STAT 460/560 Class 10: Statistical Decision Theory

Ben Bloem-Reddy

Reading: Chapter 12, [Was04].

At a high level, statistical decision theory is a framework for assessing (and comparing the performance of statistical procedures). It connects statistical tasks like parameter estimation to real-world/practical quantities such as cost, risk, etc.

As usual, $\theta \in \Theta$ is a parameter, and $\hat{\theta}$ is an estimator. In decision theoretic terms, $\hat{\theta}$ is a **decision rule** (i.e., a function that maps data to a decision, or **action**). Note that we are not limited to parameter estimation. As an example more closely related to the language, a decision rule could actually be a data-driven decision on which action to take (e.g., which experiment to run next).

For simplicity, we will look for now only at parameter estimation. The discrepancy of $\hat{\theta}$ and θ is measured by a **loss function**, $L\colon \Theta \times \Theta \to \mathbb{R}$. A common example is **squared error loss**,

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2 \; . \tag{10.1}$$

The **risk** of an estimator $\hat{\theta}$ is a function of θ,

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta[L(\theta, \hat{\theta}(X))] = \int L(\theta, \hat{\theta}(x)) f(x; \theta) \; dx \; . \tag{10.2}$$

When the loss is squared error, the resulting risk is just the MSE,

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta}(X) - \theta)^2] = \mathrm{Var}_\theta(\hat{\theta}(X)) + \mathsf{bias}_\theta(\hat{\theta}(X))^2 \; . \tag{10.3}$$

1. Comparing risk functions

How do we compare two functions? Let's start with some examples.

Example 10.1. Let $X \sim \mathcal{N}(\theta, 1)$, and consider two estimators under squared error loss: $\hat{\theta}_1 = X$ and $\hat{\theta}_2 = 3$. The risk functions are $R(\theta, \hat{\theta}_1) = 1$ and $R(\theta, \hat{\theta}_2) = (3 - \theta)^2$. If $2 < \theta < 4$ then $\hat{\theta}_2$ has lower risk; otherwise $\hat{\theta}_1$ has lower risk. So neither estimator is uniformly better than the other.

Activity 10.1. Let $X_1, \ldots, X_n \sim_{\text{IID}} \mathrm{Bern}(p)$, and consider squared error loss. Let $\hat{p}_1 = \bar{X}$. Compute the risk. (To make things easy, recall that this estimator is unbiased.)

Now suppose that we take a Bayesian approach, and use a $\mathrm{Beta}(\alpha, \beta)$ prior, with $\hat{p}_2 = \mathbb{E}[p | X^n]$. Write down $\hat{p}_2$ and calculate its risk for $\alpha = \beta = \sqrt{n}/2$.

Plot or sketch the risk functions of the two estimators. Is either one uniformly better? Do you notice anything interesting about $\hat{p}_2$?

In both of these examples, if we don't know the true value of θ (which we don't) then how should we choose between the two estimators? Much of statistical decision theory deals with this seemingly simple question.

There are two widely used ways to turn the risk function into a number. The **maximum risk** is

$$\bar{R}(\hat{\theta}) = \sup_{\theta} R(\theta, \hat{\theta}) \ . \tag{10.4}$$

The **Bayes risk** is

$$r(f, \hat{\theta}) = \int R(\theta, \hat{\theta}) f(\theta) \ d\theta \ , \tag{10.5}$$

where $f(\theta)$ is a prior PDF/PMF.

Activity 10.2. Find the maximum risk for the two estimators from Activity 10.1. Which is lower?

Find the Bayes risk for the two estimators under a uniform prior.

Exercise 10.1. Find the Bayes risk for the two estimators under a Beta($\sqrt{n}/2, \sqrt{n}/2$) prior. Show that the ratio of Bayes risks can be written as

$$\frac{r(\text{Beta}(\sqrt{n}/2, \sqrt{n}/2), \hat{p}_1)}{r(\text{Beta}(\sqrt{n}/2, \sqrt{n}/2), \hat{p}_2)} = 1 + \frac{1}{\sqrt{n}} \ . \tag{10.6}$$

Note that the Beta($\sqrt{n}/2, \sqrt{n}/2$) prior isn't likely to be motivated by any prior belief or knowledge,[1] except in the case $n = 1$. We'll see later where the prior comes from.

Based on these one-number summaries of risk, we can devise general strategies for finding "good" estimators: we might try to find $\hat{\theta}$ to minimize the maximum risk; alternatively, we might find $\hat{\theta}$ to minimize the Bayes risk. A **minimax estimator** $\hat{\theta}$ is any estimator (it may not be unique) such that

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta}) \ , \tag{10.7}$$

where the infimum is taken over all possible estimators $\tilde{\theta}$. The **Bayes estimator** with respect to prior f is any estimator $\hat{\theta}$ such that

$$r(f, \hat{\theta}) = \inf_{\tilde{\theta}} r(f, \tilde{\theta}) \ . \tag{10.8}$$

Bayes estimators are often easier to find. To see why, define the **posterior risk** as

$$r(\hat{\theta}|x) = \int L(\theta, \hat{\theta}(x)) f(\theta|x) \ d\theta \ . \tag{10.9}$$

Theorem 10.1. *The Bayes risk satisfies*

$$r(f, \hat{\theta}) = \int r(\hat{\theta}|x) m(x) \ dx \ , \tag{10.10}$$

where $m(x) = \int f(x|\theta) f(\theta) \ d\theta$. *Therefore, under squared error loss, the Bayes estimator is the posterior expectation,*

$$\hat{\theta}(x) = \int \theta f(\theta|x) \ d\theta = \mathbb{E}[\theta | X = x] \ . \tag{10.11}$$

[1] Why would our prior belief/knowledge about p depend on n?

