last week we discussed (formal) languages and grammars in terms of **set membership**
- That is, whether a particular sentence was **grammatical** or **ungrammatical**
- This is, of course, an overly simplistic view of natural language
- This week, we're going to take a more subtle approach
- The formal languages & grammars are still relevant, but we're going to add **probabilities** to them

Probability

- A **probability** tells you how likely something will happen

Probability

- A **probability** tells you how likely something will happen
- Another way of looking at it is that it is just a score for a possible event

Probability

- A **probability** tells you how likely something will happen
- Another way of looking at it is that it is just a score for a possible event
- The score is between 0 and 1 (the unit interval)
- The scores for all of the possible outcomes must add up to 1 (unity)

Probability

- A **probability** tells you how likely something will happen
- Another way of looking at it is that it is just a score for a possible event
- The score is between 0 and 1 (the unit interval)
- The scores for all of the possible outcomes must add up to 1 (unity)
- Adding up all the scores is called **normalization**

Probability

- A **probability** tells you how likely something will happen
- Another way of looking at it is that it is just a score for a possible event
- The score is between 0 and 1 (the unit interval)
- The scores for all of the possible outcomes must add up to 1 (unity)
- Adding up all the scores is called **normalization**
- A **probability distribution** is the probabilities for all the possible outcomes

Probs, cont'd

- A **uniform distribution** is where all the outcomes have the same probability

Probs, cont'd

- A **uniform distribution** is where all the outcomes have the same probability
- That's rarely true in the real world

Probs, cont'd

- A **uniform distribution** is where all the outcomes have the same probability
- That's rarely true in the real world
- But it's a good starting point, until you know better

Probs, cont'd

- A **uniform distribution** is where all the outcomes have the same probability
- That's rarely true in the real world
- But it's a good starting point, until you know better
- So, for example, suppose I say "It's raining cats and ____" .

Probs, cont'd

- A **uniform distribution** is where all the outcomes have the same probability
- That's rarely true in the real world
- But it's a good starting point, until you know better
- So, for example, suppose I say "It's raining cats and ____" . What word do you think I'll say next?

Probs, cont'd

- A **uniform distribution** is where all the outcomes have the same probability
- That's rarely true in the real world
- But it's a good starting point, until you know better
- So, for example, suppose I say "It's raining cats and ____" . What word do you think I'll say next?
- A uniform distribution would give the same probability for every English word

$$\begin{aligned} p(w_i) &= \frac{1}{|V|} \\ &= |V|^{-1} \end{aligned}$$

Probs, cont'd

- A **uniform distribution** is where all the outcomes have the same probability
- That's rarely true in the real world
- But it's a good starting point, until you know better
- So, for example, suppose I say "It's raining cats and ____" . What word do you think I'll say next?
- A uniform distribution would give the same probability for every English word

$$p(w_i) = \frac{1}{|V|} = |V|^{-1}$$

- Ok, so maybe it's not a very good idea

Unigram Distribution

- In a **unigram distribution** (or unigram model), probabilities are based on word frequencies

Unigram Distribution

- In a **unigram distribution** (or unigram model), probabilities are based on word frequencies

$$p(w_i) = \frac{\text{count}(w_i)}{\text{count}(w)}$$

Unigram Distribution

- In a **unigram distribution** (or unigram model), probabilities are based on word frequencies

$$p(w_i) = \frac{\text{count}(w_i)}{\text{count}(w)}$$

- So the word "the" will have a much higher probability than "dogs"

Unigram Distribution

- In a **unigram distribution** (or unigram model), probabilities are based on word frequencies

$$p(w_i) = \frac{\text{count}(w_i)}{\text{count}(w)}$$

- So the word "the" will have a much higher probability than "dogs"
- It doesn't take into account a word's context

- The probability of two events occurring is called the **joint probability** $p(h, w)$

- The probability of two events occurring is called the **joint probability** $p(h, w)$
- The probability of a **w**ord after a **h**istory of previous words is the **conditional probability** $p(w|h)$

- The probability of two events occurring is called the **joint probability** $p(h, w)$
- The probability of a **w**ord after a **h**istory of previous words is the **conditional probability** $p(w|h)$
- The reverse of that is the **posterior probability** $p(h|w)$

- The probability of two events occurring is called the **joint probability** $p(h, w)$
- The probability of a **w**ord after a **h**istory of previous words is the **conditional probability** $p(w|h)$
- The reverse of that is the **posterior probability** $p(h|w)$
- The **prior probability** tells us how probable the word is, in its own right $p(w)$

- The probability of two events occurring is called the **joint probability** $p(h, w)$
- The probability of a **w**ord after a **h**istory of previous words is the **conditional probability** $p(w|h)$
- The reverse of that is the **posterior probability** $p(h|w)$
- The **prior probability** tells us how probable the word is, in its own right $p(w)$
- **Likelihood** is the probability of the entire data, given our model.

