Bayesian Analysis for Natural Language Processing
Lecture 2

Shay Cohen

February 4, 2013

Administrativia

- The class has a mailing list: coms-e6998-11@cs.columbia.edu
- Need two volunteers for leading a discussion on 2/18 (each leading a discussion on one paper)
- **Next week:** guest lecture by Bob Carpenter. I will send to the mailing list two papers he will cover.

Refresher: last week

- Bayesian statistics manages uncertainty using distributions over the parameters
- Start with a prior ("half-baked idea") and update it to the posterior
- The two components of the model: prior and likelihood
- Inference: uses Bayes' theorem

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)},$$

where

$$p(x) = \int_{\theta} p(\theta)p(x|\theta)d\theta.$$

Today's class

- We will focus on **priors** in NLP and in general
- If time permits: Bayesian decision theory
- If time permits: MAP inference

All your questions and comments were great. I picked the ones which seemed most relevant to everybody.

Do Bayesian models overfit?

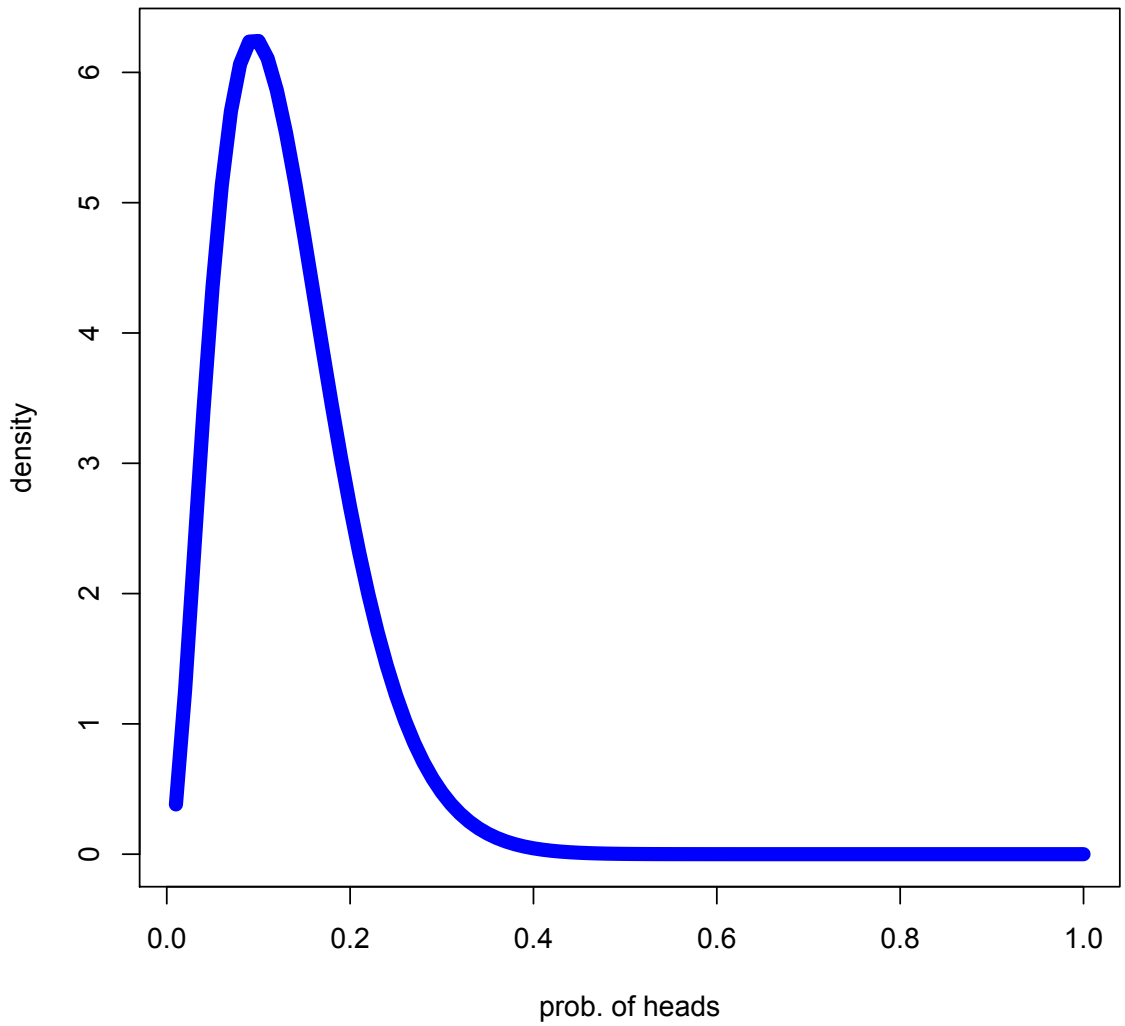
Question:

