Topics in Natural Language Processing

Shay Cohen
Institute for Language, Cognition and Computation
University of Edinburgh

Lecture 2
Reminder: the requirements for the class are presentations, brief paper responses and an essay.

- I will suggest papers and topics to cover next weekend
- They will be of different difficulty levels
- Example topics: topic models, language modeling, parsing, semantics, neural networks (your own topic?)
- Choose whatever level of difficulty you feel comfortable with, so that: (a) your presentation is clear; (b) your brief paper response is informative; (c) the essay goes into details about the topic.
Last Class

- What is learning?
- What is a statistical model?
- Basic refresher about probability

\[ \{ p(w|\theta) \mid \theta \in \Theta \} \]

\[ \theta_i > 0 \quad \sum_{i=1}^{A} \theta_i = 1 \quad \Theta \subset \mathbb{R} \]

\[ p(w|\theta) = \theta_i \quad \omega = \{ w_1, \ldots, w_m \} \]
Last class: reminder

Probability distributions, random variables, parametrisation

\[ p(w) \geq 0, \quad \sum_w p(w) = 1, \quad \Omega \ni w \]

\[ p(x=x, y=y) = \sum_w p(w) \]

\[ x(w) = x \]
\[ y(w) = y \]

\[ p(x=x) = \sum_y p(x=x, y=y) \]

\[ p(y=y) = \sum_x p(x=x, y=y) \]

\[ p(x=x | y=y) = \frac{p(x=x, y=y)}{p(y=y)} \]
Today

- What does statistical learning do?
  - Induce a model from data
  - Models tell us how data is generated
  - Learning does the "opposite"

- Two different paradigms to Statistics: frequentist and Bayesian
Approach 1: frequentist Statistics

- We need an objective function $f(\theta, w_1, \ldots, w_n)$

- The higher the value of $f$ is, the better it predicts the training data

\[
D = \{w_1, \ldots, w_n\} \quad \text{data}
\]

\[
D \rightarrow \Theta \quad \text{estimation}
\]

\[
\Theta^* = \underset{\Theta}{\arg\max} f(\Theta, w_1, \ldots, w_n)
\]
Choice of $f$: likelihood

$f(\theta, w_1, \ldots, w_n)$ is a real-valued function

\[
f(\theta, w_1, \ldots, w_n) = \prod_{i=1}^n p(w_i; \theta)
\]

We assume $w_1, \ldots, w_n$ are independent

\[
\Theta^* = \arg\max_{\theta} \prod_{i=1}^n p(w_i; \theta) \leftarrow \text{maximum likelihood}
\]
Log-likelihood

\[ L(w_1, \ldots, w_n | \theta) = \log p(\theta, w_1, \ldots, w_n) = \]
\[ = \log \left( \prod_{i=1}^{n} p(w_i | \theta) \right) = \sum_{i=1}^{n} \log p(w_i | \theta) \]

\[ \theta^* = \arg \max_{\theta} p(w_1, \ldots, w_n, \theta) = \theta = \arg \max_{\theta} L(w_1, \ldots, w_n, \theta) \]

Because \( \log \) is monotone

\[ \arg \max \neq \max \]
Next step

Estimation: maximisation of $L$. The result is the “best” $\theta$ that fits to the data according to the objective function $L$.

$$\arg\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log p(w_i | \theta)$$

average log-likelihood
Pre-historic languages

Imagine a language with two words: “argh” and “blah”
Pre-historic languages

What is $\Omega$?

$$\Omega = \{\text{ygh, blah} \}$$

What is $\Theta$?

$$\Theta = \{ (\theta_a, \theta_c) \mid \Theta \text{ is satisfied} \}$$

Actually, $\Theta = [0,1]$ \hspace{1cm} $\theta \in \Theta \Rightarrow \theta_a = \theta$

$$\theta_c = 1 - \theta$$

What is the training data?

$$w_1, \ldots, w_n \hspace{1cm} w_i \in \{ \text{ygh, blah} \}$$
Pre-historic languages

What is the likelihood objective function?

\[ p(w_i; \theta) = \begin{cases} \theta & w_i = \text{aah} \\ 1-\theta & w_i = \text{blah} \end{cases} \]

\[ \mathbb{I}(w) = \begin{cases} 1 & \text{if } w_i = \text{aah} \\ 0 & \text{if } w_i = \text{blah} \end{cases} \]

\[ p(w_i; \theta) = \theta \mathbb{I}(w_i) (1-\theta) \]

What is the log-likelihood objective?

