Topics in Natural Language Processing

Shay Cohen

Institute for Language, Cognition and Computation

University of Edinburgh

Lecture 3

Last class

Maximum likelihood estimation:

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

$$\theta^*_{\text{MLE}} = \arg\max_{\theta} L(\theta, w_1, \cdots, w_n)$$

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. Number of steps: $\log_2 n = -\log_2 \frac{1}{n}$.

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. Number of steps: $\log_2 n = -\log_2 \frac{1}{n}$.

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 $-\log_2 p(x)$?

Number of bits it takes to encode an optimal code for Ω when the true distribution is $p(x)$

Entropy:

$$H(p) = -\sum_x p(x)\log_2 p(x) = \mathbb{E}_p[|\mathrm{code}(x)|]$$

The code is a bit-by-bit description of whether we take the decision “lower” or “higher” in the game

Another view of maximum likelihood estimation

What is the “empirical distribution?”
$\tilde{p}(w)$ be a probability distribution over the domain of datapoints such that $\tilde{p}(w)$ is the fraction of the n datapoints such that they are identical to w.

$$\tilde{p}(w) = \frac{\text{count}(w; w^{(1)}, \ldots, w^{(n)})}{n}$$

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

$$\begin{aligned} L(\theta, w_1, \ldots, w_n) &= \frac{1}{n} \sum_{i=1}^{n} \log p(w_i \mid \theta) \\ &= \sum_{w \in \Omega} \tilde{p}(w) \log p(w \mid \theta) \end{aligned}$$

This is the cross entropy between $\tilde{p}$ and p

Cross-entropy

What is the definition of cross-entropy?

$$\mathrm{CE}(p, q) = -\sum_x p(x) \log q(x) = \mathbb{E}_p[-\log q(x)]$$

Cross-entropy

What is the definition of cross-entropy?

$$\mathrm{CE}(p, q) = -\sum_{x} p(x) \log q(x) = \mathbb{E}_p[-\log q(x)]$$

- Cross entropy is *not symmetric*, as such it is not “distance”, but it does tell whether p and q are close to each other

- For any given p, it is minimized when $q = p$

- It tells the expected number of bits we would use if we “encode” using q when p is the true distribution

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.

Types of Models

It is often the case that we discuss a model $p(x \mid \theta)$

Really, in NLP, you are interested in predicting some $y(x)$

Therefore, you need $p(x, y \mid \theta)$. Estimation is the same when both x and y are in the dataset. Later we will learn about incomplete data

In some cases you model also $p(y \mid x, \theta)$ (e.g. neural networks, log-linear models).

This gives the generative vs. discriminative model distinction

Types of Objectives

We showed an example of deriving the log-likelihood solution for a simple model

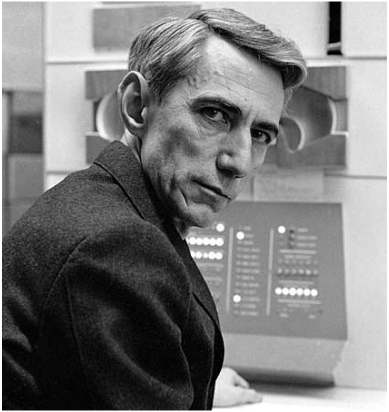
One can have more complex objective functions, and the principle would be the same

You just might not have a closed-form solution (e.g. with deep learning, log-linear models, etc.)

You need to apply an *optimisation* algorithm – more on that later

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...

Use of Bayesian Learning in NLP

Very often used in the context of *unsupervised* learning. Why?

- Hard to beat discriminative methods in the supervised case (log-linear models, deep learning, etc.)
- Flexible framework for latent variables
- Priors play much more important role in the unsupervised setting

See more discussion in the reading material

Bayes' rule

What is Bayes' rule?
If you have two random variables X and Y then

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

(show that using the chain rule)

Reminder: What does Statistics do? Invert the relationship between model and data.

Bayes' rule does the same with random variables.