Exercise 10.2. Prove Theorem 10.1 .

Exercise 10.3. For $X_1, \ldots, X_n \sim_{\text{IID}} \text{Bern}(p)$ and a $\text{Beta}(\alpha, \beta)$ prior, what is the Bayes estimator under squared error loss? Under $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$?

2. Minimax decision rules

Finding minimax decision rules is in general hard, and there is no one-size-fits-all method for doing so. In some situations, a Bayes estimator can be minimax.

Theorem 10.2. *Let $\hat{\theta}^f$ be a Bayes estimator for some prior f, and therefore*

$$r(f, \hat{\theta}^f) = \inf_{\hat{\theta}} r(f, \hat{\theta}) \ . \tag{10.12}$$

Suppose that

$$R(\theta, \hat{\theta}^f) \leq r(f, \hat{\theta}^f) \quad \textit{for all } \theta \in \Theta \ . \tag{10.13}$$

Then $\hat{\theta}^f$ is minimax and f is called a ***least favorable prior****.*

Proof. Suppose that $\hat{\theta}^f$ is not minimax. Then there is some estimator $\hat{\theta}_0$ such that

$$\sup_\theta R(\theta, \hat{\theta}_0) < \sup_\theta R(\theta, \hat{\theta}^f) \ . \tag{10.14}$$

Since the average of a function is always less than or equal to its maximum, $r(f, \hat{\theta}_0) \leq \sup_\theta R(\theta, \hat{\theta}_0)$. Then

$$r(f, \hat{\theta}_0) \leq \sup_\theta R(\theta, \hat{\theta}_0) < \sup_\theta R(\theta, \hat{\theta}^f) \leq r(f, \hat{\theta}^f) \ . \tag{10.15}$$

This is a contradiction, because $\hat{\theta}^f$ is the Bayes estimator under f. Hence, $\hat{\theta}^f$ is minimax. □

Corollary 10.3. *Suppose that $\hat{\theta}$ is a Bayes estimator under some prior f, and further that $\hat{\theta}$ has constant risk c. Then $\hat{\theta}$ is minimax.*

Exercise 10.4. Prove Corollary 10.3.

Activity 10.3. Recall that in the previous class, with $X_1, \ldots, X_n \sim_{\text{IID}} \text{Bern}(p)$ and squared error loss, we analyzed $\hat{p}_2 = \mathbb{E}[p|X^n]$ when the prior is $\text{Beta}(\sqrt{n}/2, \sqrt{n}/2)$. Argue that $\hat{p}_2$ is minimax.

3. Optional: Prediction and empirical risk minimization

When the task is parameter estimation, risk is typically a purely theoretical construct: the risk usually depends on an unknown parameter and therefore cannot be estimated from data. However, consider the problem of predicting Y from X with an estimator $\hat{Y} = \hat{\theta}(X)$. The estimator will be fit to pairs $((X_1, Y_1), \ldots, (X_n, Y_n)) \sim_{\text{IID}} P$, and the risk is

$$R(P, \hat{\theta}) = \mathbb{E}_P[L(Y, \hat{\theta}(X))] = \int L(y, \hat{\theta}(x)) dP(x, y) \ . \tag{10.16}$$

This depends on the unknown distribution P. However, it can be estimated by the **empirical risk**

$$\hat{R}(\hat{\theta}) = R(\hat{P}_n, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \hat{\theta}(X_i)) \ . \tag{10.17}$$

Minimizing this type of objective is called **empirical risk minimization**. This is the default procedure for most machine learning methods (though its implementation for different models and massive datasets is fairly complicated).

Activity 10.4. Suppose we have $\mathbb{R}$-valued pairs $((X_1, Y_1), \ldots, (X_n, Y_n)) \sim_{\text{IID}} P$, and we're using squared error loss. Show that an estimator that achieves the smallest risk is

$$\hat{\theta}(x) = \mathbb{E}_P[Y|X = x] \ . \tag{10.18}$$

Exercise 10.5. Suppose that $Y \in \{0, 1\}$, $((X_1, Y_1), \ldots, (X_n, Y_n)) \sim_{\text{IID}} P$, and we're using 0-1 loss,

$$L(Y, \hat{\theta}(X)) = \mathbf{I}\{Y \neq \hat{\theta}(X)\} \ . \tag{10.19}$$

Show that an estimator that achieves the smallest risk is

$$\hat{\theta}(x) = \begin{cases} 1 & \text{if } P(Y = 1|X = x) \geq \frac{1}{2} \\ 0 & \text{o.w.} \end{cases} \tag{10.20}$$

This is called the **Bayes classifier**.

For more on the statistical aspects of machine learning, see Hastie, Tibshirani, and Friedman [HTF09], which is now somewhat outdated but still a great place to start. Wainwright [Wai19] has more advanced material.

References

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer New York, 2009. URL: https://hastie.su.domains/ElemStatLearn/.

[Wai19] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[Was04] L. Wasserman. *All of Statistics*. Springer New York, 2004.