CS 5135/6035 Learning Probabilistic Models

Lecture 15: Natural Conjugacy, Mixture of Priors

Gowtham Atluri

October 18, 2018

Reading Material

- Kevin Murphy, Conjugate Bayesian analysis of the Gaussian distribution
  https://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf
- Gelman et al. Bayesian Data Analysis
  - Chapter 2. Single Parameter Models
- Jim Albert, Bayesian Computation With R, 2nd Ed.
  - Chapter 3. Single-Parameter Models

Topics

- Natural Conjugacy
  - Exponential Distributions
- Bayesian estimation for Gaussian
  - unknown mean and known variance
- Mixture of Priors
  - Mixture of Beta priors

Priors/Conjugacy

Prior

A prior distribution of an uncertain quantity θ is the probability distribution that would express one's uncertainty about θ before the "data" is taken into account.

Conjugate Prior

A prior $p(\theta)$ is conjugate if for $p(\theta) \in \mathcal{P}$ and $p(y|\theta) \in \mathcal{F}$, $p(\theta|y) \in \mathcal{P}$ where $\mathcal{F}$ and $\mathcal{P}$ are standard distributions.

For example, the beta distribution ($\mathcal{P}$) is conjugate to the binomial distribution with unknown probability of success ($\mathcal{F}$) since

$$\theta \sim Beta(a, b) \qquad \text{and} \qquad \theta|y \sim Beta(a + N_H, b + N_T).$$

Natural conjugate prior

Natural conjugate

A natural conjugate prior is a conjugate prior that has the same functional form as the likelihood.

- For example, the beta distribution is a natural conjugate prior since

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1} \qquad \text{and} \qquad L(\theta) \propto \theta^{y}(1-\theta)^{n-y}.$$

- Probability distributions that belong to an exponential family have natural conjugate prior distributions.
  - This is the only class of distributions that have natural conjugate prior distributions

Exponential Family

- A random variable y has a distribution from an exponential family model $\mathcal{F}$ if the density of y is of the form

$$p(y|\boldsymbol{\theta}) = h(y) exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{T}(y) - \psi(\boldsymbol{\theta})\Big)$$

Exponential Family

- A random variable y has a distribution from an exponential family model $\mathcal{F}$ if the density of y is of the form

$$p(y|\boldsymbol{\theta}) = h(y)exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T\boldsymbol{T}(y) - \psi(\boldsymbol{\theta})\Big)$$

- $\boldsymbol{\eta}(\boldsymbol{\theta})$ and $\boldsymbol{T}(y)$ are, in general, vectors of dimension same as that of $\boldsymbol{\theta}$.
- $\boldsymbol{\eta}(\boldsymbol{\theta}) = [\eta_1, \eta_2, \dots, \eta_d]^T$ is called the 'natural parameter' of the exponential family $\mathcal{F}$
- $\boldsymbol{T}(y)$ is a 'sufficient statistic' for $\boldsymbol{\theta}$
  - A sufficient statistic for θ contains all the information in the sample about θ
  - We cannot improve our knowledge about $\boldsymbol{\theta}$ by a detailed analysis of data $y_1, \dots, y_n$
- $\psi(\boldsymbol{\theta})$ is the log partition function that ensures normalization

$$\psi(\boldsymbol{\theta}) = log\int_x h(y)exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T\boldsymbol{T}(y) - \psi(\boldsymbol{\theta})\Big)$$

Exponential Family

- A random variable y has a distribution from an exponential family model $\mathcal{F}$ if the density of y is of the form

$$p(y|\boldsymbol{\theta}) = h(y)exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T\boldsymbol{T}(y) - \psi(\boldsymbol{\theta})\Big)$$

- Exponential family contains many standard distributions

Discrete	Continuous
Bernoulli	Beta
Categorical	Chi-squared
Geometric	Exponential
Poisson	Gamma
	Gaussian

Example: Beta Distribution

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

$$p(y|\boldsymbol{\theta}) = h(y)exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T\boldsymbol{T}(y) - \psi(\boldsymbol{\theta})\Big)$$

