STA 602 – Intro to Bayesian Statistics

Lecture 5

Li Ma

General conjugate prior for exponential family models

- Suppose we observe i.i.d. data from an *exponential family*,

$$X_1, X_2, \ldots, X_n \,|\, \eta \sim_{i.i.d.} p(x|\eta) = e^{\eta t(x) - K_0(\eta)} f_0(x).$$

- We place a prior on η of the following form

$$p(\eta) = c e^{a\eta + bK_0(\eta)} = c_0 e^{n_0(\eta t_0 - K_0(\eta))} = \pi_{n_0,t_0}(\eta)$$

  where n_0 and t_0 are two parameters and c_0 is the normalizing constant such that $\int p(\eta|n_0, x_0) d\eta = 1$.
- t_0 is the “prior mean” of each $t(X_i)$ (Diaconis and Ylvisaker 1979)

$$\mathrm{E}t(X_1) = \mathrm{E}\,\mathrm{E}(t(X_1) \,|\, \eta) = \mathrm{E}(K_0'(\eta)) = t_0.$$

- n_0 is the “prior sample size” that quantifies the strength of prior belief.

- Then by Bayes' theorem, the posterior is

$$p(\eta|\mathbf{x}) = c_{\mathbf{x}} e^{n_+(t_+\eta - K_0(\eta))} = \pi_{n_+,t_+}(\eta)$$

where the "posterior sample size" and "posterior mean" of each X_i is

$$n_+ = n_0 + n \quad \text{and} \quad t_+ = \frac{n_0}{n_+} t_0 + \frac{n}{n_+} \bar{t}.$$

- The prior contains information equivalent to n_0 i.i.d. observations with an average of t_0 for the $t(X_i)$.
- Recall that for repeated sampling under an exponential family distribution $\bar{t} = \sum_{i=1}^{n} t(X_i)/n$ is the *sufficient statistic*.

Example: Poisson-Gamma conjugacy

- Is this an exponential family? *Yes!*

$$p(x|\theta) = \theta^x e^{-\theta}/x! = e^{x\log\theta - \theta}/x!.$$

- What is η, $K(\eta)$, t(x), n_0 and t_0?

$$\eta = \log\theta, \qquad K_0(\eta) = e^\eta = \theta, \quad \text{and} \quad t(x) = x.$$

Thus the conjugate prior for this model takes the form

$$p(\eta) = \pi_{n_0,t_0}(\eta) = c_0 e^{n_0(\eta t_0 - K_0(\eta))} = c_0 e^{n_0 t_0 \eta - n_0 K_0(\eta)}.$$

- Apply *change of variable* to get the corresponding prior on $\theta = e^\eta$:

$$p(\theta) = p(\eta)\cdot\left|\frac{d\eta}{d\theta}\right| \propto e^{n_0 t_0 \log\theta - n_0\theta}\cdot 1/\theta = \theta^{n_0 t_0 - 1} e^{-n_0\theta}.$$

- This is exactly the Gamma(α,β) distribution with

$$\alpha = n_0 t_0 \quad \text{and} \quad \beta = n_0.$$

So β is the prior sample size and $\alpha/\beta = t_0$ the prior mean for $t(X_i) = X_i$.

Examples (you can use these as exercises)

- Beta-Binomial.
- Normal-normal.
- Poisson-Gamma.
- Exponential-Gamma and Gamma-Gamma
- …

Posterior predictive checks

- In Bayesian inference, it is often useful to check whether the model (sampling model and prior) characterizes key features of the data well.
- One strategy to do that is to use the predictive distribution for a new data set of the same size as the original one to generate "replicates" of the original data sets under the predictive distribution.
- Compare these replicates with the original data to see whether they are distinct in important ways.

- The predictive distribution distribution of a replicate data set is given by
$$p(\mathbf{x}^{(i)} \,|\, \mathbf{x}) = \int p(\mathbf{x}^{(i)}|\theta)p(\theta|\mathbf{x})d\theta.$$
where $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$.
- So one can generate a replicate data set by drawing a $\theta^{(i)}$ from $p(\theta|\mathbf{x})$, then generate
$$x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)} \sim p(\mathbf{x}|\theta^{(i)}).$$

Bringing in decision theory

- The posterior distribution summarizes all statistical information about the state of nature or parameter θ, *given* the data.
- Decision theory allows us to form notions of “good” and “optimal” in making decisions based on the posterior distribution.

