Final Exam — STA 360

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

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