Bayes' rule

What is Bayes' rule?
If you have two random variables X and Y then

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

(show that using the chain rule)

Reminder: What does Statistics do? Invert the relationship between model and data.

Bayes' rule does the same with random variables.

What if our model parameters were one random variable and our data were another random variable?
Define our "X" to be θ
Define our "Y" to be the data

Prior beliefs about models

We have a parameter space Θ and prior beliefs $p(\theta)$.

Our θ is now a random variable.

From the chain rule: $p(w, \theta) = p(\theta)p(w|\theta)$

$$p(\theta \mid w) = \frac{p(w \mid \theta)p(\theta)}{p(w)}$$

- $p(w \mid \theta)$ - the likelihood
- $p(\theta)$ - the prior
- $p(w)$ - the evidence

Posterior inference

$$p(\theta \mid w) = \frac{p(w \mid \theta)p(\theta)}{p(w)}$$

basic posterior inference

$p(w) =$

$$= \int_\theta p(w \mid \theta)p(\theta)d\theta$$

because

$$\int_\theta p(\theta \mid w)d\theta = 1$$

and therefore

$$\int_\theta p(w \mid \theta)p(\theta)/p(w)d\theta = 1$$

and therefore $p(w) = \int_\theta p(w \mid \theta)p(\theta)d\theta$

Priors

Our prior beliefs are considered in inference. There is no “correct” prior.

Is that a good or bad thing?

Priors

Our prior beliefs are considered in inference. There is no “correct” prior.

Is that a good or bad thing?

- Frequentists: probability is the frequency of an event
- Bayesians: probability denotes the state of our knowledge about an event
  - Subjectivists: probability is a personal belief
  - Objectivists: minimise human’s influence on decision making
- In practice: NLP use of Bayesian theory is largely driven by computation

Back to pre-historic languages

Language with two words: “argh” and “blah”

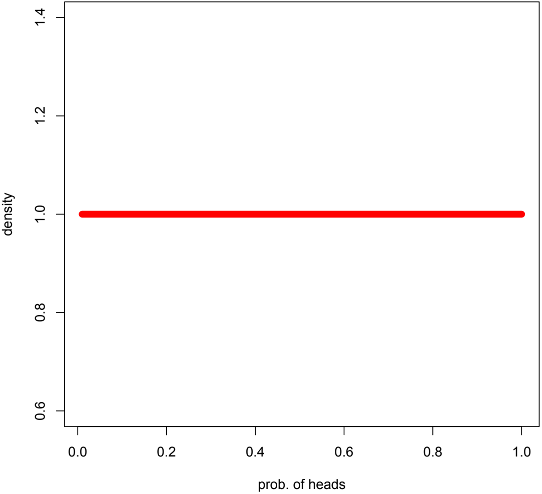
Our Ω is $\{\text{argh}, \text{blah}\}$.

Our Θ is $[0, 1]$.

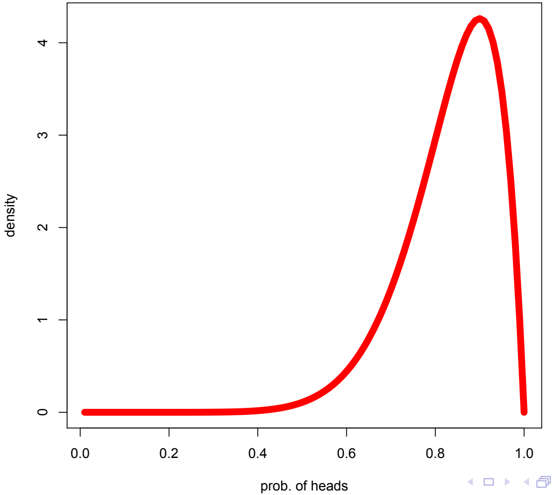
Define $I(w) = 1$ if $w = \text{argh}$ and 0 if $w = \text{blah}$.

Then, $p(w|\theta) = \theta^{I(w)}(1-\theta)^{(1-I(w))}$.

