STA 360/602L: Module 2.6

Loss functions and Bayes risk

Dr. Olanrewaju Michael Akande

STA 602L

Bayes estimate

- As we've seen by now, having posterior distributions instead of one-number summaries is great for capturing uncertainty.
- That said, it is still very appealing to have simple summaries, especially when dealing with clients or collaborators from other fields, who desire one.
- Can we obtain a single estimate of θ based on the posterior? Sure!
- Bayes estimate is the value $\hat{\theta}$, that minimizes the Bayes risk.

STA 602L

Bayes estimate

- Bayes risk is defined as the expected loss averaged over the posterior distribution.
- Put differently, a Bayes estimate $\hat{\theta}$ has the lowest posterior expected loss.
- That's fine, but what does expected loss mean?
- Frequentist risk also exists but we won't go into that here.

Loss functions

- A loss function $L(\theta, \delta(y))$ is a function of a parameter θ, where $\delta(y)$ is some decision about θ, based on just the data y.
- For example, $\delta(y) = \bar{y}$ can be the decision to use the sample mean to estimate θ, the true population mean.
- $L(\theta, \delta(y))$ determines the penalty for making the decision $\delta(y)$, if θ is the true parameter; $L(\theta, \delta(y))$ characterizes the price paid for errors.

STA 602L

Loss functions

- A common choice for example, when dealing with point estimation, is the squared error loss, which has

$$L(\theta, \delta(y)) = (\theta - \delta(y))^2.$$

- Bayes risk is thus

$$\rho(\theta, \delta) = \mathbb{E}\left[\, L(\theta, \delta(y)) \,|y\right] = \int L(\theta, \delta(y)) \cdot \pi(\theta|y) \, d\theta,$$

and we proceed to find the value $\hat{\theta}$, that is, the decision $\delta(y)$, that minimizes the Bayes risk.

STA 602L

Bayes estimator under squared error loss

- Turns out that, under squared error loss, the decision $\delta(y)$ that minimizes the posterior risk is the posterior mean.
- Proof: Let $L(\theta, \delta(y)) = (\theta - \delta(y))^2$. Then,

$$\begin{aligned}\rho(\theta, \delta) &= \int L(\theta, \delta(y)) \cdot \pi(\theta|y)\, d\theta. \\ &= \int (\theta - \delta(y))^2 \cdot \pi(\theta|y)\, d\theta.\end{aligned}$$

- Expand, then take the partial derivative of $\rho(\theta, \delta)$ with respect to $\delta(y)$.
- To be continued on the board!

STA 602L

Bayes estimator under squared error loss

- $$\rho(\theta, \delta) \int (\theta - \delta(y))^2 \cdot \pi(\theta|y) \, d\theta.$$
- Easy to see then that $\delta(y) = \mathbb{E}[\theta|x]$ is the minimizer.
- Well that's great! The posterior mean is often very easy to calculate in most cases.
- In the beta-binomial case for example, the Bayes estimate under squared error loss is just

$$\hat{\theta} = \frac{a + y}{a + b + n},$$

the posterior mean.

STA 602L

What about other loss functions?

- Clearly, squared error is only one possible loss function. An alternative is absolute loss, which has

$$L(\theta, \delta(y)) = |\theta - \delta(y)|.$$

- Absolute loss places less of a penalty on large deviations & the resulting Bayes estimate is **posterior median**.

- Median is actually relatively easy to estimate.

STA 602L

What about other loss functions?

- Recall that for a continuous random variable Y with cdf F, the median of the distribution is the value z, which satisfies

$$F(z) = \Pr(Y \leq z) = \frac{1}{2} = \Pr(Y \geq z) = 1 - F(z).$$

- As long as we know how to evaluate the CDF of the distribution we have, we can solve for z.
- Think R!

STA 602L

What about other loss functions?

- For the beta-binomial model, the CDF of the beta posterior can be written as

$$F(z) = \Pr(\theta \leq z|y) = \int_0^z \text{Beta}(\theta|a+y, b+n-y)d\theta.$$

- Then, if $\hat{\theta}$ is the median, we have that $F(\hat{\theta}) = 0.5$.

