Decision Theory

Statistical decision theory is concerned with making decisions under uncertainty. We express our uncertainties around the problem in terms of an unknown quantity or "state of nature" θ.

The particular decision made is also referred to as an "action," and we'll denote it by a, with the collection of all possible actions denoted by $\mathcal{A}$.

A loss function

$$L(\theta, a) : (\Theta \times \mathcal{A}) \rightarrow [0, \infty)$$

describes the consequences of taking action a when the true state of nature is θ. In reality, we never know the true value of the loss (at least not at the time of the decision).

Example: A drug company has developed a new pain reliever. They are trying to determine how much of the drug to produce, but they are uncertain about the proportion of the market the drug will capture (θ).

Suppose a is an estimate of θ. The company plans to produce an amount proportional to a. One possible loss function is

$$L(\theta, a) = \begin{cases} K(\theta - a) & a - \theta < 0 \\ 2K(a - \theta) & a - \theta \geq 0 \end{cases}$$

for some constant K. This loss function implies that an overestimate of demand (leading to overproduction of the drug) is considered twice as costly as an underestimate. The loss is also taken to be linear, which may be reasonable if the total cost is proportional to the number of units produced.

Many results are based on the following "standard" loss functions. These are expressed in generic "units of utility."

- Squared error loss: $L(\theta, a) = (\theta - a)^2$

- Linear loss: $L(\theta, a) = \begin{cases} K_1(\theta - a) & a - \theta < 0 \\ K_2(a - \theta) & a - \theta \geq 0 \end{cases}$

- Absolute error loss: $L(\theta, a) = |\theta - a|$ (linear loss with $K_1 = K_2$)

- L^p loss: $L(\theta, a) = |\theta - a|^p$

- Zero-one loss: $L(\theta, a) = \begin{cases} 0 & a = \theta \\ 1 & a \neq \theta \end{cases}$

Since we don't know the actual loss, we may consider an "expected loss" and then choose an "optimal" decision with respect to this. This "expected loss" is known as risk. However, there are several ways of thinking about the expectation; hence, several different risks. Note: in what follows, we'll consider estimation problems only, that is, actions $a = \hat{\theta}(x_1, ..., x_n)$. **In the following, when there is no ambiguity and for simplicity of notation, we denote $(x_1, ...x_n)$ by x.**

1. The posterior risk

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

averages over uncertainty in θ after conditioning on observations x. We may think of it as a function of x, as well as the particular form of $\hat{\theta}$. Another way to think of this is that, conditional on the observations x, we just get a single number of risk for each estimator $\hat{\theta}$ we might consider.

2. The frequentist risk (or sometimes just “risk”)

$$R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}(x)) f(x|\theta) dx$$

averages over different possible realizations x of the random variable, given that the true “state of nature” is θ. It is a function of θ, as well as the particular form of $\hat{\theta}$.

Consider two estimators, $\hat{\theta}$ and $\hat{\theta}'$. We say $\hat{\theta}'$ dominates $\hat{\theta}$ if

$$\begin{aligned} R(\theta, \hat{\theta}') &\leq R(\theta, \hat{\theta}) \text{ for all } \theta \\ R(\theta, \hat{\theta}') &< R(\theta, \hat{\theta}) \text{ for at least one } \theta \end{aligned}$$

The estimator $\hat{\theta}$ is called inadmissible if there is at least one other estimator $\hat{\theta}'$ that dominates it. Otherwise it is called admissible.

Example: Suppose $X \sim N(\theta, 1)$ and we are estimating θ under squared error loss. Consider $\hat{\theta}_c(x) = cx$. Here, X denotes a random variable, and x represents a single observed value (i.e., an observed sample of size 1).

- Calculate the risk in terms of c and θ.

- Calculate the risk when $c = 1$.

- Show that $\hat{\theta}_c$ is inadmissible when $c > 1$.

- Make a plot comparing the risk when $c = 1/2$ and $c = 1$.

3. The Bayes risk:

$$\begin{aligned} r(f,\hat{\theta}) &= \int R(\theta,\hat{\theta}) f(\theta) d\theta \\ &= \int \left[\int L(\theta,\hat{\theta}(x)) f(x|\theta) dx \right] f(\theta) d\theta \\ &= \int \int L(\theta,\hat{\theta}(x)) f(x,\theta) dx \, d\theta \\ &= \int \left[\int L(\theta,\hat{\theta}(x)) f(\theta|x) d\theta \right] f(x) dx \\ &= \int r(\hat{\theta}|x) f(x) dx \end{aligned}$$

averages over both θ and X. It depends on the particular form of $\hat{\theta}$, and the probability model for the data $(f(x|\theta))$ and the parameter θ $(f(\theta))$.

A decision rule that minimizes the Bayes risk is called a Bayes rule. The estimator $\hat{\theta}$ is a Bayes rule, or Bayes estimator (under a particular model and loss function) if

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

We can also find the Bayes estimator using the posterior risk. For each observed sample x (i.e., $x = (x_1, ..., x_n)$), let $\hat{\theta}(x)$ be the value of $\hat{\theta}$ that minimizes $r(\hat{\theta}|x)$. (Recall that for each x, $r(\hat{\theta}|x)$ returns a single number for each $\hat{\theta}$.) The estimator defined in this way is the Bayes estimator. This is because

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

and we have defined this $\hat{\theta}$ to minimize the quantity being integrated for each x; hence we've also minimized the whole integral.

Example (continued): Suppose $X \sim N(\theta, 1)$ and we are estimating θ under squared error loss. Consider $\hat{\theta}_c(x) = cx$. Here, X denotes a random variable, and x represents a single observed value (i.e., an observed sample of size 1).