> *There is a lot of talk about the power that Bayesian statistics can bring to NLP because of the additional degrees of freedom added by picking the prior (and potentially also introducing latent variable and hyperparameters). However, one concern that is apparent to me is that with this freedom there is the danger for the model to "overfit" training data and perform poorly on test data. Of course this is a theoretical concern, and if Bayesian methods perform well on test sets in practice this is not a real issue but I think it is at least a theorical issue with these methods.*

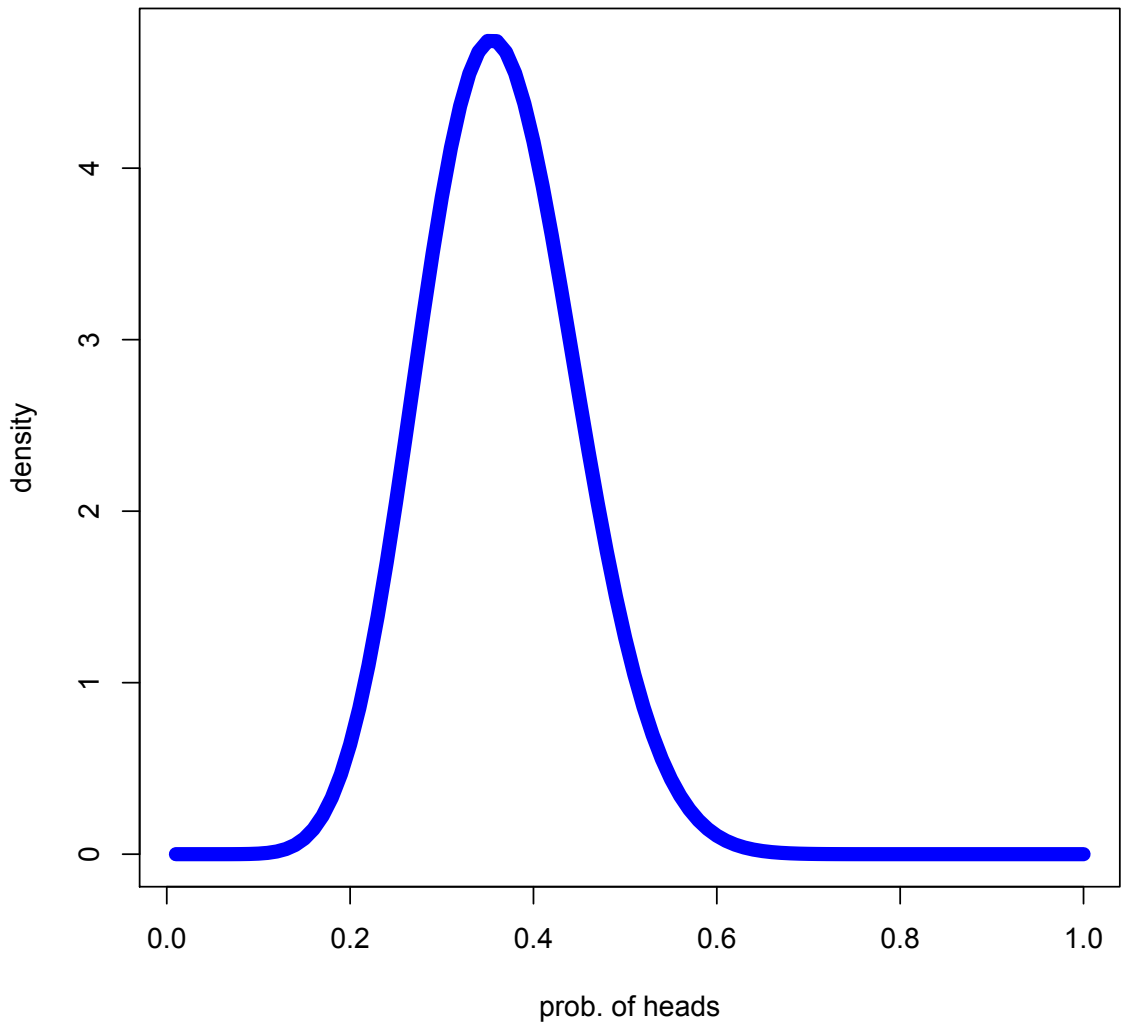
Do Bayesian models overfit?

