STA 360 Final Exam Review

Rebecca C. Steorts

Course evalautions

- Please take 10-15 mintues to fill out the course evaluations for the course.
- If students could also please fill out TA evaluations this would be very helpful.
- If there is 100 percent response for these as well, there will be an extra incentive for the class!

Review of Bayesian Methods (Module 1)

- Traditional inference
- Bayes' Theorem
- Conjugate Distributions
- Marginal likelihood and posterior predictive distribution

Module 1: https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-1/01-intro-to-Bayes.pdf

Traditional inference

Suppose I have data x and I wish to estimate an **unknown, fixed** parameter θ using traditional (frequentist) inference.

- I could use maximum likelihood estimation (MLE)
- I could find an unbiased estimator of θ.

Goal of this course: we will instead assume θ is a **unknown, random variable**, and put a distribution θ.

Bayes' Theorem

Recall Bayes' Theorem:

$$p(\theta \mid x) = \frac{p(\theta, x)}{p(x)} \tag{1}$$

$$= \frac{p(x \mid \theta)p(\theta)}{p(x)} \tag{2}$$

$$\propto p(x \mid \theta)p(\theta) \tag{3}$$

$$\tag{4}$$

This says that the **posterior** is proportional to the **likelihood** times the **prior**

Conjugacy

If the posterior distribution comes from the same family of distributions as the prior, we say that the prior and posterior are conjugate distributions.

More formally, the prior is called a conjugate family for the likelihood function.

Example: The **Beta prior** is conjugate to the **Bernoulli likelihood function.**

Marginal likelihood and posterior predictive distribution

$p(x)$: marginal likelihood

- This is the normalizing constant in Bayes' theorem, that sometimes we cannot calculate.
- When it cannot be calculated, this motivates the use of Monte Carlo methods (importance sampling, rejection sampling) or Markov monte Carlo methods (the Metropolis algorithm or Gibbs sampling).

$p(x^{\text{new}} \mid x)$: posterior predictive distribution

- This is used for predicting a new observation based on observed data.

Review of Decision Theory (Module 2)

- Loss functions
- Posterior risk
- Bayes rule
- Frequentist risk
- Integrated risk
- Admissibility

Module 2: https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-2/02-intro-to-Bayes.pdf

Loss function

- Let $\hat{\theta}$ is an estimator of θ (such as the MLE or Bayes estimator).
- A loss function $\ell(\theta, \hat{\theta})$ quantifies how far off the estimator is from the parameter of interest.

Examples: 0-1 loss, squared error loss.

Posterior risk

The posterior risk is defined as

$$\rho(\hat{\theta}, x) = E[\ell(\theta, \hat{\theta} \mid x] \tag{5}$$

Bayes rule

The Bayes rule under **squared error loss** is the posterior mean.

Frequentist risk

The frequentist risk is defined as

$$R(\hat{\theta}, \theta) = E[\ell(\theta, \hat{\theta}) \mid \theta] \tag{6}$$

$$= \int \ell(\theta, \hat{\theta}) p(x \mid \theta) \, dx \tag{7}$$

Integrated risk

$$r(\hat{\theta}) = E[\ell(\theta, \hat{\theta})] \tag{8}$$

$$= \int \int \ell(\theta, \hat{\theta}) p(x \mid \theta) p(\theta) \, dx \, d\theta \tag{9}$$

Admissibility

A decision rule is **admissible** so long as it's not being dominated everywhere.

A decision rule is **inadmissible** is one that is dominated everywhere.

Formally, $\hat{\theta}$ is admissible if there is **no other** $\hat{\theta}'$ such that

$$R(\theta, \hat{\theta}') \leq R(\theta, \hat{\theta}) \quad \text{for all} \quad \theta$$

and

$$R(\theta, \hat{\theta}') \leq R(\theta, \hat{\theta})) \quad \text{for at least one} \quad \theta.$$

Review of Normal distribution (Module 3)

- Review of univariate normal distribution and key properties
- Re-parameterization normal distribution in terms of the precision
- Uniform prior
- Normal-Normal conjugacy

Module 3: https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-3/03-normal-distribution.pdf

Normal distribution

The normal distribution $\mathcal{N}(\mu, \sigma^2)$

- with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ - (standard deviation $\sigma = \sqrt{\sigma^2}$) has p.d.f.