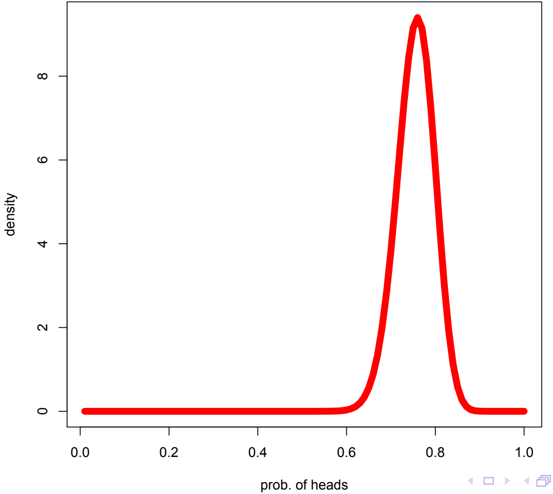
Uniform prior, 0.7 prob. for argh



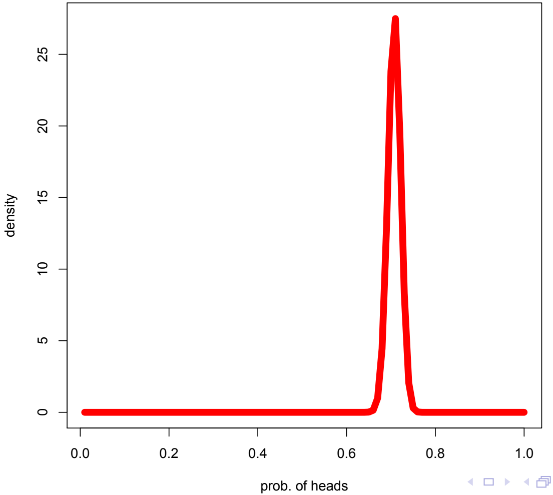
Posterior with 10 datapoints, truth is 0.7 prob. for argh



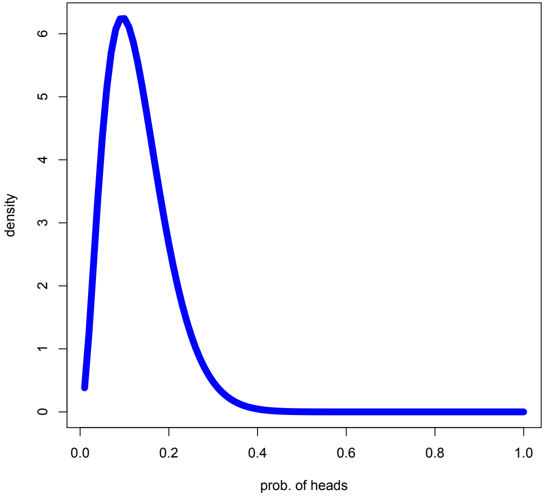
Posterior with 100 datapoints, truth is 0.7 prob. for argh



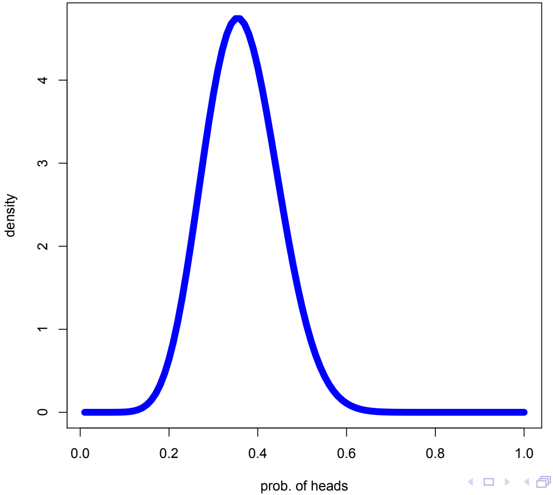
Posterior with 1000 datapoints, truth is 0.7 prob. for argh