- Bayesian models usually work in an unsupervised (transductive or non-transductive) setting
- MAP inference with certain priors can actually *alleviate* overfitting – for example, L_2 regularization corresponds to MAP inference with Gaussians

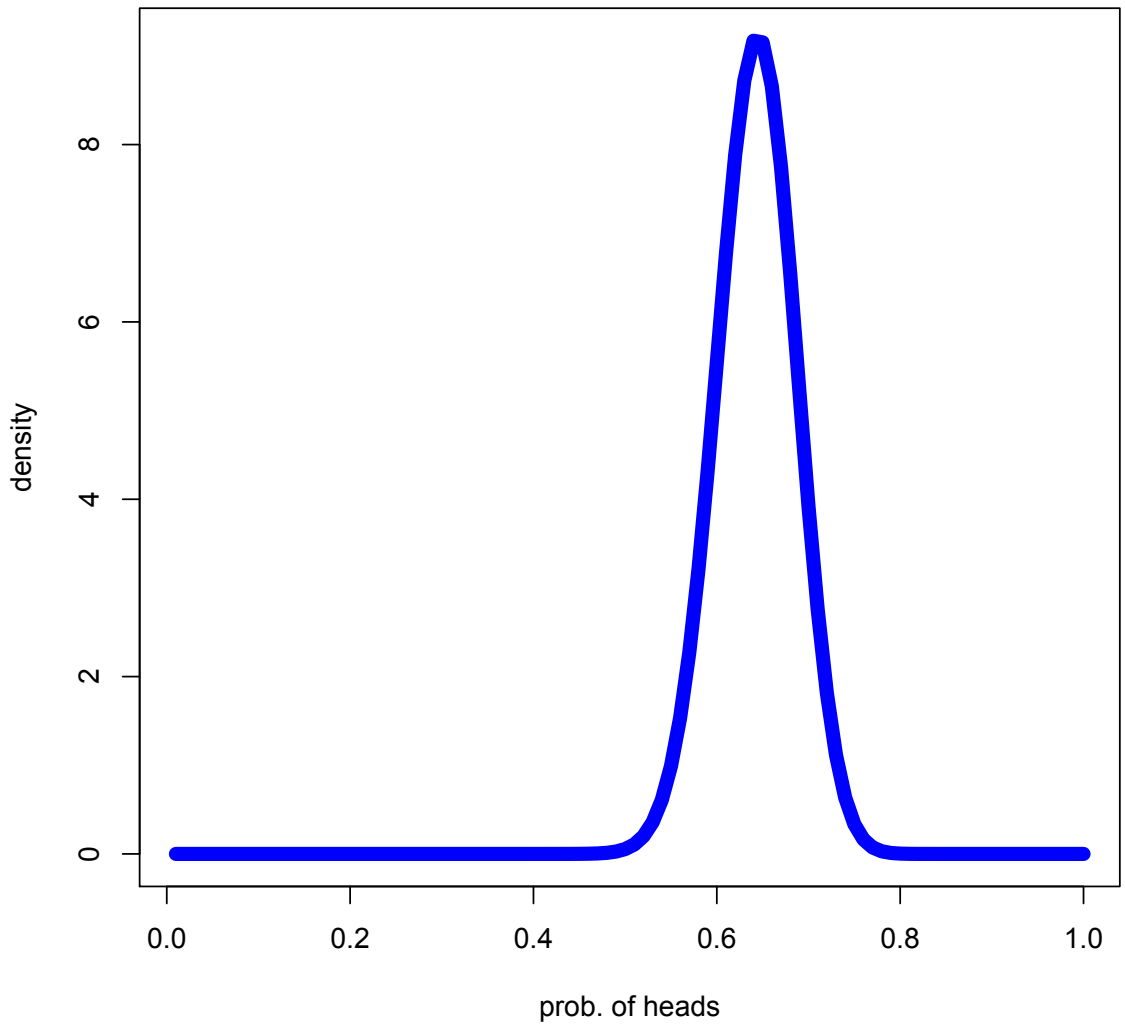
Non-uniform prior, coin with 0.7 prob. for heads



