Point estimation

Suppose we are interested in the value of a parameter θ, for example the unknown bias of a coin. We have already seen how one may use the Bayesian method to reason about θ; namely, we select a likelihood function $p(\mathcal{D} \mid \theta)$, explaining how observed data $\mathcal{D}$ are expected to be generated given the value of θ. Then we select a prior distribution $p(\theta)$ reflecting our initial beliefs about θ. Finally, we conduct an experiment to gather data and use Bayes' theorem to derive the posterior $p(\theta \mid \mathcal{D})$.

In a sense, the posterior contains all information about θ that we care about. However, the process of inference will often require us to use this posterior to answer various questions. For example, we might be compelled to choose a single value $\hat{\theta}$ to serve as a *point estimate* of θ. To a Bayesian, the selection of $\hat{\theta}$ is a *decision,* and in different contexts we might want to select different values to report.

In general, we should not expect to be able to select the true value of θ, unless we have somehow observed data that unambiguously determine it. Instead, we can only hope to select an estimate that is "close" to the true value. Different definitions of "closeness" can naturally lead to different estimates. The Bayesian approach to point estimation will be to analyze the impact of our choice in terms of a *loss function,* which describes how "bad" different types of mistakes can be. We then select the estimate which appears to be the least "bad" according to our current beliefs about θ.

Decision theory

As mentioned above, the selection of an estimate $\hat{\theta}$ can be seen as a decision. It turns out that the Bayesian approach to decision theory is rather simple and consistent, so we will introduce it in an abstract form here.

A decision problem will typically have three components. First, we have a *parameter space* (also called a *state space*) Θ, with an unknown value $\theta \in \Theta$. We will also have a *sample space* $\mathcal{X}$ representing the potential observations we could theoretically make. Finally, we have an *action space* $\mathcal{A}$ representing the potential actions we may select from. In the point estimation problem, the potential actions $\hat{\theta}$ are exactly those in the parameter space, so $\mathcal{A} = \Theta$. This might not always be the case, however. Finally, we will have a likelihood function $p(\mathcal{D} \mid \theta)$ linking potential observations to the parameter space.

After conducting an experiment and observing data, we are compelled to select an action $a \in \mathcal{A}$. We define a (deterministic) *decision rule* as a function $\delta \colon \mathcal{X} \to \mathcal{A}$ that selects an action a given the observations $\mathcal{D}$.[1] In general, this decision rule can be any arbitrary function. How do we select which decision rule to use?

To guide our selection, we will define a *loss function,* which is a function $L \colon \Theta \times \mathcal{A} \to \mathbb{R}$. The value $L(\theta, a)$ summarizes "how bad" an action a was if the true value of the parameter was revealed to be θ; larger losses represent worse outcomes. Ideally, we would select the action that minimizes this loss, but unfortunately we will never know the exact value of θ; complicating our decision.

As usual, there are two main approaches to designing decision rules. We begin with the Bayesian approach.

[1] We may also consider *randomized* decision rules, where δ maps observed data $\mathcal{D}$ to a probability distribution over $\mathcal{A}$, which we select a sample from. This can be useful, for example, when facing an intelligent adversary. Most of the expressions we will derive can be derived analogously for this case; however, we will not do so here.

Bayesian decision theory

The Bayesian approach to decision theory is straightforward. Given our observed data $\mathcal{D}$, we find the posterior $p(\theta \mid \mathcal{D})$, which represents our current belief about the unknown parameter θ. Given a potential action a, we may define the *posterior expected loss* of a by averaging the loss function over the unknown parameter:

$$\rho\big(p(\theta \mid \mathcal{D}), a\big) = \mathbb{E}\big[L(\theta, a) \mid \mathcal{D}\big] = \int_\Theta L(\theta, a) p(\theta \mid \mathcal{D}) \, \mathrm{d}\theta.$$

Because the posterior expected loss of each action a is a scalar value, it defines a total order on the action space $\mathcal{A}$. When there is an action minimizing the posterior expected loss, it is the natural choice to make:

$$\delta^*(\mathcal{D}) = \arg\min_{a \in \mathcal{A}} \rho\big(p(\theta \mid \mathcal{D}), a\big),$$

representing the action with the lowest expected loss, given our current beliefs about θ. Note that $\delta^*(\mathcal{D})$ may not be unique, in which case we can select any action attaining the minimal value. Any minimizer of the posterior expected loss is called a *Bayes action.* The value $\delta^*(\mathcal{D})$ may also be found by solving the equivalent minimization problem

$$\delta^*(\mathcal{D}) = \arg\min_{a \in \mathcal{A}} \int_\Theta L(\theta, a) p(\mathcal{D} \mid \theta) p(\theta) \, \mathrm{d}\theta;$$

the advantage of this formulation is that it avoids computing the normalization constant $p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta) p(\theta) \, \mathrm{d}\theta$ in the posterior.