- To solve for $\hat{\theta}$, apply the inverse CDF $\hat{\theta} = F^{-1}(0.5)$.

- In R, that's simply

```
qbeta(0.5,a+y,b+n-y)
```

- For other popular distributions, switch out the beta.

Loss functions and decisions

- Loss functions are not specific to estimation problems but are a critical part of decision making.
- For example, suppose you are deciding how much money to bet ($A) on Duke in the next UNC-Duke men's basketball game.
- Suppose, if Duke
  - loses (y = 0), you lose the amount you bet ($A)
  - wins (y = 1), you gain B per $1 bet
- What is a good sampling distribution for y here?
- Then, the loss function can be characterized as

  $$L(A, y) = A(1 - y) - y(BA),$$

  with your action being the amount bet A.
- When will your bet be "rational"?

STA 602L

How much to bet on Duke?

- y is an unknown state, but we can think of it as a new prediction y_{n+1} given that we have data from win-loss records $(y_{1:n})$ that can be converted into a Bayesian posterior,

  $$\theta \sim \text{beta}(a_n, b_n),$$

  with this posterior concentrated slightly to the left of 0.5, if we only use data on UNC-Duke games (UNC men lead Duke 139-112 all time).
- Actually, it might make more sense to focus on more recent head-to-head data and not the all time record.
- In fact, we might want to build a model that predicts the outcome of the game using historical data & predictors (current team rankings, injuries, etc).
- However, to keep it simple for this illustration, go with the posterior above.

STA 602L

How much to bet on Duke?

- The Bayes risk for action A is then the expectation of the loss function,

$$\rho(A) = \mathbb{E}\left[\, L(A, y) \,\middle|\, y_{1:n} \right].$$

- To calculate this as a function of A and find the optimal A, we need to marginalize over the **posterior predictive distribution** for y.

- Why are we using the posterior predictive distribution here instead of the posterior distribution?

- As an aside, recall from Module 2.3 that

$$p(y_{n+1}|y_{1:n}) = \frac{a_n^{y_{n+1}} b_n^{1-y_{n+1}}}{a_n + b_n}; \quad y_{n+1} = 0, 1.$$

- Specifically, that the posterior predictive distribution here is $\text{Bernoulli}(\hat{\theta})$, with

$$\hat{\theta} = \frac{a_n}{a_n + b_n}$$

- By the way, what do a_n and b_n represent?

STA 602L

How much to bet on Duke?

- With the loss function $L(A, y) = A(1-y) - y(BA)$, and using the notation y_{n+1} instead of y (to make it obvious the game has not been played), the Bayes risk (expected loss) for bet A is

$$\begin{aligned}
\rho(A) &= \mathbb{E}\left[\, L(A, y_{n+1}) \,|y_{1:n}\right] \\
&= \mathbb{E}\left[A(1-y_{n+1}) - y_{n+1}(BA) \,|y_{1:n}\right] \\
&= A\,\mathbb{E}\left[1 - y_{n+1}|\, y_{1:n}\right] - (BA)\,\mathbb{E}\left[y_{n+1}|y_{1:n}\right] \\
&= A\,\left(1 - \mathbb{E}\left[y_{n+1}|y_{1:n}\right]\right) - (BA)\,\mathbb{E}\left[y_{n+1}|y_{1:n}\right] \\
&= A\,\left(1 - \mathbb{E}\left[y_{n+1}|y_{1:n}\right]\,(1+B)\right).
\end{aligned}$$

STA 602L

How much to bet on Duke?

- Hence, your bet is rational as long as

$$\mathbb{E}\left[y_{n+1} \middle| y_{1:n}\right](1+B) > 1$$

$$\frac{a_n(1+B)}{a_n + b_n} > 1.$$

- Clearly, there is no limit to the amount you should bet if this condition is satisfied (the loss function is clearly too simple).
- Loss function needs to be carefully chosen to lead to a good decision - finite resources, diminishing returns, limits on donations, etc.
- Want more on loss functions, expected loss/utility, or decision problems in general? Consider taking a course on decision theory (STA623?).

What's next?

Move on to the readings for the next module!

STA 602L