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

for $x \in \mathbb{R}$

Normal distribution (re-parametrization)

It is often more convenient to write the p.d.f. in terms of the **precision**, or inverse variance, $\lambda = 1/\sigma^2$ rather than the variance.

In this parametrization, the p.d.f. is

$$\mathcal{N}(x \mid \mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\tfrac{1}{2}\lambda(x-\mu)^2\right)$$

since $\sigma^2 = 1/\lambda = \lambda^{-1}$.

Uniform prior

We considered $X_{1:n} \mid \theta \overset{iid}{\sim} \mathcal{N}(x \mid \theta, \sigma^2)$,

where $p(\theta) \propto 1$, which is the uniform prior over the real line.

We showed that

$$\theta \mid x_{1:n} \sim \mathcal{N}(\bar{x}, \sigma^2/n).$$

Normal-Normal conjugacy

We considered the following model:

$$X_1, \ldots, X_n \mid \theta \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \lambda^{-1}).$$

Assume the precision $\lambda = 1/\sigma^2$ is known and fixed, and θ is given a $\mathcal{N}(\mu_0, \lambda_0^{-1})$ prior:

$$\theta \sim \mathcal{N}(\mu_0, \lambda_0^{-1})$$

i.e., $p(\theta) = \mathcal{N}(\theta \mid \mu_0, \lambda_0^{-1})$. This is sometimes referred to as a **Normal–Normal** model.

Normal-Normal conjugacy

We derived that

$$p(\theta|x_{1:n}) = \mathcal{N}(\theta \mid M, L^{-1}),$$

where

$$L = \lambda_0 + n\lambda \quad \text{and} \quad M = \frac{\lambda_0\mu_0 + \lambda\sum_{i=1}^{n} x_i}{\lambda_0 + n\lambda}.$$

Review of Normal-Gamma model (Module 4)

- Consider our first hiearchical model with more than two levels

Module 4: https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-4/04-normal-gamma.pdf

Normal-Gamma model

Assume that the likelihood is

$X_1, \ldots, X_n \mid \mu, \lambda \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \lambda^{-1})$ and assume **both**

- the mean μ and
- the precision $\lambda = 1/\sigma^2$ are **unknown**.

Given the likelihood, we place the following priors:

$$\begin{aligned} \boldsymbol{\mu}|\lambda &\sim \mathcal{N}(m, (c\lambda)^{-1}) \\ \boldsymbol{\lambda} &\sim \text{Gamma}(a, b), \end{aligned}$$

where m, c, a, b are known.

Normal-Gamma model

The joint p.d.f. of μ, λ can be written as

$$p(\mu, \lambda) = p(\mu|\lambda)p(\lambda) = \mathcal{N}(\mu \mid m, (c\lambda)^{-1}) \, \text{Gamma}(\lambda \mid a, b)$$

which we denote by

$$\text{NormalGamma}(\mu, \lambda \mid m, c, a, b).$$

Normal-Gamma model

We derived

$$\boldsymbol{\mu}, \boldsymbol{\lambda} | x_{1:n} \sim \text{NormalGamma}(M, C, A, B) \tag{10}$$

i.e., $p(\mu, \lambda | x_{1:n}) = \text{NormalGamma}(\mu, \lambda \mid M, C, A, B)$, where

$$\begin{aligned} M &= \frac{cm + \sum_{i=1}^n x_i}{c + n} \\ C &= c + n \\ A &= a + n/2 \\ B &= b + \tfrac{1}{2}(cm^2 - CM^2 + \textstyle\sum_{i=1}^n x_i^2). \end{aligned}$$

Case study on IQ scores

- We considered a case study on IQ scores (for two groups) and talked about when we might consider the Normal-Gamma model over the Normal-Normal model.
- This case study is a good one to review as it combines many different concepts in the course.