Point estimation

- A very common statistical problem is to "guess" the value of a parameter θ *based on observed data* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$.
- *Functions of the data* that are used for guessing the values of a parameter are called *estimators* for the parameter. Common notations: $\hat{\theta}(\mathbf{X})$, $\delta(\mathbf{X})$, etc. It emphasizes the randomness under repeated experiments.
- If the observed data is $\mathbf{X} = \mathbf{x}$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the realized value of an estimator $\delta(\mathbf{X})$ is $\delta(\mathbf{x})$, which is called an *estimate*.
- In other words, *estimators* are rules that specify how to guess for the parameter based on the data. So they are functions of the data.
- *Estimates* are the specific guesses of the parameter generated after observing the data according to the rules. That is, the are the corresponding functions evaluated at the actual observed data.

How to make "good" estimates/estimators?

- What is a criterion for *good* estimators?
- A good estimator should be such that the estimate and the actual parameter θ are *"likely to be close"*.

What does "likely to be close" mean?

This depends on which view about inference you are taking …

1. The Bayesian view:
   - Both the parameter θ and data $\mathbf{X}$ are random variables.
   - *After* we have observed the data $\mathbf{X} = \mathbf{x}$, only θ is random, and its distribution is the posterior distribution $p(\theta|\mathbf{x})$.
   - In this case, we want to pick an estimate $\delta(\mathbf{x})$ such that *a posteriori* the parameter θ, which is random, will likely take values close to the estimate $\delta(\mathbf{x})$.

Note that here the parameter is random while the estimate $\delta(\mathbf{x})$ is a fixed number given the observed data $\mathbf{X} = \mathbf{x}$.

What does “likely to be close” mean?

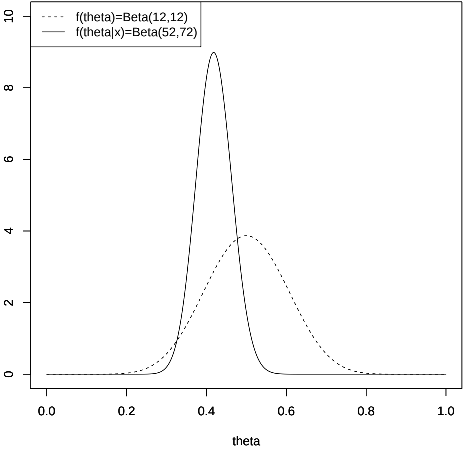
2. The sampling view:
    - The parameter is a *fixed* unknown number. The only random quantities are the data.
    - After data is observed, however, nothing is random. No matter what estimator $\delta(\mathbf{X})$ we are considering, we cannot judge how close the parameter θ is to a *single realization of* the estimate $\delta(\mathbf{x})$ after $\mathbf{X} = \mathbf{x}$ is observed.
    - In this case, we want to choose an estimator $\delta(\mathbf{X})$ that will “with high probability” take values close to the underlying fixed θ.
    - Such a probabilistic statement can only be made *before the experiment*, or under repeated experiment.

Note that this is a “before the experiment” *view, in contrast to the “after the experiment” view taken by the Bayesian perspective.*

The Bayesian estimation problem

- Come back to the political poll example.
- With a Beta(α, β) prior on θ, after observing $X = x$, the posterior distribution of θ is Beta($\alpha + x, \beta + n - x$).
- What would be a good estimate for θ based on this posterior distribution?

Example: Under the $Beta(12, 12)$ prior



- Given $X = 40$, the (posterior) distribution of θ is $\text{Beta}(52, 72)$.
- Which value will you pick as a guess of θ?
- How about the mean, the median, or the mode of the posterior distribution?

