Module 2: Introduction to Bayesian Statistics, Part II

Rebecca C. Steorts

Exercise

Suppose $X \mid \theta \stackrel{iid}{\sim} Bin(n, \theta)$ and $\theta \mid Beta(a, b)$. Derive the posterior distribution of θ. Now derive the marginal distribution $p(x)$. How does this differ from the Bernoulli-Beta example? Is one a special case of the other?

Agenda

- What is decision theory?
- General setup
- Bayesian approach
- Frequentist and Integrated Risk
- Examples

General setup

Assume an unknown state S (a.k.a. the state of nature). Assume

- we receive an observation x,
- we take an action a, and
- we incur a real-valued loss $\ell(S, a)$.

S	state (unknown)
x	observation (known)
a	action
$\ell(s, a)$	loss

Bayesian approach

- S is a random variable,
- the distribution of x depends on S,
- and the optimal decision is to choose an action a that minimizes the ***posterior expected loss***,

$$\rho(a, x) = \mathbb{E}(\ell(S, a)|x).$$

In other words, $\rho(a, x) = \sum_s \ell(s, a)p(s|x)$ if S is a discrete random variable, while if S is continuous then the sum is replaced by an integral.

Bayesian approach (continued)

1. A ***decision procedure*** δ is a systematic way of choosing actions a based on observations x. Typically, this is a deterministic function $a = \delta(x)$ (but sometimes introducing some randomness into a can be useful).
2. A ***Bayes procedure*** is a decision procedure that chooses an a minimizing the posterior expected loss $\rho(a, x)$, for each x.
3. Note: Sometimes the loss is restricted to be nonnegative, to avoid certain pathologies.

Example 1

1. State: $S = \boldsymbol{\theta}$
2. Observation: $x = x_{1:n}$
3. Action: $a = \hat{\theta}$
4. Loss: $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ (quadratic loss, a.k.a. square loss)

What is the optimal decision rule?

- Goal: Minimize the posterior risk
- First note that

$$\ell(\theta, \hat{\theta}) = \theta^2 - 2\theta\hat{\theta} + \hat{\theta}^2$$

- It then follows that

$$\rho(\hat{\theta}, x_{1:n}) = \mathbb{E}(\ell(\boldsymbol{\theta}, \hat{\theta})|x_{1:n}) = \mathbb{E}(\boldsymbol{\theta}^2|x_{1:n}) - 2\hat{\theta}\mathbb{E}(\boldsymbol{\theta}|x_{1:n}) + \hat{\theta}^2,$$

  which is a convex function of $\hat{\theta}$.

Setting the derivative with respect to $\hat{\theta}$ equal to 0, and solving, we find that the minimum occurs at $\hat{\theta} = \mathbb{E}(\boldsymbol{\theta}|x_{1:n})$, **the posterior mean**.

Resource allocation for disease prediction

Suppose public health officials in a small city need to decide how much resources to devote toward prevention and treatment of a certain disease, but the fraction θ of infected individuals in the city is unknown.

Resource allocation for disease prediction (continued)

Suppose they allocate enough resources to accomodate a fraction c of the population.

- If c is too large, there will be wasted resources, while if it is too small, preventable cases may occur and some individuals may go untreated.
- After deliberation, they tentatively adopt the following loss function:

$$\ell(\theta, c) = \begin{cases} |\theta - c| & \text{if } c \geq \theta \\ 10|\theta - c| & \text{if } c < \theta. \end{cases}$$

Resource allocation for disease prediction (continued)

- By considering data from other similar cities, they determine a prior $p(\theta)$. For simplicity, suppose $\boldsymbol{\theta} \sim \text{Beta}(a, b)$ (i.e., $p(\theta) = \text{Beta}(\theta|a, b)$), with $a = 0.05$ and $b = 1$.
- They conduct a survey assessing the disease status of $n = 30$ individuals, $x_1, \ldots, x_n$.

This is modeled as $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, which is reasonable if the individuals are uniformly sampled and the population is large. Suppose all but one are disease-free, i.e., $\sum_{i=1}^{n} x_i = 1$.

The Bayes procedure

The Bayes procedure is to minimize the posterior expected loss

$$\rho(c, x) = \mathbb{E}(\ell(\boldsymbol{\theta}, c)|x) = \int \ell(\theta, c) p(\theta|x) d\theta$$

where $x = x_{1:n}$.

1. We know $p(\theta|x)$ as an updated Beta, so we can numerically compute this integral for each c.
2. Figure 1 shows $\rho(c, x)$ for our example.
3. The minimum occurs at $c \approx 0.08$, so under the assumptions above, this is the optimal amount of resources to allocate.
4. How would one perform a sensitivity analysis of the prior assumptions?

