CS 5135/6035 Learning Probabilistic Models

Lecture 14: Conjugacy, Posterior Summarization

Gowtham Atluri

October 18, 2018

Reading Material

- Larry Wasserman, Lecture Notes 14 Bayesian Inference
http://www.stat.cmu.edu/~larry/=stat705/Lecture14.pdf
- Gelman et al. Bayesian Data Analysis
  - Chapter 2. Single Parameter Models

Topics

- Informative prior
- Solving for the posterior
  - hard vs. easy way
- Conjugacy
- Summarizing the posterior
  - Point estimate
  - Interval estimate

Bayesian Parameter Estimation

For point or interval estimation of a parameter θ in a model M based on data y, Bayesian inference is based off

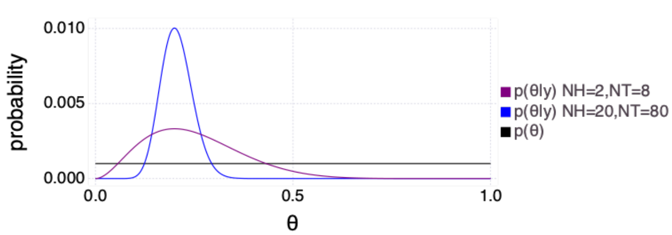
$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \propto p(y|\theta)p(\theta)$$

where

- $p(\theta)$ is the prior distribution for the parameter,
- $p(\theta|y)$ is the posterior distribution for the parameter,
- $p(y|\theta)$ is the statistical model (or likelihood), and
- $p(y)$ is the prior predictive distribution (or marginal likelihood).

Bayesian Estimation: Uninformative prior

- *Scenario:* Is the coin used in betting a fair coin?
  - We saw 2 heads and 8 tails in 10 trials
- A flat prior $p(\theta) = k$
- Likelihood $p(y|\theta) = \theta^{N_H}(1-\theta)^{N_T}$
- The posterior is $p(\theta|y_1, \dots, y_n) \propto \theta^{N_H}(1-\theta)^{N_T}$
  - This is a Beta distribution $Beta(N_H + 1, N_T + 1)$
- As more samples are available, less is the uncertainty in the posterior



Choosing Prior

- How do we construct/choose prior distributions?
- Two interpretations:
  - *Population* interpretation
    - Prior distribution represents a population of possible parameter values from which θ has been drawn
  - *Knowledge* interpretation
    - We must express our knowledge about θ as if its value could be thought of as a random realization from the prior distribution.
- In many applications there is no perfectly relevant population of θ's from which the current θ has been drawn.

General guidelines:

- Prior distribution should include all possible values of θ
- Prior need not be realistically concentrated around the 'true' value.

Information about θ contained in the data will far outweigh any reasonable prior specification.

Informative Prior

- Let us consider using a *Beta* prior $\theta \sim Beta(\alpha, \beta)$

$$Beta(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Interpretation of *information* in the prior

- Compare this prior to the previous posterior under uniform prior
- $Beta(a, b)$ is equivalent to $a-1$ priori successes and $b-1$ prior failures.

Hyperparameters

- Parameters of the priori distribution are referred to as hyperparameters
  - These are assumed to be known
- Beta prior is indexed by two hyperparameters (a, b)
- We are essentially fixing two features of the dist. (e.g., mean and variance)

Obtaining the posterior

The hard way:

- Derive $p(y)$
- Derive $p(\theta|y) = p(y|\theta)p(\theta)/p(y)$

The easy way:

- Derive $f(\theta) \propto p(y|\theta)p(\theta)$
- Recognize $f(\theta)$ as the kernel of some distribution

Definition

The kernel of a probability density (mass) function is the form of the pdf (pmf) with any terms not involving the random variable omitted.

For example, $\theta^{N_H}(1-\theta)^{N_T}$ is the kernel of a Beta distribution.