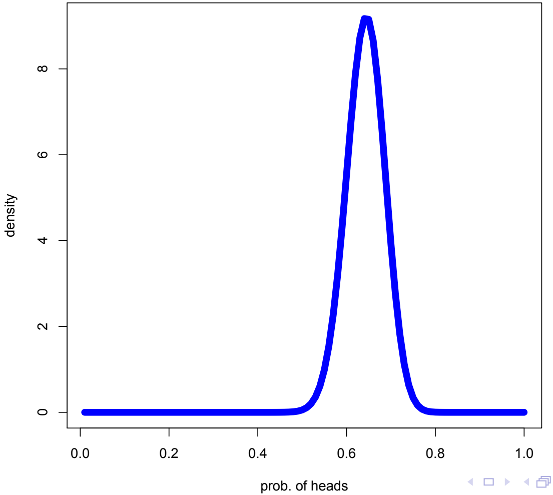
Non-uniform prior, truth is 0.7 prob. for argh



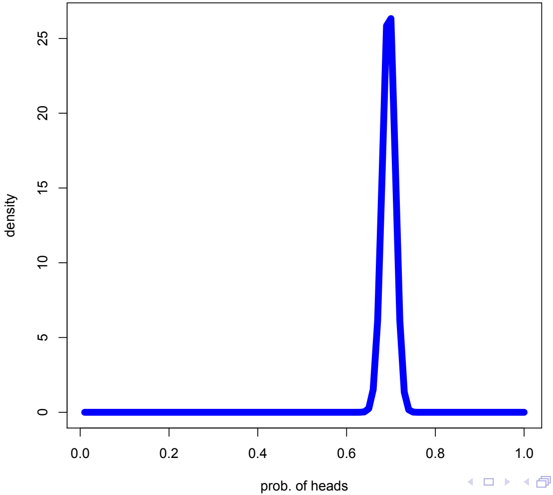
Posterior with 10 datapoints, truth is 0.7 prob. for argh



Posterior with 100 datapoints, truth is 0.7 prob. for argh



Posterior with 1000 datapoints, truth is 0.7 prob. for argh



Priors for binary outcomes

$p(\theta) \propto \theta^{\alpha}(1-\theta)^{\beta}$ $\quad$ $p(w|\theta) = \theta^{I(w)}(1-\theta)^{(1-I(w))}$

What is the posterior?
$D = \{w_1, \ldots, w_n\}$

$$p(\theta \mid w_1, \ldots, w_n) = \frac{p(w_1, \ldots, w_n \mid \theta)p(\theta)}{p(w_1, \ldots, w_n)}$$

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

$$\propto \prod_{i=1}^{n} \theta^{I(w_i)}(1-\theta)^{1-I(w_i)} \times \theta^{\alpha}(1-\theta)^{\beta}$$

$p(\theta)$ is also called the Beta distribution $\mathrm{Beta}(\alpha, \beta)$

$$= \frac{\theta^{\sum I(w_i)} \times (1-\theta)^{n-\sum I(w_i)}}{p(w_1,\ldots,w_n)} \times \theta^{\alpha}(1-\theta)^{\beta}$$

$$= \frac{\theta^{\alpha+\sum I(w_i)} \times (1-\theta)^{\beta+n-\sum I(w_i)}}{p(w_1,\ldots,w_n)}$$

Note that $p(w_1,\ldots,w_n)$ does not depend on θ, so we get that the "posterior" is:

$$\mathrm{Beta}(\sum_{i=1}^{n} I(w_i) + \alpha, n - \sum_{i=1}^{n} I(w_i) + \beta)$$

Maximum a posteriori estimate (MAP)

“Bayesian estimation”: find θ^* that maximises the posterior:

$$\theta^*_{\text{MAP}} = \arg\max_\theta p(\theta \mid w_1, \ldots, w_n)$$

$$= \arg\max_\theta \theta^{a+\alpha}(1-\theta)^{b+\beta}$$

Same process as with MLE (derivative and equating to 0):

$$\theta^*_{\text{MAP}} = \frac{a+\alpha}{a+b+\alpha+\beta} = \frac{a+\alpha}{n+\alpha+\beta}$$

where $a = \sum_i I(w_i)$ and $b = n - \sum_i I(w_i)$

MAP and posteriors

In general,

- Priors are especially important when the amount of data is small
- As there is more data, the prior becomes less influential on the posterior
- Under some mild conditions, the posterior is a distribution concentrated around the MLE