Review of Monte Carlo and MCMC (Modules 5-7)

- Monte Carlo (naive, rejection, and importance sampling)
- Markov Chain Monte Carlo (Gibbs and Metropolis)

Modules 5-7: \url{https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-5/05-monte-carlo.pdf

https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-6/06-metropolis.pdf

https://github.com/resteorts/modern-bayes/tree/master/lecturesModernBayes20/lecture-7

Goal of Monte Carlo and MCMC

Recall Bayes' Theorem:

$$p(\theta \mid x) = \frac{p(\theta, x)}{p(x)} \tag{11}$$

$$= \frac{p(x \mid \theta)p(\theta)}{p(x)} \tag{12}$$

As our models become more complex, we cannot compute $p(x)$, which calls for the use of needing to approximate the posterior distribution.

Monte Carlo

- If were not faced with a high-dimensional problem, then we can resort to using Monte Carlo. This is appealing as it's simple, fast, and easy.
- We illustrated this in the case study of the IQ scores in model 4, where we took some very simple Monte Carlo samples to help solve our case study.

Markov Chain Monte Carlo

- If we are faced with high-dimensional problems, then we need to utilize MCMC (either Gibbs or Metropolis).
- To use Gibbs, we must derive the full conditional distributions. Gibbs is appealing as it's relatively straight forward.
- We typically turn to Metropolis last as it's more complex to implement.

Key concepts

- Understanding when Monte Carlo can be used and when it cannot.
- Being able to work through Gibbs sampling problems. This means deriving the joint likelihood and full conditional distributions.
- Understanding latent variable or data augmentation problems.
- Understanding when you cannot use Gibbs and must use Metropolis (example: logistic regression on o-rings from Module 9).

Multivariate Normal

- Multivariate Notation
- Multivariate Normal and inverse Wishart
- MVN-MVN conjugacy
- MVN-inverseWishart conjugacy
- MVN-MVN-inverseWishart model

Module 8:

\url:{https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-8/08-multivariate-norm.pdf}

https://github.com/resteorts/modern-bayes/blob/master/lecturesModernBayes20/lecture-8/08-missing-data.pdf

Multivariate Notation

Assume that $\boldsymbol{y}_{p\times 1} \sim (\mu_{p\times 1}, \Sigma_{p\times p})$.

$$\boldsymbol{y}_{p\times 1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix}.$$

$$\boldsymbol{\mu}_{p\times 1} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

$$\Sigma_{p\times p} = Cov(\boldsymbol{y}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{pmatrix}.$$

MVN-MVN

Suppose that

$$\boldsymbol{y} = (y_1 \dots y_n)^T \mid \theta \sim MVN(\theta, \Sigma).$$

Let

$$\pi(\boldsymbol{\theta}) \sim MVN(\boldsymbol{\mu}, \Omega).$$

MVN-MVN

We derived that

$$\boldsymbol{\theta} \mid \boldsymbol{y}, \Sigma \sim MVN(A_n^{-1} b_n, A_n^{-1}) = MVN(\mu_n, \Sigma_n),$$

where

$$A_n = A_o + A_1 = \Omega^{-1} + n\Sigma^{-1}$$

and

$$b_n = b_o + b_1 = \Omega^{-1}\mu + \Sigma^{-1} n\bar{y}.$$

MVN-inverseWishart

We then considered the following model:

$$\boldsymbol{y} \mid \boldsymbol{\theta}, \Sigma \sim MVN(\boldsymbol{\theta}, \Sigma).$$
$$\Sigma \sim \text{inverseWishart}(\nu_o, S_o^{-1}),$$

MVN-inverseWishart

We showed that

$$\Sigma \mid \boldsymbol{y}, \boldsymbol{\theta} \sim \text{inverseWishart}(\nu_o + n, [S_o + S_\theta]^{-1} =: S_n),$$

where $S_\theta = \sum_i (\boldsymbol{y}_i - \boldsymbol{\theta})(\boldsymbol{y}_i - \boldsymbol{\theta})^T$.

MVN-MVN-inverseWishart

We then considered the following hierachical model:

$$\boldsymbol{y} \mid \boldsymbol{\theta}, \Sigma \sim MVN(\boldsymbol{\theta}, \Sigma).$$

$$\boldsymbol{\theta} \sim MVN(\mu, \Omega)$$

$$\Sigma \sim \text{inverseWishart}(\nu_o, S_o^{-1}),$$

where we used semi-conjuacy from the previous two derivations, which provide us with full conditional distributions. (Work through this on your own as an exercise for the final exam.)