Here y is the rand. var. and parameters $\boldsymbol{\theta} = [\alpha, \beta]$

Beta distribution can be rewritten as:

$$[y(1-y)]^{-1}\exp(\alpha\log(y) + \beta\log(1-y) + \log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta))$$

Now we can see that this is in the exponential family form, where

- $h(y) = (y(1-y))^{-1}$
- $\boldsymbol{\eta}(\boldsymbol{\theta}) = [\alpha, \beta]^T$
- $\boldsymbol{T}(y) = [\log(y), \log(1-y)]^T$
- $\psi(\boldsymbol{\theta}) = -\log\Gamma(\alpha+\beta) + \log\Gamma(\alpha) + \log\Gamma(\beta)$

Example: Gaussian $\rightarrow h(y)exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T\boldsymbol{T}(y) - \psi(\boldsymbol{\theta})\Big)$

Consider the univariate Gaussian

$$\mathcal{N}(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}exp\Big(-\frac{1}{2\sigma^2}(y-\mu)^2\Big)$$

Here y is the rand. var. and parameters $\boldsymbol{\theta} = [\mu, \sigma^2]$

This can be expanded as

$$exp\Big(-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log 2\pi\sigma^2\Big)$$

Now we can see that this is in the exponential family form, where

- $h(y) = 1$
- $\boldsymbol{\eta}(\boldsymbol{\theta}) = [1/\sigma^2,\ \mu/\sigma^2]^T$
- $\boldsymbol{T}(y) = [-y^2/2,\ y]^T$
- $\psi(\boldsymbol{\theta}) = \frac{\mu^2}{2\sigma^2} + \log(2\pi\sigma^2)$

This parameterization is not unique (rescale $T_i(x)$, inversely scale η_i).

Natural conjuage prior

Exponential family:

$$p(y|\boldsymbol{\theta}) = h(y)exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T\boldsymbol{T}(y) - \psi(\boldsymbol{\theta})\Big)$$

Likelihood:

$$p(y|\boldsymbol{\theta}) = \Big(\prod_{i=1}^{n} h(y_i)\Big)exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T\sum_{i=1}^{n}\boldsymbol{T}(y_i) - n\psi(\boldsymbol{\theta})\Big)$$

Prior:

$$p(\boldsymbol{\theta}|\boldsymbol{\chi},\boldsymbol{\nu}) \propto exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T\boldsymbol{\chi} - \boldsymbol{\nu}\psi(\boldsymbol{\theta})\Big)$$

where

- $\boldsymbol{\nu}$ corresponds to the # psuedo-observations prior contributes
- $\boldsymbol{\chi}$ is the # psuedo-observations that contribute to the sufficient statistic

$$p(y|\boldsymbol{\theta}) \propto exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T[\chi + \sum_{i=1}^{n}\boldsymbol{T}(y_i)] - (n+\nu)\psi(\boldsymbol{\theta})\Big)$$

Posterior: is also in the same form as the prior

Estimating parameters of a Gaussian (only unknown is μ)

- Given a training data $y = \{y_1, \dots, y_n\}$ drawn *i.i.d* from a Gaussian $\mathcal{N}(y|\mu,\sigma^2)$ with unknown mean μ and a given variance σ^2

$$\mathcal{N}(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}exp\Big(-\frac{1}{2\sigma^2}(y-\mu)^2\Big)$$

- Likelihood is written as

$$p(y|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\Big)$$

- As we know σ^2 and μ is the only unknown, posterior can be written as

$$p(\mu|y) \propto p(y|\mu,\sigma^2)p(\mu|\mu_0,\sigma_0^2)$$

$p(\mu|\mu_0,\sigma_0^2)$ is the prior with hyperparameters μ_0 and σ_0^2.

