Bayesian Analysis for Natural Language Processing
Lecture 3

Shay Cohen

February 18, 2013

Administrativia

- There is now a schedule for paper discussions until April 1st
- If you are in there, you should have received an email from me

Refresher: two weeks ago (conjugate priors)

"Conjugacy" here describes a relationship between the likelihood and a **family** of prior distributions ($\mathcal{P}$): the posterior belongs to $\mathcal{P}$

This definition by itself is not sufficient

- Is a singleton set that consists of delta-function (all mass is concentrated on a single parameter) a conjugate prior?
- Is the family of all possible priors a conjugate prior?

Want other properties: richness, interpretability, computationally efficient

Refresher cont'd: conjugate priors

- The most common example of conjugacy in NLP: Dirichlet-multinomial conjugacy
- Dirichlet can be interpreted in various ways, e.g. pseudo-counts
- Small values of hyperparameters lead to sparsity

Today

- Rachel is going to discuss chapter 3 (Bayesian estimation)
- Armineh is going to discuss a paper comparing Bayesian estimation procedures

Feel free to participate in the discussion and ask questions

Being fully Bayesian

Question:

> *On pages 27 and 28, Shay explains the fully Bayesian approach is often not necessary for the purposes of NLP and that point estimation is sufficent. I'm still not sure why being fully Bayesian is less appropriate in an NLP setting. Perhaps from the standpoint of applications, computational speed is always preferable. At the same time, it seems to me that a lot of development in the field would rest on the results from algorithms that try to more fully express a probabilistic process through understanding the parameters in the model. Please clarify this methodological approach.*

L2 regularization

Question:

With the L2 regularization, the MAP estimates behave like penalized maximum likelihood estimates. But how much does the prior variance (sigma) affect this objective function?

MCMC

Question:

In general I am not familiar with the unsupervised learning setting(section 3.2.1.3), especially in regard to latent variables. Why is estimation in this situation intractable? Additionally I don't understand how methods like MCMC are able to infer a distribution that we don't know. In the latent variable case, is this just finding some general distribution for dividing data into n latent states?

Laplace Approximation

Question:

> *when using the Laplace approximation, it is better to change the parametrization of the prior so that theta(sub i) is defined on a real line." Would you please try to explain both mechanically and intuitively why this is the case as I am not comfortable with Laplace approximations.*