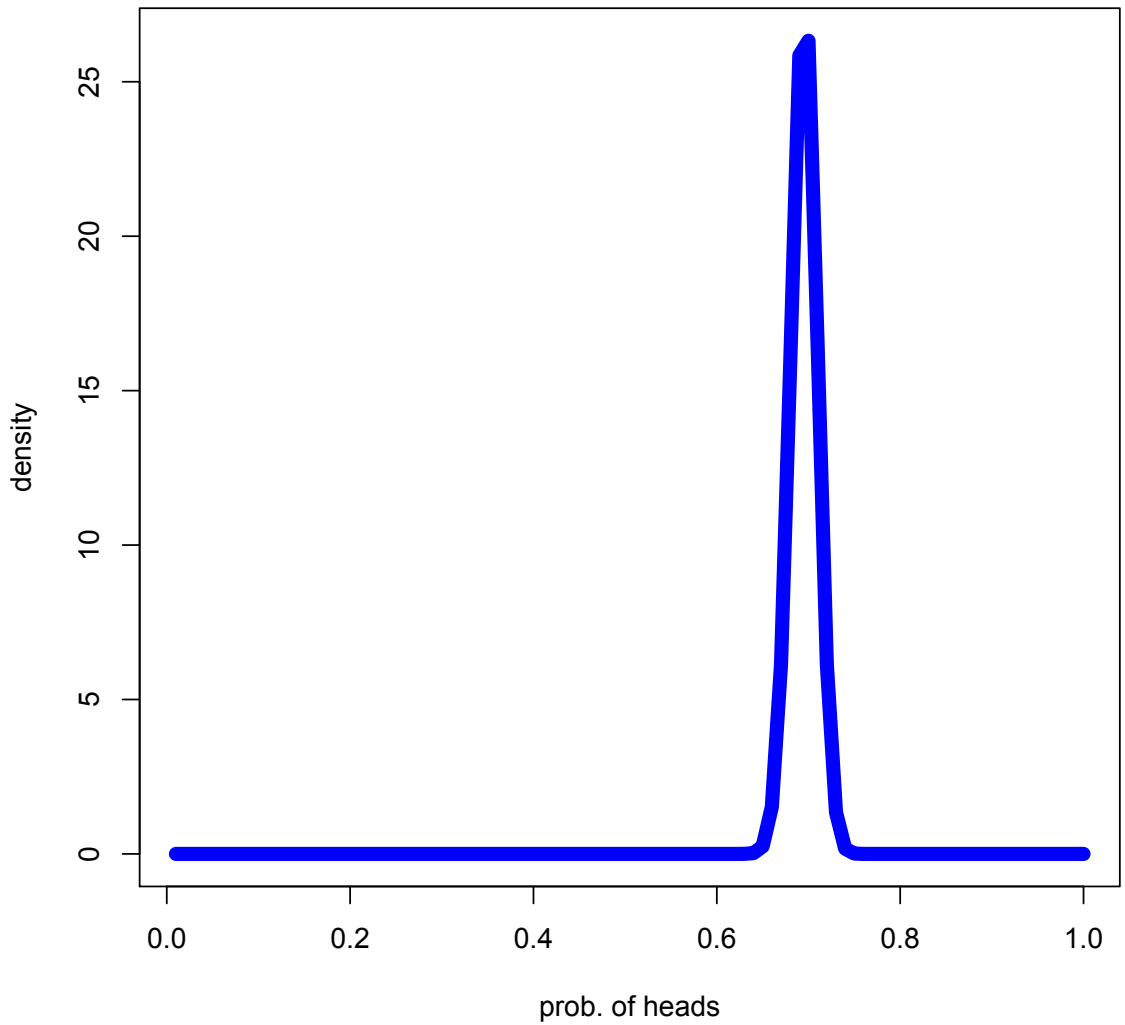
Posterior after 10 tosses, coin with 0.7 prob. for heads



