Topics in Natural Language Processing

Shay Cohen

Institute for Language, Cognition and Computation

University of Edinburgh

Lecture 2

Administrativia

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

Θ $\quad \{ p(w|\theta) \mid \theta \in \Theta \}$

$\theta_i \geq 0 \quad \sum_{i=1}^{d} \theta_i = 1 \quad \theta \in \mathbb{R}^d$

$p(w_i|\theta) = \theta_i \quad \Omega = \{w_1, \ldots, w_m\}$

Last class: reminder

Probability distributions, random variables, parametrisation

$p(w) \geq 0 \quad \sum_w p(w) = 1 \quad \Omega \ni w$

$X : \Omega \to A$

$Y : \Omega \to B$

$$\boxed{p(X=x, Y=y)} = \sum_{\substack{w \\ X(w)=x \\ Y(w)=y}} p(w)$$

X - sentence

Y - pos tags

$$p(X=x) = \sum_y p(X=x, Y=y)$$

$$p(Y=y) = \sum_x p(X=x, Y=y)$$

$$p(X=x \mid Y=y) = \frac{p(X=x, Y=y)}{p(Y=y)}$$

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

$D = \{w_1, \ldots w_n\}$ data

$D \longmapsto \theta$

estimation

$$\theta^* = \operatorname*{argmax}_{\theta} f(\theta, w_1 .. w_n)$$

Choice of f: likelihood

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

$$f(\theta, w_1 \ldots w_n) = p(w_1 \ldots w_n | \theta) = \prod_{i=1}^{n} p(w_i | \theta)$$

We assume $w_1, \ldots w_n$ are independent

$$\theta^* = \operatorname*{argmax}_{\theta} \prod_{i=1}^{n} p(w_i | \theta) \leftarrow \text{maximising likelihood}$$

Log-likelihood

$$L(w_1 \ldots w_n \mid \theta) = \log f(\theta, w_1 \ldots w_n) =$$

$$= \log\left(\prod_{i=1}^{n} p(w_i \mid \theta)\right) = \sum_{i=1}^{n} \log p(w_i \mid \theta)$$

$$\boxed{\begin{array}{l}\log(a \times b) = \\ = \log(a) + \\ + \log(b)\end{array}}$$

$$\boxed{\begin{array}{c} a \geq b \\ \Updownarrow \\ \log a \geq \log b \end{array}}$$

monotone

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, f(w_1 \ldots w_n, \theta) =$$

$$= \underset{\theta}{\operatorname{argmax}}\, L(w_1 \ldots w_n, \theta)$$

Because log is monotone

$$\operatorname{argmax} \neq \max$$

Next step

Estimation: maximisation of $\cancel{f}\,L$. The result is the “best” θ that fits to the data *according to the objective function* $\cancel{f}\,L$

$$\arg\max_{\theta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log p(w_i \mid \theta)}_{\text{average log-likelihood}}$$

Pre-historic languages

Imagine a language with two words: “argh” and “blah”

Pre-historic languages

What is Ω?

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

What is Θ?

$\theta_a \quad \theta_b \quad \boxed{\theta_a + \theta_b = 1 \qquad \theta_a, \theta_b \geq 0}$ (*)

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

actually,

$\Theta = [0,1] \qquad \theta \in \Theta \Rightarrow \begin{array}{l} \theta_a = \theta \\ \theta_b = 1 - \theta \end{array}$

What is the training data?

$w_1 \dots w_n \qquad w_i \in \{ \text{ugh}, \text{blah} \}$

Pre-historic languages

$a + b = n$

What is the likelihood objective function?

$$p(w_i \mid \theta) = \begin{cases} \theta & w_i = \text{argh} \\ 1-\theta & w_i = \text{blah} \end{cases} \qquad I(w) = \begin{cases} 1 & \text{if } w_i = \text{argh} \\ 0 & \text{if } w_i = \text{blah} \end{cases}$$

$$p(w_i \mid \theta) = \theta^{I(w_i)} (1-\theta)^{1-I(w_i)}$$

What is the log-likelihood objective?

$$\log p(w_i \mid \theta) = I(w_i) \log \theta + (1 - I(w_i)) \log(1-\theta)$$

$$L(w_1 \ldots w_n \mid \theta) = \sum_{i=1}^{n} \log p(w_i \mid \theta) =$$

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

$$= \log \theta \times \Big(\underbrace{\sum_{i=1}^{n} I(w_i)}_{a} \Big) + \log(1-\theta) \times \Big(\underbrace{\sum_{i=1}^{n} (1 - I(w_i))}_{b} \Big) = a \log \theta + b \log(1-\theta)$$

$$\boxed{\log(ab) = \log(a) + \log(b)}$$

$$\boxed{\log(a^b) = b \log(a)}$$

Pre-historic languages

count of argh in D

count of blah in D

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

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

$$\boxed{\begin{aligned} \frac{\partial \log \theta}{\partial \theta} &= \frac{1}{\theta} \\ \frac{\partial \log(1-\theta)}{\partial \theta} &= -\frac{1}{1-\theta} \end{aligned}}$$

How to maximise this?

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

$$a(1-\theta) - b\theta = 0$$

$$a - a\theta - b\theta = a - (a+b)\theta = 0$$

$$a = (a+b)\theta$$

$$\theta^* = \frac{a}{\underbrace{a+b}_{n}}$$

maximum likelihood estimate

$$\theta^* = \frac{a}{n} \qquad 1-\theta^* = \frac{b}{n} = 1 - \frac{a}{n}$$

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?

$\lceil \log_2 20 \rceil = 5$

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 Ω. What is $-\log_2 p(x)$?

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

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

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

$$= -\sum_x p(x) \log_2 p(x)$$ ← "entropy"

Another view of maximum likelihood estimation

What is the “empirical distribution?”

$$\tilde{p}(w) = \frac{\text{count}(w \text{ in data})}{n}$$

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

$$L(\theta, w_1 \ldots w_n) = \frac{1}{n} \sum_{i=1}^{n} \log p(w_i|\theta) =$$

$$= \sum_{w \in \Omega} \tilde{p}(w) \log p(w_i|\theta)$$

$$\theta^* \quad \underset{\theta}{\operatorname{argmin}} - \sum_{w \in \Omega} \tilde{p}(w) \log p(w_i|\theta)$$

$$= \underset{\theta}{\operatorname{argmin}} \; CE(\tilde{p}, p(w|\theta))$$

Cross-entropy

What is the definition of cross-entropy?

$$CE(p_1, p_2) = -\sum_w p_1(w) \underline{\log p_2(w)}$$

Likelihood maximisation

By doing maximum likelihood maximisation we:

- Choose the parameters that make the data most probable,

  or, from an information-theoretic perspective:

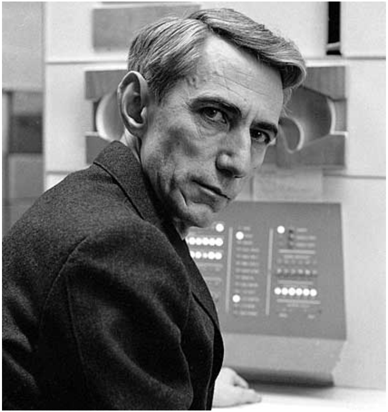
- Choose the parameters that make the encoding of the data most succinct (bit-wise),

  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?