Estimating parameters of a Gaussian (only unknown is μ)

- We need to determine the posterior
$$p(\mu|y) \propto p(y|\mu,\sigma^2)p(\mu|\mu_0,\sigma_0^2)$$
- We know data is drawn *i.i.d* from a Gaussian $\mathcal{N}(y|\mu,\sigma^2)$
$$p(y|\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left[\frac{1}{2\sigma^2}(y-\mu)^2\right]$$
- As the likelihood function is a Gaussian, if we choose a Gaussian prior, the posterior will also be a Gaussian.
$$p(\mu) = (2\pi\sigma_0^2)^{-1/2}\exp\left[\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]$$
- Plugging in the likelihood function and prior, the posterior is
$$p(\mu|y) \propto \exp\left[\frac{1}{2\sigma^2}(y-\mu)^2\right]\times\exp\left[\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]$$

Estimating parameters of a Gaussian (only unknown is μ)

$$p(\mu|y) \propto \exp\left[\frac{1}{2\sigma^2}(y-\mu)^2\right]\times\exp\left[\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]$$

- Since the product of two Gaussians is a Gaussian, we write the posterior Gaussian as
$$\mathcal{N}(\mu_p,\sigma_p^2) \propto \exp[-\frac{1}{2\sigma_p^2}(\mu-\mu_p)^2] = \exp[-\frac{1}{2\sigma_p^2}(\mu^2-2\mu\mu_p+\mu_p^2)]$$
where μ_p and σ_p are parameters of the posterior Gaussian form.
- We rewrite our above posterior as...
$$p(\mu|y) \propto \exp\left[-\frac{\mu^2}{2}\Big(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\Big)+\mu\Big(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_i y_i}{\sigma^2}\Big)-\Big(\frac{\mu_0^2}{2\sigma_0^2}+\frac{\sum_i y_i^2}{2\sigma^2}\Big)\right]$$
- By matching coefficients of μ^2 from the above two eqns...
$$-\frac{\mu^2}{2\sigma_p^2} = -\frac{\mu^2}{2}\Big(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\Big); \quad \frac{1}{\sigma_p^2} = \frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}; \quad \sigma_p^2 = \frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}$$

Estimating parameters of a Gaussian (only unknown is μ)

Posterior
$$p(\mu|y) \propto \exp\left[-\frac{\mu^2}{2}\Big(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\Big)+\mu\Big(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_i y_i}{\sigma^2}\Big)-\Big(\frac{\mu_0^2}{2\sigma_0^2}+\frac{\sum_i y_i^2}{2\sigma^2}\Big)\right]$$

The Gaussian form we want to transform it to is...
$$\exp[-\frac{1}{2\sigma_p^2}(\mu^2-2\mu\mu_p+\mu_p^2)] = \exp[-\frac{1}{2\sigma_p^2}(\mu-\mu_p)^2]$$

Matching the coefficients of μ we get...
$$\frac{-2\mu\mu_p}{\sigma_p^2} = \mu\Big(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_i y_i}{\sigma^2}\Big); \quad \frac{\mu_p}{\sigma_p^2} = \frac{\mu_0}{\sigma_0^2}+\frac{\sum_i y_i}{\sigma^2}; \quad \mu_p = \sigma_p^2\Big(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_i y_i}{\sigma^2}\Big)$$

This process of matching the first power and second powers of μ is called **completing the square**

Estimating parameters of a Gaussian (only unknown is μ)

- Given a training data $y = \{y_1,\ldots,y_n\}$ drawn *i.i.d* from a Gaussian $\mathcal{N}(y|\mu,\sigma^2)$ with unknown mean μ and a given variance σ^2
$$\mathcal{N}(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}exp\Big(-\frac{1}{2\sigma^2}(y-\mu)^2\Big)$$
- Choosing a Gaussian prior over μ
$$p(\mu) = (2\pi\sigma_0^2)^{-n/2}\exp\left[\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]$$
- Our posterior over parameter μ
$$\mathcal{N}(\mu|\mu_p,\sigma_p^2) = \frac{1}{\sqrt{2\pi\sigma_p^2}}exp\Big(-\frac{1}{2\sigma_p^2}(\mu-\mu_p)^2\Big)$$

