Bayesian Statistics

Priors and Model Selection

Valentin Zambelli

Seminar Data Modeling Humboldt University Berlin, WS 16/17

Table of contents

1. Motivation
2. Choosing a good prior
3. Mixtures of conjugate priors
4. Analyzing the posterior
5. Selecting a model
6. Conclusion

Motivation

You have a coin that is assumed to be fair ($\Theta = 0.5$). You flip it 10 times and get 7 heads.

How likely is it that the coin is actually fair?

Building a Bayesian Model

$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{p(D)}$$

1. **Select Prior**
2. Formulate Likelihood
3. **Analyze posterior**

Choosing a good prior

Uninformative vs. Informative

The prior represents our belief about Θ. If we don't know what Θ should be we want the prior to be uninformative.

Possible uninformative priors

- **Uniform prior**
  - Problem: Not Invariant to parameterization
- **Prior that contains no information at all**
  - Beta(0,0) - Problem: Not a proper prior (integrates to ∞)
- **Jeffrey's prior**
  - General purpose technique to create uninformative priors
  - $P(\Theta) = \sqrt{\mathbf{I}(\Theta)}$, where $\mathbf{I}$ is the Fisher Information
  - Example: $Beta(\frac{1}{2}, \frac{1}{2})$

Conjugacy

If the **prior** distribution is **conjugate** to the likelihood distribution then the **posterior** will come from the same distribution as the prior.

Benefits of Conjugacy

- Posterior is guaranteed to come from a **known family** of distributions (more importantly: well behaved family)
- Gives a simple **closed-form** solution for the posterior
- Intuitively shows how the likelihood updates the prior

Example of Beta-Binomial Conjugacy

$$Beta(a,b) = \frac{\Theta^{a-1}(1-\Theta)^{b-1}}{B(a,b)} \tag{1}$$

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{2}$$

$$p(D|\Theta) \propto \Theta^{S}(1-\Theta)^{N-S} \tag{3}$$

$$p(\Theta) \propto Beta(a,b) \tag{4}$$

Example of Beta-Binomial Conjugacy

$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{p(D)} \tag{5}$$

$$p(\Theta|D) \propto \frac{\Theta^{a+S-1}(1-\Theta)^{N+b-S-1}}{B(a+S, b+N-S)} \tag{6}$$

which is simply a $Beta(a+S, b+N-S)$ distribution

Intuition behind proof: Express Beta and Binomial with Gamma distributions. After some integration all but the above terms get canceled out.

Overview conjugate families

Beta

Bernoulli, Binomial

Dirichlet

Categorical, Multinomial

Gamma

Poisson, Exponential

Gaussian

Gaussian (depends an unknown parameter)

Mixtures of conjugate priors

Prior belief: A coin is either fair or is skewed towards heads.

Constructing a mixed conjugate prior

$$p(\Theta) = \sum_k p(z = k) p(\Theta | z = k) \tag{7}$$

$$p(\Theta | D) = \sum_k (p(z = k | D) p(\Theta | D, z = k) \tag{8}$$

where $p(z = k|D)$ are the **mixing weights** (e.g Z=[0.6,0.4]):

$$p(Z = k | D) = \frac{p(Z = k) p(D | Z = k)}{\sum_k' p(Z = k') p(D | Z = k')} \tag{9}$$

where $p(D|Z = k')$ is the marginal likelihood of the data under the k-th model

Example: Mixed Beta prior for Binomial

Mixture prior, $a_1 = b_1 = 20, a_2 = b_2 = 10$:

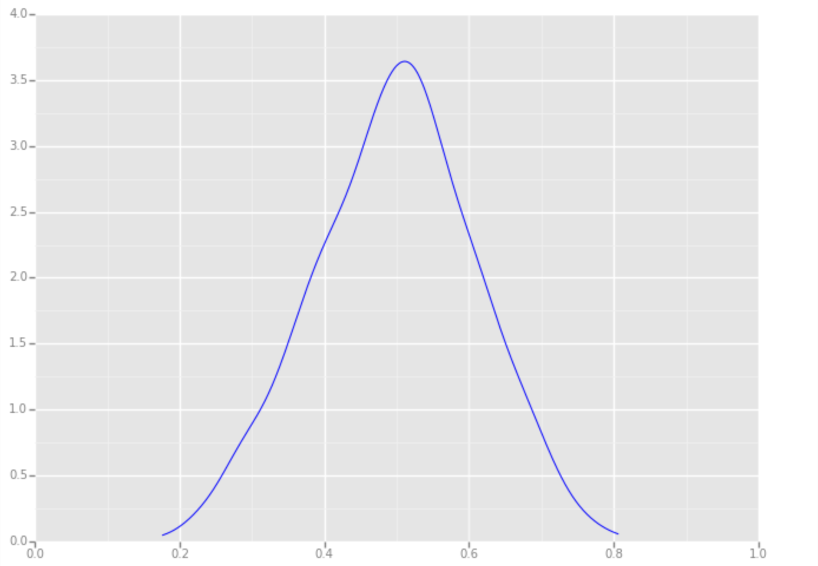
$$p(\Theta) = 0.5 Beta(\Theta | a_1, b_1) + 0.5 Beta(\Theta | a_2, b_2)$$

a’ and b’ are the updated parameters, mixture weights according to (9)

$$p(\Theta | D) = 0.346 Beta(a_1', b_1') + 0.654 Beta(a_2', b_2')$$

Analyzing the posterior

$Beta(10, 10)$ Posterior Distribution



Using a posterior distribution

Point Estimates

- Mean
- Median
- Mode (MAP Estimate)