Derive the posterior - the hard way

Likelihood $p(y|\theta) = \theta^{N_H}(1-\theta)^{N_T}$ Prior $p(\theta) = Beta(a, b)$

$$\begin{aligned} p(y) &= \int p(y|\theta)p(\theta)d\theta \\ &= \int \left(\theta^{N_H}(1-\theta)^{N_T}\right)\left(\tfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right)d\theta \\ &= \tfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int \theta^{a+N_H-1}(1-\theta)^{b+N_T-1}d\theta \\ &= \tfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \tfrac{\Gamma(a+N_H)\Gamma(b+N_T)}{\Gamma(a+N_H+b+N_T)} \end{aligned}$$

which is known as the Beta-binomial distribution.

$$\begin{aligned} p(\theta|y) &= p(y|\theta)p(\theta)/p(y) \\ &= \frac{\left(\theta^{N_H}(1-\theta)^{N_T}\right)\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right)}{\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+N_H)\Gamma(b+N_T)}{\Gamma(a+N_H+b+N_T)}\right)} \\ &= \tfrac{\Gamma(a+N_H+b+N_T)}{\Gamma(a+N_H)\Gamma(b+N_T)}\theta^{a+N_H-1}(1-\theta)^{b+N_T-1} \\ &= Beta(a+N_H, b+N_T) \end{aligned}$$

Thus $\theta|y \sim Beta(a+N_H, b+N_T)$.

Derive the posterior - the easy way

Likelihood $p(y|\theta) = \theta^{N_H}(1-\theta)^{N_T}$ Prior $p(\theta) = Beta(a, b)$

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \left(\theta^{N_H}(1-\theta)^{N_T}\right)\left(\tfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right) \\ &\propto \tfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a+N_H-1}(1-\theta)^{b+N_T-1} \\ &\propto \theta^{a+N_H-1}(1-\theta)^{b+N_T-1} \end{aligned}$$

Thus $\theta|y \sim Beta(a+N_H, b+N_T)$.

Note that the posterior follows the same parametric form as the prior.

Conjugacy

If the posterior is of the same parametric form as the prior, then we call the prior the conjugate distribution for the likelihood distribution.

$$\underbrace{p(\theta|y)}_{\text{posterior}} \propto \underbrace{p(y|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$

For a prior on parameter θ, with hyperparameter α, $p(\theta|\alpha)$,
the posterior given data y is the same form as the prior, but with updated hyperparameters, $p(\theta|y, \alpha) = p(\theta|\alpha')$

Example: For a Bernoulli family likelihood $p(y|\theta) = \theta^{N_H}(1-\theta)^{N_T}$ and a Beta prior with hyperparameters a, b, $p(\theta|a, b) = Beta(a, b)$

Posterior is $p(\theta|y, a, b) \propto p(y|\theta)p(\theta|a, b) = Beta(a+N_H, b+N_T)$

Conjugacy

Discrete distributions

Sample Space	Sampling Dist.	Conjugate Prior	Posterior
$y \in \{0, 1\}$	*Bernoulli*	*Beta*	*Beta*
$y = \mathbb{Z}_+$	*Poisson*	*Gamma*	*Gamma*
$y = \mathbb{Z}_{++}$	*Geometric*	*Gamma*	*Gamma*
$y = \mathbb{H}_K$	*Multinomial*	*Dirichlet*	*Dirichlet*

Continuous distributions

Sampling Dist.	Conjugate Prior	Posterior
$Exponential(\theta)$	$Gamma(\alpha, \beta)$	*Gamma*
$\mathcal{N}(\mu, \sigma^2)$, known σ^2	$\mathcal{N}(\mu_0, \sigma_0^2)$	*Gaussian*
$\mathcal{N}(\mu, \sigma^2)$, known μ	*InvGamma*	*InvGamma*
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, known $\boldsymbol{\Sigma}$	$\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0^2)$	*Gaussian*
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, known $\boldsymbol{\mu}$	*InvWishart*	*InvWishart*

Conjugacy

Advantages

- Interpretability
  - E.g., when data y are generated from $Bernoulli(\theta)$
  - a $Beta(a, b)$ prior is equivalent to $a - 1$ priori successes and $b - 1$ prior failures
- Computational convenience
  - We can easily determine the posterior!

```
function Beta_posterior(a,b,NH,NT)
    # a and b are the hyperparameters of the Beta prior
    # NH are the number of successful trials
    return Beta(a+NH, b+NT)
end
```


Summarizing the posterior

- Posterior distribution contains all the *current* info. about the parameter θ
- Ideally one may report the entire probability distribution $p(\theta|y)$
  - A graphical display is useful
- Bayesian estimation provides flexibility of summarizing posterior
- Two ways:
  - Point Estimate: most likely guess
    - mean
    - median
    - mode
  - Interval Estimate
    - Equal-tailed
    - One-sided
    - Highest posterior density

Bayes Risk

The Bayes Risk of an estimate $\hat{\theta}$ can be assessed by how much we believe we missed the true θ.