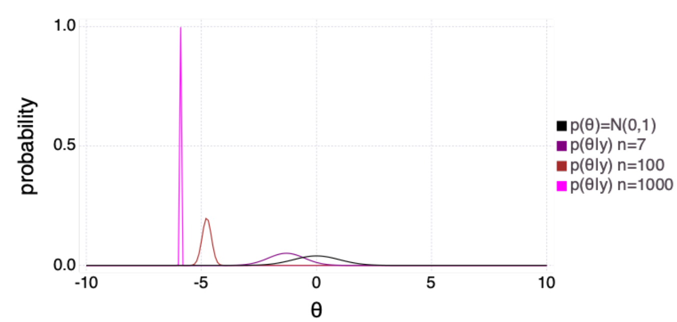
where
$$\mu_p = \sigma_p^2\Big(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_i y_i}{\sigma^2}\Big); \qquad \sigma_p^2 = \frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}$$

Estimating parameters of a Gaussian - alternative view

- Let us consider the inverse of variance, also referred to as **precision**.
$$\lambda = \frac{1}{\sigma^2}; \quad \lambda_0 = \frac{1}{\sigma_0^2}; \quad \lambda_p = \frac{1}{\sigma_p^2}$$
- The parameter of the posterior from our computation were
$$\sigma_p^2 = \frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}; \quad \mu_p = \sigma_p^2\Big(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_i y_i}{\sigma^2}\Big)$$
- We can now rewrite the posterior as $p(\mu|y,\lambda) = \mathcal{N}(\mu|\mu_p,\lambda_p)$ where
$$\lambda_p = \lambda_0 + n\lambda \qquad \mu_p = \frac{\mu_0\lambda_0+\lambda\sum_i y_i}{\lambda_p} = (1-w)\mu_0 + w\mu_{ML} \text{where } w = \frac{n\lambda}{\lambda_p}$$
- The precision of the posterior λ_p is the precision of the prior λ_0 plus one contribution of data precision λ for each observed data point
- Mean of the posterior is a convex combination of prior and MLE.

Revisiting the temperature example

The temperatures, in Celsius, in Minneapolis during the first week of March 2018 are observed as $(-2.5, -9.9, -12.1, -8.9, -6.0, -4.8, 2.4)$
What is the distribution from which this data was generated (assuming it was Gaussian and $\sigma^2 = 25$)?



Mixture of Priors

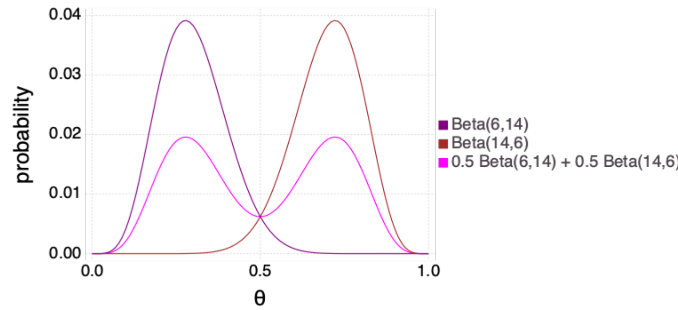
- Suppose we know that a coin has significant bias, but we don't know if the coin is biased towards heads or tails
- If θ is the probability that a coin lands heads
  - we believe that θ is either in the neighborhood of 0.3 or 0.7
  - we believe that it is equally likely that θ is in one of the two neighborhoods
- This belief can be modeled using the prior density

$$p(\theta) = \pi p_1(\theta) + (1-\pi)p_2(\theta))$$

  - where $p_1(\theta) = Beta(6, 14)$ and $p_2(\theta) = Beta(14, 6)$
  - mixing probability is 0.5.

Mixture of Beta Priors

$\frac{1}{2}Beta(6, 14) + \frac{1}{2}Beta(14, 6)$



Mixtures of conjugate priors are conjugate

The general case:
Mixture of priors: $[p_1(\theta), p_2(\theta), \ldots, p_k(\theta)]$
Selection probabilities: $[\pi_1, \pi_2, \ldots, \pi_k]$

Let $\pi_i = P(H_i)$ and $p_i(\theta) = p(\theta|H_i)$,

$$\theta \sim \sum_{i=1}^{k} \pi_i p_i(\theta) \qquad \sum_{i=1}^{k} \pi_i = 1$$

We want to derive the equation for $p(\theta|y)$ when a mixture of priors is used.