- The probability of two events occurring is called the **joint probability** $p(h, w)$
- The probability of a **w**ord after a **h**istory of previous words is the **conditional probability** $p(w|h)$
- The reverse of that is the **posterior probability** $p(h|w)$
- The **prior probability** tells us how probable the word is, in its own right $p(w)$
- **Likelihood** is the probability of the entire data, given our model. The higher this is, the better it is at predicting the data...

How good are you at guessing?

- Likelihood is usually a really really small number. Why?

How good are you at guessing?

- Likelihood is usually a really really small number. Why?
- People often transform this really small number to a large negative number via the logarithm function ($\log_2$)
- This is then called the **log likelihood**

How good are you at guessing?

- Likelihood is usually a really really small number. Why?
- People often transform this really small number to a large negative number via the logarithm function ($\log_2$)
- This is then called the **log likelihood**
- Not content there, “people” then remove the negative sign, making it a positive number
- Then they’ll divide that number by the number of words in the data. Why?

How good are you at guessing?

- Likelihood is usually a really really small number. Why?
- People often transform this really small number to a large negative number via the logarithm function ($\log_2$)
- This is then called the **log likelihood**
- Not content there, "people" then remove the negative sign, making it a positive number
- Then they'll divide that number by the number of words in the data. Why?
- This gives us the average number of binary choices the model made to predict each outcome in the data, or **cross entropy**

I'm Perplexed

- Cross entropy also tells us how different two distributions are, like the model and the data $H(\theta, \mathbf{x})$

I'm Perplexed

- Cross entropy also tells us how different two distributions are, like the model and the data $H(\theta, \mathbf{x})$
- If we take the binary exponent of the cross entropy (2^H), we get **perplexity**

I'm Perplexed

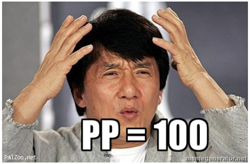
- Cross entropy also tells us how different two distributions are, like the model and the data $H(\theta, \mathbf{x})$
- If we take the binary exponent of the cross entropy (2^H), we get **perplexity**
- This is how confused the model is, on average

I'm Perplexed

- Cross entropy also tells us how different two distributions are, like the model and the data $H(\theta, \mathbf{x})$
- If we take the binary exponent of the cross entropy (2^H), we get **perplexity**
- This is how confused the model is, on average
- In other words, perplexity (PP or PPL) is the average number of words that the model guessed for each trial in the data

I'm Perplexed

- Cross entropy also tells us how different two distributions are, like the model and the data $H(\theta, \mathbf{x})$
- If we take the binary exponent of the cross entropy (2^H), we get **perplexity**
- This is how confused the model is, on average
- In other words, perplexity (PP or PPL) is the average number of words that the model guessed for each trial in the data



Entropy

- **Entropy** measures how predictable the data is

Entropy

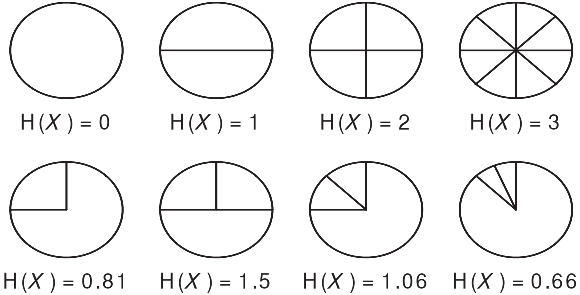
- **Entropy** measures how predictable the data is

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

Entropy

- **Entropy** measures how predictable the data is

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



(courtesy of Koehn, 2010)

Maximum Likelihood Estimation

- **Maximum likelihood estimation** (MLE) just uses counts/frequencies of seen events in the train data
- Why is this type of estimation called *maximum likelihood*?

Maximum Likelihood Estimation

- **Maximum likelihood estimation** (MLE) just uses counts/frequencies of seen events in the train data
- Why is this type of estimation called *maximum likelihood*?
- Are there ever any unseen events in language data?

Maximum Likelihood Estimation

- **Maximum likelihood estimation** (MLE) just uses counts/frequencies of seen events in the train data
- Why is this type of estimation called *maximum likelihood*?
- Are there ever any unseen events in language data?
- How could we handle unseen events (not seen before in the training set)?

Rule-based and Statistical NLP

- Rule-based models are a subset of statistics-based models

Rule-based and Statistical NLP

- Rule-based models are a subset of statistics-based models
- Rules have an unsmoothed, uniform distribution – each rule has an equiprobable chance

Rule-based and Statistical NLP

- Rule-based models are a subset of statistics-based models
- Rules have an unsmoothed, uniform distribution – each rule has an equiprobable chance
- Grammatical: $p(x) > 0$, ungrammatical: $p(x) = 0$