More formally, Bayes Risk is computed as the expectation of the loss function $L(\theta, \hat{\theta})$ over the posterior $p(\theta|y)$.

$$Risk = \int L(\theta, \hat{\theta})\ p(\theta|y)$$

Common estimators:

- Mean: $\hat{\theta}_{Bayes} = E[\theta|y]$ minimizes $L\left(\theta, \hat{\theta}\right) = \left(\theta - \hat{\theta}\right)^2$
- Median: $\int_{\hat{\theta}_{Bayes}}^{\infty} p(\theta|y)d\theta = \frac{1}{2}$ minimizes $L\left(\theta, \hat{\theta}\right) = \left|\theta - \hat{\theta}\right|$
- Mode: $\hat{\theta}_{Bayes} = \text{argmax}_{\theta}\, p(\theta|y)$ is obtained by minimizing $L\left(\theta, \hat{\theta}\right) = -\mathbb{I}\left(|\theta - \hat{\theta}| < \epsilon\right)$ as $\epsilon \to 0$, also called maximum a posterior (MAP) estimator.

Bayes Risk with quadratic loss

$$Risk = \int L(\theta, \hat{\theta})\ p(\theta|y) = \int \underbrace{(\hat{\theta} - \theta)^2}_{\text{Quadratic Loss}}\ p(\theta|y)$$

Let μ_p and σ_p^2 denote the mean and variance of the posterior distribution.

$$= \int (\hat{\theta} - \mu_p + \mu_p - \theta)^2\ p(\theta|y)$$
$$= \int (\hat{\theta} - \mu_p)^2\ p(\theta|y) + \int (\mu_p - \theta)^2\ p(\theta|y) + 2 \int (\hat{\theta} - \mu_p)(\mu_p - \theta) p(\theta|y) d\theta$$
$$= (\hat{\theta} - \mu_p)^2 + \sigma_p^2 + 2(\hat{\theta} - \mu_p) \int (\mu_p - \theta) p(\theta|y) d\theta$$
$$= (\hat{\theta} - \mu_p)^2 + \sigma_p^2 + 0$$

Bayes risk of an estimate $\hat{\theta}$: $(\hat{\theta} - \mu_p)^2 + \sigma_p^2$

Point Estimate: Mean (Bayes Estimate)

The Bayes risk of the estimate of the parameter $\hat{\theta}$

$$MSE = (\hat{\theta} - \mu_p)^2 + \sigma_p^2$$

This is minimized when $\hat{\theta} = \mu_p$.
i.e., estimate $\hat{\theta}$ is the same as the mean of the posterior.

For this reason, some Bayesians prefer to use the mean of the posterior probability.

$$\mu_p = E_{p(\theta|y)}(\theta) = \int \theta p(\theta|y)$$

The estimate μ_p is referred to as the Bayes estimate.

Point Estimate: Mean (Bayes Estimate)

Example: Coin toss experiment

- Posterior $p(\theta|y) = \text{Beta}(a + N_H, b + N_T)$
- Bayes Estimate is $E_{p(\theta|y)}(\theta)$, i.e, mean of the posterior.
- Mean of $Beta(a, b) = \frac{a}{a+b}$
- Bayes Estimate is $\frac{a+N_H}{a+b+N_H+N_T}$

Point Estimate: Bayes Risk with Absolute Error Loss

$$Risk = \int L(\theta, \hat{\theta})\ p(\theta|y) = \int \underbrace{|\hat{\theta} - \theta|}_{\text{Abs Error Loss}}\ p(\theta|y)$$

$$= \int_{-\infty}^{\hat{\theta}} (\hat{\theta} - \theta)\ p(\theta|y) + \int_{\hat{\theta}}^{\infty} (\theta - \hat{\theta})\ p(\theta|y)$$

$$= \hat{\theta} \int_{-\infty}^{\hat{\theta}} p(\theta|y) - \int_{-\infty}^{\hat{\theta}} \theta p(\theta|y) + \int_{\hat{\theta}}^{\infty} \theta p(\theta|y) - \hat{\theta} \int_{\hat{\theta}}^{\infty} p(\theta|y)$$

$$= 0, \text{ when } \int_{-\infty}^{\hat{\theta}} p(\theta|y) = \int_{\hat{\theta}}^{\infty} p(\theta|y)$$

This happens when both sides of $\hat{\theta}$ are equal in area.
This happens at the **median**.