Mixtures of conjugate priors are conjugate

$$\theta \sim \sum_{i=1}^{k} \pi_i p_i(\theta) \qquad \sum_{i=1}^{k} \pi_i = 1$$

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

$$\begin{aligned} p(\theta|y) &= \frac{1}{p(y)}p(y|\theta)p(\theta) = \frac{1}{p(y)}p(y|\theta)\sum_{i=1}^{k}\pi_i p_i(\theta) \\ &= \frac{1}{p(y)}\sum_{i=1}^{k}\pi_i p(y|\theta)p_i(\theta) = \frac{1}{p(y)}\sum_{i=1}^{k}\pi_i p_i(y)p_i(\theta|y) \\ &= \sum_{i=1}^{k}\frac{\pi_i p_i(y)}{p(y)}p_i(\theta|y) = \sum_{i=1}^{k}\frac{\pi_i p_i(y)}{\sum_{j=1}^{k}\pi_j p_j(y)}p_i(\theta|y) \end{aligned}$$

Mixture of Beta Priors

Recall:

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

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

- Mixture of priors $p(\theta) = \pi Beta(a_1, b_1) + (1-\pi)Beta(a_2, b_2)$

Then

$$p(\theta|y) = \pi' Beta(a_1 + N_H, b_1 + N_T) + (1-\pi')Beta(a_2 + N_H, b_2 + N_T)$$

$$\pi' = \frac{\pi p_1(y)}{\pi p_1(y) + (1-\pi)p_2(y)} \qquad p_i(y) = \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)}\frac{\Gamma(a_i + N_H)\Gamma(b_i + N_T)}{\Gamma(a_i + N_H + b_i + N_T)}$$

Mixture of Priors

- Mixture of priors

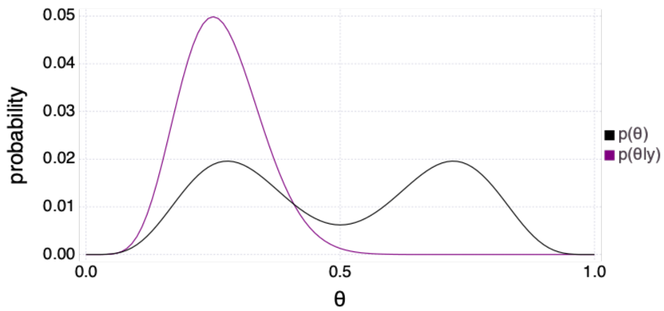
$$p(\theta) = \pi p_1(\theta) + (1-\pi)p_2(\theta)$$

  - where $p_1(\theta) = Beta(6, 14)$ and $p_2(\theta) = Beta(14, 6)$
  - mixing probability is 0.5.
- Data: 2 Heads, 8 Tails
- Posterior

$$p(\theta|y) = \pi' Beta(6+2, 14+8) + (1-\pi')Beta(14+2, 6+8)$$

$$\pi' = \frac{0.5p_1(y)}{0.5p_1(y) + 0.5p_2(y)} \quad p_i(y) = \frac{(a_i + b_i - 1)!}{(a_i - 1)!(b_i - 1)!}\frac{(a_i + N_H - 1)!(b_i + N_T -}{(a_i + N_H + b_i + N_T - 1)}$$

Mixture of Priors



Summary

- Natural conjugacy
  - Prior and likelihood are of the same form
  - Exists only for Exponential family of distributions
- Bayesian estimation for Gaussian
  - Unknown mean, and known variance
  - Determining parameters of posterior by *completing the square*
- Mixture of priors
  - Mixture of priors are also conjugate
  - Need to update mixing coefficients in the posterior
  - Coin-toss example with a mixture of Beta priors