Posterior expected loss for disesase prevelance

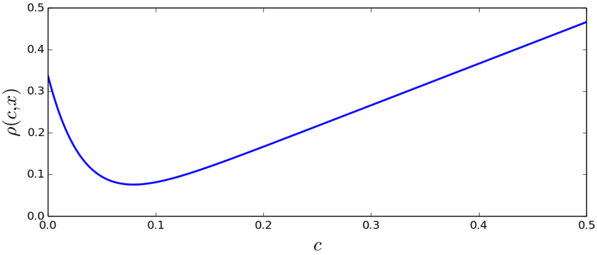

Figure 1: Posterior expected loss for the disease prevalence example. Think about the shape of $\ell(\theta, c)$ as a function of c, for some fixed θ. Imagine how it changes as θ goes from 0 to 1, and think about taking a weighted average of these functions, with weights determined by $p(\theta|x)$.

Frequentist and Integrated Risk

1. Consider a decision problem in which $S = \boldsymbol{\theta}$.
2. The ***risk*** (or ***frequentist risk***) associated with a decision procedure δ is

$$R(\theta, \delta) = \mathbb{E}(\ell(\boldsymbol{\theta}, \delta(X)) \mid \boldsymbol{\theta} = \theta),$$

   where X has distribution $p(x|\boldsymbol{\theta})$. In other words,

$$R(\theta, \delta) = \int \ell(\theta, \delta(x))\, p(x|\theta)\, dx$$

   if X is continuous, while the integral is replaced with a sum if X is discrete.
3. The ***integrated risk*** associated with δ is

$$r(\delta) = \mathbb{E}(\ell(\boldsymbol{\theta}, \delta(X)) = \int R(\theta, \delta)\, p(\theta)\, d\theta.$$

Example: Resource allocation, revisited

1. The frequentist risk provides a useful way to compare decision procedures in a prior-free way.
2. In addition to the Bayes procedure above, consider two other possibilities: choosing $c = \bar{x}$ (sample mean) or $c = 0.1$ (constant).

Example: Resource allocation, revisited

3. Figure 2 shows each procedure as a function of $\sum x_i$, the observed number of diseased cases. For the prior we have chosen, the Bayes procedure always picks c to be a little bigger than $\bar{x}$.

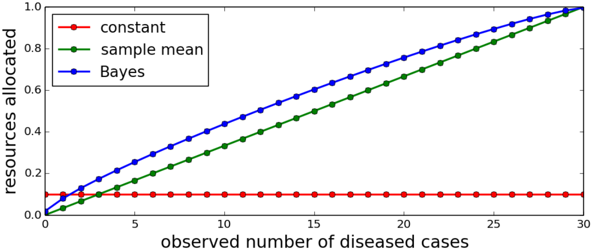

Figure 2: Resources allocated c, as a function of the number of diseased individuals observed, $\sum x_i$, for the three different procedures.

Example: Resource allocation, revisited

4. Figure 3 shows the risk $R(\theta, \delta)$ as a function of θ for each procedure. Smaller risk is better. (Recall that for each θ, the risk is the expected loss, averaging over all possible data sets. The observed data doesn't factor into it at all.)

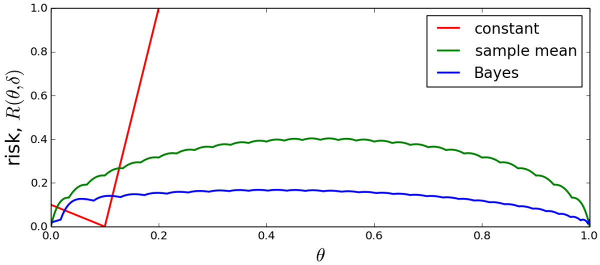

Figure 3: Risk functions for the three different procedures.

Example: Resource allocation, revisited

5. The constant procedure is fantastic when θ is near 0.1, but gets very bad very quickly for larger θ. The Bayes procedure is better than the sample mean for nearly all θ's. These curves reflect the usual situation—some procedures will work better for certain θ's and some will work better for others.
6. A decision procedure is called ***admissible*** if there is no other procedure that is at least as good for all θ and strictly better for some. That is, δ is admissible if there is no δ' such that

$$R(\theta, \delta') \leq R(\theta, \delta)$$

   for all θ and $R(\theta, \delta') < R(\theta, \delta)$ for at least one θ.
7. Bayes procedures are admissible under very general conditions.
8. Admissibility is nice to have, but it doesn't mean a procedure is necessarily good. Silly procedures can still be admissible—e.g., in this example, the constant procedure $c = 0.1$ is admissible too!