- Derive the full joint to see why it's difficult to sample from.
- Derive the full conditionals (Hint: use derivations we have done before.)
- Write out the Gibbs sampler (and you can check this as we did this in class).
- What diagnostics would you need to look at to make sure that your sampler has not failed to converge?

Linear Regression

- Setup
- Multivariate Linear Regression
- Multivariate Bayesian Linear Regression

https://github.com/resteorts/modern-bayes/blob/master/lectures ModernBayes20/lecture-9/9-linear-regression.pdf

Setup

Let's assume that we have data points (x_i, y_i) available for all $i = 1, \ldots, n$.

- y is the response variable

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1}$$

- $\boldsymbol{x}_i$ is the ith row of the design matrix $X_{n \times p}$.

Consider the regression coefficients

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{p \times 1}$$

Multivariate Linear Regression

The Normal regression model specifies that

- $E[Y \mid \boldsymbol{x}_i]$ is linear and
- the sampling variability around the mean is independently and identically (iid) drawn from a normal distribution

$$Y_i = \beta^T \boldsymbol{x}_i + \boldsymbol{\epsilon}_i \tag{13}$$

$$\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} \text{Normal}(0, \sigma^2) \tag{14}$$

This implies $Y_i \mid \beta, \boldsymbol{x}_i \sim \text{Normal}(\beta^T \boldsymbol{x}_i, \sigma^2)$.

We can re-write this as

$$\boldsymbol{y} \mid X, \beta, \sigma^2 \sim \text{MVN}(X\beta, \sigma^2 I_p).$$

Multivariate Bayesian Linear Regression

Let

$$\boldsymbol{y} \mid \beta \sim \text{MVN}(X\beta, \sigma^2 I) \tag{15}$$
$$\beta \sim \text{MVN}(\beta_0, \Sigma_0) \tag{16}$$

We derived (exercise 3)

$$\beta \mid \boldsymbol{y}, X \sim \text{MVN}(\beta_n, \Sigma_n), \text{ where}$$

$$\beta_n = E[\beta \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2] = (\Sigma_o^{-1} + (X^T X)/\sigma^2)^{-1}(\Sigma_o^{-1}\beta_0 + \boldsymbol{X}^T \boldsymbol{y}/\sigma^2)$$
$$\Sigma_n = \text{Var}[\beta \mid \boldsymbol{y}, X, \sigma^2] = (\Sigma_o^{-1} + (X^T X)/\sigma^2)^{-1}$$

Remark: If $\Sigma_o^{-1} << (X^T X)^{-1}$ then $\beta_n \approx \hat{\beta}_{ols}$

If $\Sigma_o^{-1} >> (X^T X)^{-1}$ then $\beta_n \approx \beta_0$

The g-prior

To improve our model by doing the **least amount of calculus**, we can put a *g-prior* on β (not β_0).

The g-prior on β has the following form:

$$\beta \mid \boldsymbol{X}, \sigma^2 \sim MVN(0, g\, \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}),$$

where g is a constant, such as $g = n$.

It can be shown that (Zellner, 1986):

1. g shrinks the coefficients and can prevent overfitting to the data
2. if $g = n$, then as n increases, inference approximates that using $\hat{\beta}_{ols}$

The g-prior

Under the g-prior, it follows that

$$\beta_n = E[\beta \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2] \tag{17}$$

$$= \left(\frac{X^T X}{g\sigma^2} + \frac{X^T X}{\sigma^2} \right)^{-1} \frac{X^T y}{\sigma^2} \tag{18}$$

$$= \frac{g}{g+1}(X^T X)^{-1} X^T y = \frac{g}{g+1}\hat{\beta}_{ols} \tag{19}$$

$$\Sigma_n = \mathrm{Var}[\beta \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2] \tag{20}$$

$$= \left(\frac{X^T X}{g\sigma^2} + \frac{X^T X}{\sigma^2} \right)^{-1} = \frac{g}{g+1}\sigma^2 (X^T X)^{-1} \tag{21}$$

$$= \frac{g}{g+1}\mathrm{Var}[\hat{\beta}_{ols}] \tag{22}$$

Logistic Regression

https://github.com/resteorts/modern-bayes/blob/master/lectures ModernBayes20/lecture-9/9-logistic-regression.pdf