Intervals

- Credible Intervals

Point Estimates for Beta Distribution

Estimate	Formula	Loss
Mean	$\frac{a}{a+b}$	Square loss
Median	$\frac{a-\frac{1}{3}}{a+b-\frac{2}{3}}$	Absolute loss
Mode	$\frac{a-1}{a+b-2}$	0-1 loss

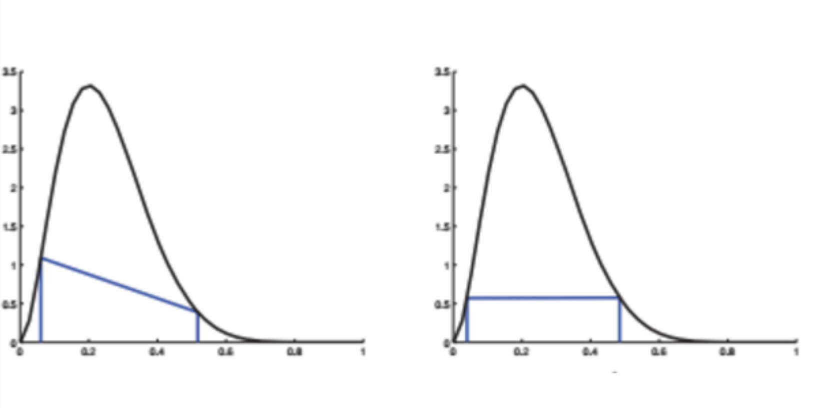
MAP Estimate

- Choosing the mode is called **Maximum a posteriori (MAP)** estimation
- MAP is the most popular choice due to computational convenience
- Several drawbacks:
    - Mode is an untypical point
    - Not invariant to reparameterization

Credible intervals

- Bayesian version of confidence intervals
- Region C under the probability density curve that contains $1 - \alpha$ of the posterior probability mass
- Different versions:
    - **Central Interval:** $(1 - \alpha)/2$ in each tail
    - **Highest posterior density region:** Set of most probable points that constitute $100(1 - \alpha)\%$ of the probability mass

Central vs High Posterior density intervals



Source: Murphy, 2013

Baysian vs. Frequentist

Difference between Credible Intervals and Confidence Intervals?

Selecting a model

Bayesian Model Selection

Instead of cross validation we can compute the posterior over all models

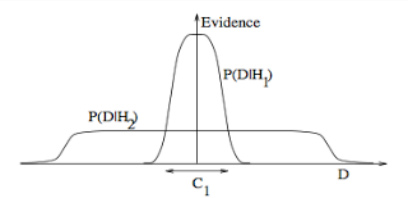
$$p(m|D) = \frac{p(D|m)p(m)}{\sum_{m \in M} p(m, D)} \tag{10}$$

For uniform priors: Pick model with maximal *evidence/marginal likelihood*

$$p(D|m) = \int p(D|\Theta)p(\Theta|m)d\Theta \tag{11}$$

Bayesian Occam's razor

- Bayesian Model Selection naturally guards against overfitting
- Since probability mass integrates to 1, more complex models have a lower probability for a specific dataset



Source: MacKay, 1995

Computing the Evidence

Conjugate Priors have closed-form solution

$$p(D) = \frac{Z_n}{Z_0 Z_l}$$

$Z_n =$ normalizing constant in posterior
$Z_0 =$ normalizing constant in prior
$Z_l =$ constants in likelihood

Otherwise we need to use sampling or approximation **(BIC)**

Evidence for Beta-Binomial model

$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{p(D)} \tag{12}$$

$$p(\Theta|D) = \frac{1}{p(D)}\frac{1}{B(a,b)}\Theta^{a-1}(1-\Theta)^{b-1}\binom{N}{N_1}\Theta^{N_1}(1-\Theta)^{N_0} \tag{13}$$

$$p(\Theta|D) = \binom{N}{N_1}\frac{1}{p(D)}\frac{1}{B(a,b)}\Theta^{a+N_1-1}(1-\Theta)^{b+N_0-1} \tag{14}$$

Evidence for Beta-Binomial model

Divide by Θ term

$$\frac{1}{B(a+N_1, b+N_0)} = \binom{N}{N_1} \frac{1}{p(D)} \frac{1}{B(a,b)} \tag{15}$$

$$p(D) = \binom{N}{N_1} \frac{B(a+N_1, b+N_0)}{B(a,b)} \tag{16}$$

Conclusion

Summary - What have we learned

- When choosing a prior make sure to try and pick a (mixed) conjugate prior that encodes just the information you have
- There are various ways to use the posterior distribution based on the loss you want to minimize
- When selecting a Bayesian model we look at the evidence it provides given the data

Questions?

Excellent video lectures:
https://goo.gl/oPb6pG

References I

H. S. S. A. Gelman, J.B. Carlin.
Bayesian Data Analysis.
2014.

K. Murphy.
Machine Learning: A Probabilistic Perspective.
2013.