Conjugacy of prior and likelihood

$p(\theta) \propto \theta^{\alpha}(1-\theta)^{\beta}$ $\quad$ $p(w|\theta) = \theta^{I(w)}(1-\theta)^{(1-I(w))}$

Prior is “hyperparametrised”. What is the posterior?

$$p(\theta \mid w) \propto \theta^{a+\alpha}(1-\theta)^{b+\beta}$$

$$p(\theta) \sim \mathrm{Beta}(\alpha, \beta) \Rightarrow p(\theta \mid w_1, \ldots, w_n) \sim \mathrm{Beta}(a+\alpha, b+\beta)$$

Conjugacy of prior to a likelihood

We are given a prior family and a likelihood family.

Each probability distribution in the likelihood family is parametrised by a parameter in the prior family

We say that they are conjugate to each other if the posterior of a likelihood for a prior in the family stays a member of the prior family

“It all stays in the family...”

Conjugacy – always useful?

Trivial non-useful example of conjugacy

Prior family is:

$$P = \{\textit{All distributions over } \Theta\}$$

Posterior is always in P. As such, P is conjugate to any likelihood. Useful?

Conjugacy – always useful?

Another trivial non-useful example of conjugacy

Prior family is:

$$P = \{p(\theta) \mid p(\theta) \textit{ places all mass on a single point } \theta_0\}$$

Posterior is always in P. As such, P is conjugate to any likelihood. Useful?

Conjugacy: summary

Conjugacy is useful when:

- The prior is not too poor
- It is easy to calculate the posterior hyperparameters

In many cases, conjugate priors lead to treating the hyperparameters from the prior as “pseudo-observations”

(Such is the case with additive smoothing)

Minimum Description Length and MAP

What is $-\log_2 p(\theta|w_1,\ldots,w_n)$?
# of bits that is required to encode θ when $w_1,\ldots,w_n$ is known.

Minimum Description Length and MAP

What is $-\log_2 p(\theta|w_1, \ldots, w_n)$?
# of bits that is required to encode θ when $w_1, \ldots, w_n$ is known.
What is $-\log_2 p(\theta)$?
# bits that is required to encode θ apriori

Minimum Description Length and MAP

What is $-\log_2 p(\theta|w_1,\ldots,w_n)$?
# of bits that is required to encode θ when $w_1,\ldots,w_n$ is known.
What is $-\log_2 p(\theta)$?
# bits that is required to encode θ apriori
What is $-\log_2 p(w_1,\ldots,w_n|\theta)$?
# bits that is required to encode the data if we think θ is “correct”

Minimum Description Length and MAP

What is $-\log_2 p(\theta|w_1,\ldots,w_n)$?
# of bits that is required to encode θ when $w_1,\ldots,w_n$ is known.
What is $-\log_2 p(\theta)$?
# bits that is required to encode θ apriori
What is $-\log_2 p(w_1,\ldots,w_n|\theta)$?
# bits that is required to encode the data if we think θ is "correct"
MAP: $\theta^* = \arg\min_\theta -\log_2 p(\theta) - \log_2 p(w_1,\ldots,w_n|\theta)$

Encoding θ^* requires separately:

- Encoding the hypothesis according to the prior
- Encoding the data according to the hypothesis

That's the "minimum description length" criterion

Learning from Incomplete Data

Just as a side note:

- Bayesian analysis in NLP is especially uesful for learning from incomplete data
- It presents a flexible framework for introducting latent variables into a model
- Posterior inference becomes more complex

Summary

Bayesian analysis:

- Only uses Bayes' rule to do inference
- Posterior is a *distribution* over parameters
- Can summarise the posterior, e.g. MAP, to get a point estimate
- Need to be careful about choice of prior
- Especially important with small amounts of data
- MAP has a connection to minimum description length (MDL)

Discussion

Bayesian analysis or frequentist analysis?