Notice that a Bayes action is tied to a particular set of observed data $\mathcal{D}$. This does not limit its utility terribly; after all, we will always have a particular set of observed data at hand when making a decision. However we may extend the notion of Bayes actions in a natural way to define an entire decision rule. We define the *Bayes rule,* a decision rule, by simply always selecting a (not necessarily unique) Bayes action given the observed data. Note that the second formulation above can often be minimized even when $p(\theta)$ is not necessarily a probability distribution (such priors are called *improper* but are often encountered in practice). A decision rule derived in this way from an improper prior is called a *generalized Bayes rule.*

In the case of point estimation, the decision rule δ may be more naturally written $\hat{\theta}(\mathcal{D})$. Then, as above, the point estimation problem reduces to selecting a loss function and deriving the decision rule $\hat{\theta}$ that minimizes the expected loss at every point. A decision rule that minimizes posterior expected loss for every possible set of observations $\mathcal{D}$ is called a *Bayes estimator.*

We may derive Bayes estimators for some common loss functions. As an example, consider the common loss function

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

which represents the squared distance between our estimate and the true value. For the squared loss, we may compute:

$$\begin{aligned}
\mathbb{E}\big[L(\theta, \hat{\theta}) \mid \mathcal{D}\big] &= \int (\theta - \hat{\theta})^2 p(\theta \mid \mathcal{D}) \, \mathrm{d}\theta \\
&= \int \theta^2 p(\theta \mid \mathcal{D}) \, \mathrm{d}\theta - 2\hat{\theta} \int \theta p(\theta \mid \mathcal{D}) \, \mathrm{d}\theta + \hat{\theta}^2 \int p(\theta \mid \mathcal{D}) \, \mathrm{d}\theta \\
&= \int \theta^2 p(\theta \mid \mathcal{D}) \, \mathrm{d}\theta - 2\hat{\theta}\mathbb{E}[\theta \mid \mathcal{D}] + \hat{\theta}^2.
\end{aligned}$$

We may minimize this expression by differentiating with respect to $\hat{\theta}$ and equating to zero:

$$\frac{\partial \mathbb{E}\big[L(\theta, \hat{\theta}) \mid \mathcal{D}\big]}{\partial \hat{\theta}} = -2\mathbb{E}[\theta \mid \mathcal{D}] + 2\hat{\theta} = 0,$$

from which we may derive $\hat{\theta} = \mathbb{E}[\theta \mid \mathcal{D}]$. Examining the second derivative, we see

$$\frac{\partial^2 \mathbb{E}\big[L(\theta, \hat{\theta}) \mid \mathcal{D}\big]}{\partial \hat{\theta}^2} = 2 > 0,$$

so this is indeed a minimum. Therefore we have shown that the Bayes estimator in the case of squared loss is the posterior mean $\hat{\theta}(\mathcal{D}) = \mathbb{E}[\theta \mid \mathcal{D}]$.

A similar analysis shows that the Bayes estimator for the absolute deviation loss $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ is the posterior median, and the Bayes estimators for a relaxed 0–1 loss:

$$L(\theta, \hat{\theta}; \varepsilon) = \begin{cases} 0 & |\theta - \hat{\theta}| < \varepsilon; \\ 1 & |\theta - \hat{\theta}| \geq \varepsilon, \end{cases}$$

converge to the posterior mode for small ε.

The posterior mode, also called the *maximum a posteriori* (MAP) estimate of θ and written $\hat{\theta}_{\text{MAP}}$, is a rather common estimator used in practice. The reason is that optimization is almost always easier than integration. In particular, we may find the MAP estimate by maximizing the *unnormalized* posterior

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\mathcal{D} \mid \theta) p(\theta),$$

where we have avoided computing the normalization constant $p(\mathcal{D}) = \int p(\mathcal{D} \mid \theta) p(\theta) \, \mathrm{d}\theta$. An important caveat is that the MAP estimator is not invariant to nonlinear transformations of θ!

Frequentist decision theory

The frequentist approach to decision theory is somewhat different. As usual, in classical statistics it is not allowed to place a prior distribution on a parameter such as θ; rather, it is much more common to use the likelihood $p(\mathcal{D} \mid \theta)$ to hypothesize what data might look like for different values of θ, and use this analysis to drive your action.

The frequentist approach to decision theory involves the notion of *risk functions*. The *frequentist risk* of a decision function δ is defined by

$$R(\theta, \delta) = \int_{\mathcal{X}} L\big(\theta, \delta(\mathcal{D})\big) p(\mathcal{D} \mid \theta) \, \mathrm{d}\mathcal{D},$$

that is, it represents the expected loss incurred when repeatedly using the decision rule δ on different datasets $\mathcal{D}$ as a function of the unknown parameter θ.

To a Bayesian, the frequentist risk is a very strange notion: we know the exact value of our data $\mathcal{D}$ when we make our decision, so why should we average over other datasets that we haven't seen? The frequentist counterargument to this is typically that we might know $\mathcal{D}$ but can't know $p(\theta)$!

Notice that whereas the posterior expected loss was a scalar defining a total order on the action space $\mathcal{A}$, which could be extended to naturally define an entire decision rule, the frequentist risk is a function of θ and the entire decision rule δ. It is very unusual for there to be a single decision rule δ that works the best for every potential value of θ. For this reason, we must decide on some mechanism to use the frequentist risk to select a "good" decision rule. There are many proposed mechanisms for doing so, but we will simply quickly describe two below.