Therefore median of the posterior is the Bayesian estimate to minimize Absolute Error Loss

Point Estimate: Most probable A Posteriori

The Most probable A Posteriori (MAP) setting is that which maximizes the posterior.

$$\theta^* = \arg\max_{\theta} p(\theta|y)$$

Example: Coin-toss experiment

- Posterior $p(\theta|y) \propto \theta^{a+N_H-1}(1-\theta)^{b+N_T-1}$
- To maximize the posterior $p(\theta|y)$, take log on both sides and differentiate w.r.t. θ
- $\log p(\theta|y) \propto (a + N_H - 1)\log\theta + (b + N_T - 1)\log(1-\theta)$
- $$\frac{\partial}{\partial\theta}\log p(\theta|y) = \frac{a + N_H - 1}{\theta} - \frac{b + N_T - 1}{1-\theta} = 0$$
- $$\hat{\theta} = \frac{\alpha + N_H - 1}{N_H + N_T + \alpha + \beta - 2}$$

MAP Estimate

The Most probable A Posteriori (MAP) setting is that which maximizes the posterior.

$$\hat{\theta} = \arg\max_{\theta} p(\theta|y)$$

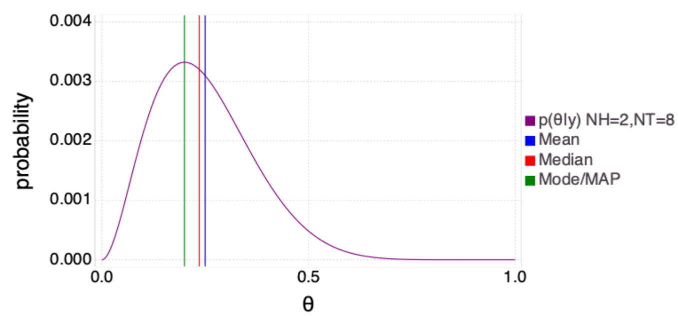
$$Risk = \int L(\theta, \hat{\theta})\ p(\theta|y)$$

What Loss function is a MAP estimate optimizing?

$$L(\theta, \hat{\theta}) = \begin{cases} 0, & \text{if } |\theta - \hat{\theta}| < \epsilon \\ 1, & \text{otherwise} \end{cases}$$

As $\epsilon \to 0$, the Bayes estimator approaches the MAP estimator. This is only for discrete cases.

Point Estimation




```
d = Beta(3,9);
[Base.mean(d) Base.median(d) Distributions.modes(d)]
```

```
## 1×3 Array{Float64,2}:
```


Interval estimation

Definition

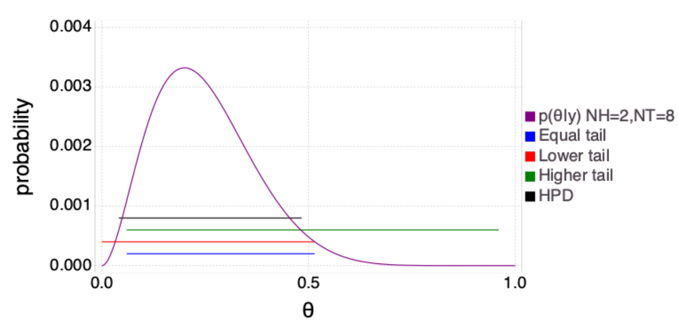
A $100(1-a)\%$ credible interval is any interval (L,U) such that

$$1 - a = \int_L^U p(\theta|y) d\theta.$$

Some typical intervals are

- Equal-tailed: $a/2 = \int_{-\infty}^{L} p(\theta|y)d\theta = \int_U^{\infty} p(\theta|y)d\theta$
- One-sided: either $L = -\infty$ or $U = \infty$
- Highest posterior density (HPD): $p(L|y) = p(U|y)$ for a uni-modal posterior which is also the shortest interval
  - one with the smallest interval width among all credible intervals

Interval estimation




```
quantile(Beta(3,9),[0.025,0.975])'
```

```
## 1×2 RowVector{Float64,Array{Float64,1}}:
##  0.0602177  0.517756
```

Summary

- Conjugacy
  - If the posterior is of the same parametric form as the prior, then we call the prior the conjugate distribution for the likelihood distribution.
- Conjugacy helps with
  - interpretation
  - computaitonal convenience
- Summarizing the posterior
  - Point estimate
    - mean
    - median
    - mode
  - Interval estimate
    - Equal-tailed
    - One-sided
    - Highest posterior density