- For example, if we choose the mean as the estimate. With $\alpha = 12$ and $\beta = 12$, given $X = 40$, this estimate is

$$\frac{52}{52+72} = \frac{13}{31}.$$

- If we had observed $X = 50$ instead of $X = 40$, then we would have had a different posterior distribution, namely Beta(62,62) distribution.
- The estimate $\delta(50)$ would instead be

$$\frac{62}{62+62} = \frac{1}{2}.$$

Our first estimator based on the posterior distribution

- We choose the estimate depending on the value of the observed data x.
- More generally, for any observed $X = x$, we can estimate θ by

$$E(\theta|x) = \frac{\alpha + x}{\alpha + x + \beta + (n - x)} = \frac{\alpha + x}{\alpha + \beta + n}.$$

- We have just constructed an *estimator*

$$\delta(X) = E(\theta|X) = \frac{\alpha + X}{\alpha + \beta + n}.$$

  This is what the Bayesian would do if the experiment is repeated.
- What is the posterior mode estimate/estimator?

Constructing estimates and estimators by minimizing posterior expected distance

- Can we make a formal rule in building estimators to achieve the *"likely closeness"* between the parameter and the estimate?
- Yes! How about choosing an estimate such that the expected *distance* between θ and the estimate is *as small as possible*.
- In particular, given the posterior distribution $p(\theta|\mathbf{x})$, we can choose an estimate a such that the expected distance between θ and a

$$\mathrm{E}(|\theta - a|\,\big|\,\mathbf{x}) = \int_{-\infty}^{\infty} |\theta - a| p(\theta|\mathbf{x}) d\theta.$$

  is minimized.
- That is, we can define an estimate $\delta^*(\mathbf{x})$ such that

$$\delta^*(\mathbf{x}) = \mathrm{argmin}_a \mathrm{E}(|\theta - a|\,\big|\,\mathbf{x})$$

- Estimates constructed this way are called *Bayes estimates*.

More generally (a decision theory setup)

- Different notions of distance can be adopted. We introduce a distance (or *loss*) function

$$L(\theta, a).$$

- Examples of common *loss* functions include:
  1. $L(\theta, a) = |\theta - a|$ is called the *absolute error loss*.
  2. $L(\theta, a) = (\theta - a)^2$ is called the *squared error loss*.
  3. $L(\theta, a) = \mathbf{1}(|\theta - a| > \Delta)$ is called the *step error loss*.
- The Bayes estimate is the value of a that minimizes the *posterior expectation* of the loss

$$\delta^*(\mathbf{x}) = \operatorname{argmin}_a \mathrm{E}\left(L(\theta, a)|\mathbf{x}\right) = \operatorname{argmin}_a \int_{-\infty}^{\infty} L(\theta, a) p(\theta|\mathbf{x}) d\theta$$

- For example, the Bayes *estimate* under the squared error loss is

$$\delta^*(\mathbf{x}) = \operatorname{argmin}_a \mathrm{E}((\theta - a)^2 \big| \mathbf{x})$$

Loss as the "cost" in decision making

- One can think of the loss function as characterizing the cost of choosing a as the estimate for θ. (Draw a graph.)
  - Consider the situation in which the statistician is making certain *decisions* based on the estimates.
  - The loss function characterizes the cost of choosing a as the estimate for a parameter θ.
  - So one can design custom-made losses for specific problems.
  - Think about the political poll example. What might be a realistic loss function for that?
  - The above simple loss functions are mostly chosen for their mathematical simplicity, especially the squared error loss.

The steps in Bayes estimation (or other decision problems)

1. Choose the distribution of the data given the parameter, $p(\mathbf{x}|\theta)$.
2. Specify a prior distribution for the parameter, $p(\theta)$.
3. After observing the data $\mathbf{X} = \mathbf{x}$, apply Bayes theorem to get the poterior distribution of the parameter, $p(\theta|\mathbf{x})$.
4. Choose a loss function that specifies the distance between the parameter and the estimates.
5. Choose a number a that minimizes the expected distance $\mathrm{E}(L(\theta, a)\big|\mathbf{x})$. This a is our *Bayes estimate* given data $\mathbf{X} = \mathbf{x}$, $\delta^*(\mathbf{x})$.
6. The corresponding *estimator* $\delta^*(\mathbf{X})$ emphasizes the randomness of this decision under repeated experiments, and is called the *Bayes estimator*.