Bayes risk

One solution to the problem of comparing decision rules is to place a prior distribution on the unknown θ and compute the average risk:

$$r\big(p(\theta), \delta\big) = \mathbb{E}\big[R(\theta, \delta)\big] = \int_\Theta R(\theta, \delta) p(\theta) \, \mathrm{d}\theta = \int_\Theta \int_\mathcal{X} L\big(\theta, \delta(\mathcal{D})\big) p(\mathcal{D} \mid \theta) p(\theta) \, \mathrm{d}\theta \, \mathrm{d}\mathcal{D}.$$

The function $r\big(p(\theta), \delta\big)$ is called the *Bayes risk* of δ under the prior $p(\theta)$. Again, the Bayes risk is scalar-valued, so we induce a total order on all decision rules, making identifying a unique decision rule easier. Any δ minimizing Bayes risk is called a *Bayes rule.* We have seen this term before! It turns out that given a prior $p(\theta)$, the Bayesian procedure described above for defining a decision function by selecting an action with minimum posterior expected loss is guaranteed to minimize Bayes risk and therefore produce a Bayes rule with respect to $p(\theta)$. Note, however, that it is unusual in the Bayesian perspective to first find an entire decision rule δ and then apply it to a particular dataset $\mathcal{D}$. Instead, it is almost always easier to minimize the expected posterior loss only at the actual observed data. After all, why would we need to know what decision we would make with other data?

Admissibility

Another criterion for selecting between decision rules in the frequentist framework is called *admissibility.* In short, it is often difficult to identify a single best decision rule, but it can sometimes be easy to discard some bad ones, for example if they can be shown to always be no better than (and sometimes worse than) another rule.

Let δ_1 and δ_2 be two decision rules. We say that δ_1 *dominates* δ_2 if:

- $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all $\theta \in \Theta$, and
- there exists at least one θ for which $R(\theta, \delta_1) < R(\theta, \delta_2)$.

If there is a decision rule δ that is not dominated by any other rule, it is called *admissible.* One interesting result tying Bayesian and frequentist decision theory is the following:

- Every Bayes rule is admissible.
- Every admissible decision rule is a generalized Bayes rule for some (possibly improper) prior $p(\theta)$.

So, in a sense, all admissible frequentist decision rules can be equivalently derived from a Bayesian perspective.

Examples

Here we give two quick examples of applying Bayesian decision theory.

Classification with 0–1 loss

Suppose our observations are of the form (x, y), where x is an arbitrary input, and $y \in \{0, 1\}$ is a binary label associated with x. In classification, our goal is to predict the label y' associated with a new input x'. The Bayesian approach is to derive a model giving probabilities $\Pr(y' = 1 \mid x', \mathcal{D})$.

Suppose this model is provided for you. Notice that this model is not conditioned on any additional parameters θ; we have integrated them out via

$$\Pr(y' = 1 \mid x', \mathcal{D}) = \int \Pr(y' = 1 \mid x', \mathcal{D}, \theta) p(\theta \mid \mathcal{D}) \, \mathrm{d}\theta.$$

Given a new datapoint x', which label a should we predict? Notice that the prediction of a label is actually a *decision*. Here our action space is $\mathcal{A} = \{0, 1\}$, enumerating the two labels we can predict. Our parameter space is the same: the only uncertainty we have is the unknown label y'.

Let us suppose a simple loss function for this problem:

$$L(y', a) = \begin{cases} 0 & a = y'; \\ 1 & a \neq y'. \end{cases}$$

This loss function, called the *0–1 loss,* is common in classification problems. We pay a constant loss for every mistake we make. In this case, the expected loss of each possible action is simple to compute:

$$\begin{aligned} \mathbb{E}\big[L(y', a = 1) \mid x', \mathcal{D}\big] &= \Pr(y' = 0 \mid x', \mathcal{D}); \\ \mathbb{E}\big[L(y', a = 0) \mid x', \mathcal{D}\big] &= \Pr(y' = 1 \mid x', \mathcal{D}). \end{aligned}$$

The Bayes action is then to predict the class with the highest probability. This is not so surprising. Notice that if we change the loss to have different costs of mistakes (so that $L(0, 1) \neq L(1, 0)$), then the Bayes action might compel us to select the less-likely class to avoid a potentially high loss for misclassification!

Hypothesis testing

The Bayesian approach to hypothesis testing is also rather straightforward. A *hypothesis* is simply a subset of the parameter space $\mathcal{H} \subseteq \Theta$. The Bayesian approach allows one to compute the posterior probability of a hypothesis directly:

$$\Pr(\theta \in \mathcal{H} \mid \mathcal{D}) = \int_{\mathcal{H}} p(\theta \mid \mathcal{D}) \, \mathrm{d}\theta.$$

Now, equipped with a loss function, the posterior expected loss framework above can be applied directly to select between different hypotheses.

See last week's notes for a brief discussion of frequentist hypothesis testing.