\[ \log p(w_i; \theta) = \mathbb{I}(w_i) \log \theta + (1-\mathbb{I}(w_i)) \log (1-\theta) \]

\[ L(w_1, ..., w_n; \theta) = \sum_{i=1}^{n} \log p(w_i; \theta) = \]

\[ = \sum_{i=1}^{n} \mathbb{I}(w_i) \log \theta + \sum_{i=1}^{n} (1-\mathbb{I}(w_i)) \log (1-\theta) = \]

\[ = \log \theta \times \left( \sum_{i=1}^{n} \mathbb{I}(w_i) \right) + \log (1-\theta) \times \left( \sum_{i=1}^{n} (1-\mathbb{I}(w_i)) \right) = a \log \theta + b \log (1-\theta) \]
Pre-historic languages

Log-likelihood: \( L(\theta, w_1, \ldots, w_n) = a \log \theta + b \log(1 - \theta) \)

The maximisation problem: \( \theta^* = \arg \max_\theta L(\theta, w_1, \ldots, w_n) \)

How to maximise this?

\[
\frac{\partial L}{\partial \theta} = \frac{a}{\theta} - \frac{b}{1-\theta} = 0 \quad /\theta(1-\theta)
\]

\[
a(1-\theta) - b \theta = 0
\]

\[
a - a \theta - b \theta = a - (a+b) \theta = 0
\]

\[
a = (a+b) \theta
\]

\[
\theta = \frac{a}{a+b} \quad \text{maximum likelihood estimate}
\]

\[
\hat{\sigma} = \frac{a}{n}
\]

\[
1 - \hat{\theta} = \frac{b}{n} = 1 - \frac{a}{n}
\]
Maximisation of log-likelihood

How to maximise the log-likelihood?

we take the derivative of the log-likelihood and set it to 0.
Principle of maximum likelihood estimation

- Objective function: log-likelihood (or likelihood)
- Estimation: maximise the log-likelihood with respect to the set of parameters
A guessing game

I choose a random number between 1 and 20. You need to guess it, and each time you make a guess I tell you whether your guess is higher or lower than my number. What is your strategy to guess the number as quickly as possible?

\[ \log_{20} 5 = \frac{\log 5}{\log 20} \]
A guessing game

I choose a random number between 1 and 20. You need to guess it, and each time you make a guess I tell you whether your guess is higher or lower than my number. What is your strategy to guess the number as quickly as possible?

binary search

I choose a random number \( x \) between 1 and 20 from a distribution \( p(x) \). You know \( p \) and need to guess the number. What is your strategy?
What does log-probability mean?

Let $p$ be a probability distribution over $\Omega$. What is $-\log_2 p(x)$?

$|\text{code}(x)| = -\log_2 p(x)$

$\text{code}(x)$ = sequence of 0's and 1's telling whether we make the choice of "left" or "right" to the averaged mid-point

$E[|\text{code}|] = E[-\log_2 p(x)] = -\sum_x p(x) \log_2 p(x) \triangleq \text{"entropy"}$
Another view of maximum likelihood estimation

What is the “empirical distribution?”

\[ \hat{p}(w) = \frac{\text{count}(w \text{ in data})}{n} \]

Rewriting the objective function \( L(\theta, w_1, \ldots, w_n) \)

\[ L(\theta, w_1, \ldots, w_n) = \frac{1}{n} \sum_{i=1}^{n} \log p(w_i | \theta) = \sum_{w \in \Omega} \hat{p}(w) \log p(w_i | \theta) \]

\[ \theta^* = \arg\min_\theta \sum_{w \in \Omega} \hat{p}(w) \log p(w_i | \theta) \]

\[ = \arg\min_\theta \text{CE}(\hat{p}, p(w_i | \theta)) \]
Cross-entropy

What is the definition of cross-entropy?

\[
CE(p_1, p_2) = - \sum_w p_1(w) \log p_2(w)
\]
Likelihood maximisation

By doing maximum likelihood maximisation we:

- Choose the parameters that make the data most probable,
  
or, from an information-theoretic perspective:

- Choose the parameters that make the encoding of the data most succinct (bit-wise),

  in other words, we

- Minimize the cross-entropy between the empirical distribution and the model we choose.
A bit of history

One of the earliest experiments with statistical analysis of language – measuring entropy of English

2-3 bits are required for English
Approach 2: the Bayesian approach

- History: 1700s. Seminal ideas due to Thomas Bayes and Pierre-Simon Laplace

- A lot has changed since then...
Next class

- The core ideas in Bayesian inference
- Structure in NLP - what type of computational structures are used and how?