Bayes estimator under squared error loss

It turns out that with *squared error loss*, the Bayes estimate given $\mathbf{X} = \mathbf{x}$ is exactly the posterior mean of θ. That is the mean of the posterior distribution:

$$\delta^*(\mathbf{x}) = \mathrm{E}(\theta|\mathbf{x}),$$

as long as this expecation is well-defined and finite.

The corresponding Bayes estimator is

$$\delta^*(\mathbf{X}) = \mathrm{E}(\theta|\mathbf{X}).$$

Example: Political poll revisited

- Let us go back to our political poll example and find the Bayes estimator for the fraction θ under squared error loss.
- With a Beta(α, β) prior on θ, the Bayes estimator is

$$\delta^*(X) = \mathrm{E}(\theta|X) = \frac{\alpha + X}{\alpha + \beta + n}.$$

- That is, it minimizes the posterior expected squared error loss for any observed data $X = x$.
- Now let's see why the Bayes estimate for squared error loss is the posterior mean.

Bayes estimate under squared error loss

Let Y be a random variable with a finite mean $\mu_Y = E[Y]$. Then for any number a,

$$\begin{aligned} \mathrm{E}(L(Y,a)) &= \mathrm{E}(Y-a)^2 \\ &= \mathrm{E}(Y-\mu_Y+\mu_Y-a)^2 \\ &= \mathrm{E}(Y-\mu_Y)^2 + 2\mathrm{E}(Y-\mu_Y)(\mu_Y-a) + (\mu_Y-a)^2 \\ &= \mathrm{Var}(Y) + (\mu_Y-a)^2. \end{aligned}$$

This is minimized when $a = \mu_Y$.

- Now let the random variable Y be our parameter θ.
- Given $\mathbf{X} = \mathbf{x}$, its distribution is the posterior distribution $p(\theta|\mathbf{x})$.
- Therefore the value a that minimizes $\mathrm{E}(L(\theta,a)|\mathbf{x})$ is $\mathrm{E}(\theta|\mathbf{x})$.

- One can show through more complex arguments that when $L(\theta, a)$ is the absolute error loss, the number a that minimizes $\mathrm{E}(L(\theta, a)|\mathbf{x})$ is the median of posterior distribution $p(\theta|\mathbf{x})$.
- Thus the Bayes estimate

$$\delta^*(\mathbf{x}) = \text{the median of } p(\theta|\mathbf{x}).$$

- The Bayes estimator is

$$\delta^*(\mathbf{X}) = \text{the median of } p(\theta|\mathbf{X}).$$

- For the political poll example, given $X = 40$,
  - The Bayes estimate $\delta^*(40)$ is the median of Beta$(\alpha + 40, \beta + 60)$.
  - The Bayes estimator $\delta^*(X)$ is the median of Beta$(\alpha + X, \beta + n - X)$.

Question: What is the corresponding Bayes estimator for the step error loss?*

$$L(\theta, a) = \begin{cases} 1 & \text{if } |\theta - a| > \Delta \\ 0 & \text{otherwise.} \end{cases}$$

What happens when $\Delta \downarrow 0$?

The air pollutant example with a single reading

- The posterior distribution of θ, given a single measurement $X = x$ is $\mathrm{N}(\mu_1, \tau_1^2)$ with

$$\mathrm{E}(\theta|X = x) = \mu_1 = \left(\frac{1/\tau_0^2}{1/\tau_0^2 + 1/\sigma^2}\right)\mu_0 + \left(\frac{1/\sigma^2}{1/\tau_0^2 + 1/\sigma^2}\right)x.$$

- This is both the mean and the median of the posterior distribution.
- Bayes estimator under squared error loss is

$$\delta(X) = \left(\frac{1/\tau_0^2}{1/\tau_0^2 + 1/\sigma^2}\right)\mu_0 + \left(\frac{1/\sigma^2}{1/\tau_0^2 + 1/\sigma^2}\right)X.$$

- What is the Bayes estimator under absolute error loss?
- How about under the step error loss?

Decision theory under the sampling view

- Because we can no longer take the “after-the-experiment” point of view, evaluating the performances of the corresponding statistical procedure, using loss functions, must be done differently—under the “repeated experiment” or “before-the-experiment” point of view.
- This is useful even for evaluating estimators that arise from a Bayesian perspective.
  - How would the Bayesian do on average under multiple experiments?
- We will see an example next.