Posterior after 100 tosses, coin with 0.7 prob. for heads



Posterior after 1000 tosses, coin with 0.7 prob. for heads



Effect of priors

Question:

> *In practice is there a trend that bayesian methods probably achieve their best performance with the lesser number of examples than the frequentist methods take to achieve their best performance? The motivation for this question is if the bayesian methods are given good priors to start with I get the feel that the parameter search space would be lesser compared to the frequentist methods and hence they require fewer examples to learn a good model.*

Effect of priors

- Priors have great effect when the sample size is not too large
- In many cases, the posterior converges to a distribution with mass concentrated around the maximum likelihood estimate
- In NLP, we usually don't have enough samples, so the prior can have large effect

Conjugate priors

What are conjugate priors?

Dirichlet prior

Question:

> *I was wondering if there was any work done of the theoretic justification of conjugate priors specifically the Dirichlet priors. That is, is there any theoretical work indicating that Dirichlet is a good prior for any reasons other than computation efficiency?*

Dirichlet prior

- Main reason for choosing Dirichlet is computational convenience
- But... Dirichlet priors can encourage sparsity
- They are also interpretable as adding extra counts to empirical counts – a technique known as “additive smoothing”

Conjugacy to multinomial

Question:

Is the Dirichlet distribution the only conjugate prior with respect to the multinomial distribution? In general, can a distribution have multiple conjugate prior associated with it?

Conjugacy to multinomial

- Is a family of priors with a single delta function on some parameter conjugate?
- Is a family of all possible priors conjugate?
- Is a mixture of conjugate priors also conjugate?

Prior selection

Question:

How to choose priors? Mostly, tractability is the key for choosing priors. I am not sure advantages and disadvantages of each prior from the aspect of NLP.

Prior selection

- Do we have reasons to believe one parameter is more likely than the other? If not, use non-informative prior
- Can we interpret the prior?
- Are we looking to model some aspect of the parameters? Sparsity? Correlation?
- Is there an interpretation to the prior that is useful for our data?
- Empirically: Trial-and-error
- Sometimes selection is done automatically

Prior design

Question:

> *can one manually design a prior to suit his needs? Or is it the case that one must always choose from the existing distributions over the real? I suppose unlike discrete distributions (where you can just specify the probability of each item), priors are hard to hand-design if possible at all*

Prior design

- Prior elicitation - not done much in NLP, if at all
- Empirical Bayesian setting - “learning” the hyperparameters

Dirichlet vs. Logistic Normal

Question:

> *You mentioned that the logistic normal distribution allows modeling for dependence as opposed to the Dirichlet distribution. I didn't quite get how so.*

When Bayesian Statistics works well?

Question:

It would be great if we could see an example comparing Bayesian and MLE (Frequentist) estimators. Specifically, one where (empirically) the bayesian estimator is better than the MLE, and vice versa. Also, are there any examples where Bayesian statistics leads to very bad errors? (I'm guessing that the best way to proceed is - analyze the data, pick a very good prior, and then adopt the model. But from the plots in class it seems like with enough data you wash out the prior - this always happens? How important is the prior then?)

When Bayesian Statistics works well?

- If most signal is in the likelihood, Bayesians and frequentists get similar results
- This happens when:
  - Lots of data is available
  - Complete data is available (again, in large amounts)
  - Prior is non-informative
- Bayesian Statistics enables “diversity” of the parameters, mananging uncertainty using distributions

Hyperparameters

Question:

What are hyperparameters?

Hyperparameters

- Priors usually come from a parametric family
- Priors define a distribution over parameters, and the parameters controlling the distribution are hyperparameters
- Empirical Bayes vs. hierarchical Bayesian models

Do we need anything but conjugate priors?

Question:

Aren't we always selecting distributions to model our data because they are elegant/convenient? I have difficulty understanding why we can assume a variable is distributed Gaussian, for example, and we should not assume that its mean is also distributed Gaussian? Isn't that the whole assumption of a generative model? If the mean is not distributed Gaussian, we could always put in an uninformative prior. Could you provide an NLP based example in which a conjugate prior is not an appropriate means of expressing the distribution of a variable in question?

Do we need anything but conjugate priors?

- Balance: efficiency and accuracy of model
- Dirichlet is sometimes a weak choice of prior