Topics in Natural Language Processing

Shay Cohen

Institute for Language, Cognition and Computation

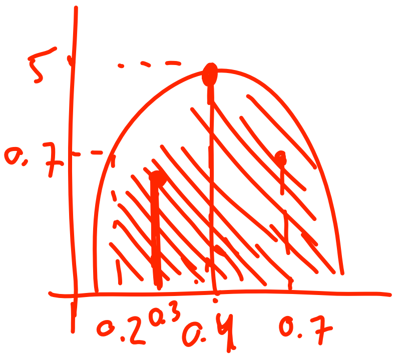
University of Edinburgh

Lecture 4

Administrativia

- You should have been added to the Piazza forum
- Today is the last day to send in the papers for presentations. I will send an email over the weekend with assignments.

$p(\theta)$ - density like a prob. dist. over a continuous space, $[0,1]$



we want from a density

$$p(\theta) \geq 0$$

$$\int_\theta p(\theta)\, d\theta = 1$$

if this is

$p(\theta \mid w_1 \ldots w_n)$

we just choose

$$\boxed{\theta^* = 0.4}$$

Last class

Bayesian inference: $\underset{\text{prior}}{p(\theta)}$ $\qquad p(w_1 \ldots w_n \mid \theta) =$

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

$$p(\theta \mid w_1 \ldots w_n) \underset{\text{Bays}}{=} \boxed{\frac{p(\theta)\, p(w_1 \ldots w_n \mid \theta)}{p(w_1 \ldots w_n)}}$$

$$\Rightarrow \quad p(w_1 \ldots w_n) = \int_\theta \underbrace{p(w_1 \ldots w_n \mid \theta)\, p(\theta)}\, d\theta$$

MAP - maximum a posteriori

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

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

$$\underline{p(\theta) \propto \theta^{\alpha}(1-\theta)^{\beta}} \leftarrow \text{Beta} \qquad \underline{p(w|\theta) = \theta^{I(w)}(1-\theta)^{(1-I(w))}}$$

Prior is "hyperparametrised". What is the posterior?

$$p(\theta | w_1, \ldots, w_n) \propto \theta^{\alpha + a}(1-\theta)^{\beta + b}$$

$$\underbrace{p(\theta | w_1, \ldots, w_n)}_{\text{posterior}} = \boxed{\frac{\theta^{\alpha+a}(1-\theta)^{\beta+b}}{\int_{\theta'} \theta'^{\alpha+a}(1-\theta')^{\beta+b}\, d\theta'}} \leftarrow \text{Beta}$$

"new α" $\quad \alpha + a$

"new β" $\quad \beta + b$

Definition of Conjugacy

$[0,1]$ (α, β)

Let P be a set of priors hyperparametised by a set $\mathcal{A}$, for a parameter space Θ. Therefore, each $p \in P$ is a probability distribution $p(\theta \mid \alpha)$. Let M be a model over Ω such that each $p \in M$ is a probability distribution $p(w \mid \theta)$. We say, P **is conjugate to** M, if for any choice of $\alpha \in \mathcal{A}$ and data $w_1, \ldots, w_n$ it holds that $p(\theta \mid w_1, \ldots, w_n, \alpha) \in P$.

Definition of Conjugacy

Let P be a set of priors hyperparametised by a set $\mathcal{A}$, for a parameter space Θ. Therefore, each $p \in P$ is a probability distribution $p(\theta \mid \alpha)$. Let M be a model over Ω such that each $p \in M$ is a probability distribution $p(w \mid \theta)$. We say, P **is conjugate to** M, if for any choice of $\alpha \in \mathcal{A}$ and data $w_1, \ldots, w_n$ it holds that $p(\theta \mid w_1, \ldots, w_n, \alpha) \in P$.

Previous example (argh-blah example):

$$M = \left\{ p(w \mid \theta) \mid \theta \in [0,1] \right\}, \quad p(w \mid \theta) = \begin{cases} \theta & \text{argh} \\ 1-\theta & \text{blah} \end{cases}$$

$$P = \left\{ p(\theta) = \theta^{\alpha}(1-\theta)^{\beta} \mid \alpha \geq 0, \beta \geq 0 \right\}$$

Posterior new hyperparameters:

$$\underline{\theta^{\underbrace{\alpha+a}_{\alpha'}}(1-\theta)^{\underbrace{\beta+b}_{\beta'}}}$$

Conjugacy – always useful?

Trivial non-useful example of conjugacy

$P = \{\text{set of all distributions over } \Theta\}$

$M = \text{some model}$

Conjugacy – always useful?

Another trivial non-useful example of conjugacy

choose some θ $\quad \theta = 0.7$

$$P = \left\{ p(\theta) \text{ s.t. } p(\theta) = \begin{cases} 1 & \text{if } \theta = 0.7 \\ 0 & \text{o/w} \end{cases} \right\}$$

$$p(\theta | w) = \frac{p(\theta)\, p(w|\theta)}{p(w)} = \begin{cases} 1 & \theta = 0.7 \\ 0 & \text{o/w} \end{cases}$$

Conjugacy: summary

Conjugacy is useful when:

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

(α, β)

Minimum Description Length and MAP

What is $-\log_2 p(\theta|w_1,\ldots,w_n)$?

Minimum Description Length and MAP

What is $-\log_2 p(\theta|w_1,\ldots,w_n)$?

What is $-\log_2 p(\theta)$?

Minimum Description Length and MAP

What is $-\log_2 p(\theta|w_1, \ldots, w_n)$? # bits that are required using a "good" code to encode θ given $w_1 \ldots w_n$

What is $-\log_2 p(\theta)$? # bits that are required using a good code to encode θ a priori

What is $-\log_2 p(w_1, \ldots, w_n|\theta)$? # bits that are required to encode the data if we think θ generated it

$$\theta^* = \arg\max_\theta p(\theta | w_1 \ldots w_n)$$

Minimum Description Length and MAP

What is $-\log_2 p(\theta|w_1, \ldots, w_n)$?

What is $-\log_2 p(\theta)$?

What is $-\log_2 p(w_1, \ldots, w_n|\theta)$?

MAP: $\theta^* = \arg\max_\theta \log_2 p(\theta) + \log_2 p(w_1, \ldots, w_n|\theta)$

$$\theta^* = \arg\max_\theta p(\theta|w_1 \ldots w_n) =$$
$$= \arg\max_\theta \frac{p(\theta)\, p(w_1 \ldots w_n|\theta)}{p(w_1 \ldots w_n)}$$
$$= \arg\max_\theta \log p(\theta) + \lg p(w_1 \ldots w_n|\theta)$$
$$= \arg\min_\theta -\log p(\theta) - \lg p(w_1 \ldots w_n|\theta)$$

Encoding θ^* requires separately:

- Encoding the hypothesis according to the prior

- Encoding the data according to the hypothesis

That's "minimum description length" criterion

$$\theta^* = \arg\max_\theta \underbrace{\log p(\theta)} + \sum \log p(w_i \mid \theta)$$

$$\theta_{MLE} = \arg\max_\theta \sum_i \log p(w_i \mid \theta)$$

Regularization

$$\theta_{MLE} = \arg\max_\theta \sum_i \log p(w_i \mid \theta) - \lambda \sum_i \theta_i^2$$

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)