Foundations of Natural Language Processing
Lecture 5b
Language Models:
MLE and the Sparse Data Problem

Alex Lascarides

School of informatics

Probabilities of Word Sequences

Last time:

- Probabilities of word sequences useful for ASR, spelling correction, word prediction (texting) . . .

Now: How do we estimate the likelihood of a sequence of n words from corpus data?

But how to estimate these probabilities?

- We want to know the probability of word sequence $\vec{w} = w_1 \ldots w_n$ occurring in English.

- Assume we have some **training data**: large corpus of general English text.

- We can use this data to **estimate** the probability of $\vec{w}$ (even if we never see it in the corpus!)

Probability theory vs estimation

- Probability theory can solve problems like:
  - I have a jar with 6 blue marbles and 4 red ones.
  - If I choose a marble uniformly at random, what's the probability it's red?

Probability theory vs estimation

- Probability theory can solve problems like:
  - I have a jar with 6 blue marbles and 4 red ones.
  - If I choose a marble uniformly at random, what's the probability it's red?

- But often we don't know the true probabilities, only have data:
  - I have a jar of marbles.
  - I repeatedly choose a marble uniformly at random and then replace it before choosing again.
  - In ten draws, I get 6 blue marbles and 4 red ones.
  - On the next draw, what's the probability I get a red marble?

- First three facts are evidence.

- The question requires estimation theory.

Notation

- I will often omit the random variable in writing probabilities, using $P(x)$ to mean $P(X = x)$.

- When the distinction is important, I will use
  - $P(x)$ for *true* probabilities
  - $\hat{P}(x)$ for *estimated* probabilities
  - $P_{\mathrm{E}}(x)$ for estimated probabilities using a particular estimation method E.

- But since we almost always mean estimated probabilities, I may get lazy later and use $P(x)$ for those too.

Example estimation: M&M colors

What is the proportion of each color of M&M?

- In 48 packages, I find[1] 2620 M&Ms, as follows:

Red	Orange	Yellow	Green	Blue	Brown
372	544	369	483	481	371

- How to estimate probability of each color from this data?

[1] Data from: https://joshmadison.com/2007/12/02/mms-color-distribution-analysis/

Relative frequency estimation

- Intuitive way to estimate discrete probabilities:

$$P_{\mathrm{RF}}(x) = \frac{C(x)}{N}$$

where $C(x)$ is the count of x in a large dataset, and $N = \sum_{x'} C(x')$ is the total number of items in the dataset.

Relative frequency estimation

- Intuitive way to estimate discrete probabilities:

$$P_{\mathrm{RF}}(x) = \frac{C(x)}{N}$$

where $C(x)$ is the count of x in a large dataset, and $N = \sum_{x'} C(x')$ is the total number of items in the dataset.

- M&M example: $P_{\mathrm{RF}}(\text{red}) = \frac{372}{2620} = .142$

- This method is also known as **maximum-likelihood estimation** (MLE) for reasons we'll get back to.

MLE for sentences?

Can we use MLE to estimate the probability of $\vec{w}$ as a sentence of English? That is, the prob that some sentence S has words $\vec{w}$?

$$P_{\text{MLE}}(S = \vec{w}) = \frac{C(\vec{w})}{N}$$

where $C(\vec{w})$ is the count of $\vec{w}$ in a large dataset, and
N is the total number of sentences in the dataset.

Sentences that have never occurred

the Archaeopteryx soared jaggedly amidst foliage

vs

jaggedly trees the on flew

- Neither ever occurred in a corpus (until I wrote these slides).
  $\Rightarrow$ $C(\vec{w}) = 0$ in both cases: MLE assigns both zero probability.

- But one is grammatical (and meaningful), the other not.
  $\Rightarrow$ Using MLE on full sentences doesn't work well for language model estimation.

The problem with MLE

- MLE thinks anything that hasn't occurred will never occur (P=0).
- Clearly not true! Such things can have differering, and non-zero, probabilities:
  - My hair turns blue
  - I ski a black run
  - I travel to Finland
- And similarly for word sequences that have never occurred.

Summary

- Maximum Likelihood Estimate (MLE) approach to learning LMs from data.
- Sparse Data Problem: the training corpus can never be truly representative of all English usage!
  Test data may feature word sequences that are absent from training data.

Next Time: We start to deal with the Sparse Data Problem: Assumptions about conditional independence