- Calculate the Bayes risk using the prior $\theta \sim N(0, \tau^2)$.

- Find the Bayes rule among estimators $\hat{\theta}_c$.

- Find the Bayes risk of this estimator.

We can calculate the Bayes rule explicitly for several standard loss functions.

- Squared error loss: posterior mean
- Absolute error loss: posterior median
- Zero-one loss: posterior mode

Recall the different risk functions:

1. Posterior risk (depends on x and the form of $\hat{\theta}$)

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

2. Frequentist risk (depends on θ and the form of $\hat{\theta}$)

$$R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}(x)) f(x|\theta) dx$$

3. Bayes risk (depends on the form of $\hat{\theta}$)

$$r(f, \hat{\theta}) = \int \int L(\theta, \hat{\theta}(x)) f(x, \theta) dx\, d\theta$$

Completely equivalently, we could write

1. Posterior risk: $r(\hat{\theta}|x) = E_{\theta|X}[L(\theta, \hat{\theta}(x))]$

2. Frequentist risk: $R(\theta, \hat{\theta}) = E_{X|\theta}[L(\theta, \hat{\theta}(X))]$

3. Bayes risk: $r(f, \hat{\theta}) = E_{\theta,X}[L(\theta, \hat{\theta}(X))]$

By iterated expectation, we also have that

$$r(f, \hat{\theta}) = E_{\theta}[E_{X|\theta}[L(\theta, \hat{\theta}(X))]] = E_{\theta}[R(\theta, \hat{\theta})]$$

and

$$r(f, \hat{\theta}) = E_X[E_{\theta|X}[L(\theta, \hat{\theta}(X))]] = E_X[r(\hat{\theta}|X)]$$

Example: Suppose $X_1, \ldots, X_n | \sigma^2 \overset{iid}{\sim} N(\theta, \sigma^2)$, where θ is known. Let the prior distribution for σ^2 be inverse gamma with parameters a and b. The prior PDF is

$$f(\sigma^2; a, b) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} \exp\{-b/\sigma^2\}$$

- Find the posterior distribution for σ^2.
- What is the Bayes estimator under squared error loss?
- What is the Bayes estimator under absolute error loss?
- What is the Bayes estimator under zero-one loss?

You may use the fact the mean of an $InverseGamma(a, b)$ distribution is $b/(a-1)$ when $a > 1$, the mode is $b/(a+1)$, and the median is not available in closed form.

One final note about Bayes rules: under weak conditions, they are admissible. The intuition for this is that if there existed a rule that had lower risk, it would also have lower Bayes risk.

Here is one set of conditions:

Suppose that $\Theta \subseteq \mathbb{R}$ and that $R(\theta, \hat{\theta})$ is a continuous function of θ for every $\hat{\theta}$. Let f be a prior density that assigns positive probability to any open subset of Θ. Let $\hat{\theta}^f$ be a Bayes rule, with finite Bayes risk. Then $\hat{\theta}^f$ is admissible.

We'll now consider a different strategy for choosing an action, called a **minimax rule**. To motivate this, consider the following example.

An investor is deciding whether or not to purchase $\$1000$ of risky ZZZ bonds. If the investor buys the bonds, they can be redeemed at maturity for a net gain of $\$500$. There could, however, be a default on the bonds, in which case the original $\$1000$ investment would be lost. If the investor doesn't buy the bonds, she will put her money in a "safe" investment, for which she will be guaranteed a net gain of $\$300$ over the same time period. She estimates the probability of a default to be 0.1.

- Describe the parameter space Θ and the space of possible actions $\mathcal{A}$.
- What is the prior distribution?
- For each possible $\theta \in \Theta$ and $a \in \mathcal{A}$, compute the loss.
- Is any action inadmissible?

In the previous example, the Bayes rule is for the investor to buy the bonds, since this minimizes her expected loss (maximizes her expected gain) relative to the prior distribution for a default occurring.

However, suppose the investor is very conservative, and wants to choose a strategy to minimize the "worst case scenario." This is known as the minimax strategy – it minimizes the maximum loss that could occur.

Writing the frequentist risk of action a as $R(\theta, a)$, the maximum risk

$$\bar{R}(a) = \sup_{\theta} R(\theta, a)$$

Which action in the example minimizes $\bar{R}(a)$?

In the estimation context, our possible actions are estimators $\hat{\theta}$. Then the maximum risk is

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

A decision rule that minimizes the maximum frequentist risk is called a minimax rule. The estimator $\hat{\theta}$ is a minimax rule (under a particular loss function) if

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

Example (continued): Suppose $X \sim N(\theta, 1)$ and we are estimating θ under squared error loss. Consider $\hat{\theta}_c(x) = cx$.

- Calculate $\sup_{\theta} R(\theta, \hat{\theta}_c)$.
- Use this to determine the minimax estimator of θ.
- Let the prior distribution for θ be $N(a, b)$. Determine the Bayes estimator of θ.

Geometry of Bayes and Minimax Points for Finite Ω

Given a finite parameter space $\Omega = \{\theta_1, \cdots, \theta_k\}$, we define the risk set as $S \subseteq \mathbb{R}^k$ such that

$$S = \{(y_1, \cdots, y_k) : y_i = R(\theta_i, \delta) \text{ for } \delta \in \mathcal{A}\}$$

Lemma. The risk set S is always convex when $\mathcal{A}$ has randomized estimators.

In this setting, a prior of θ can be considered as a finite vector

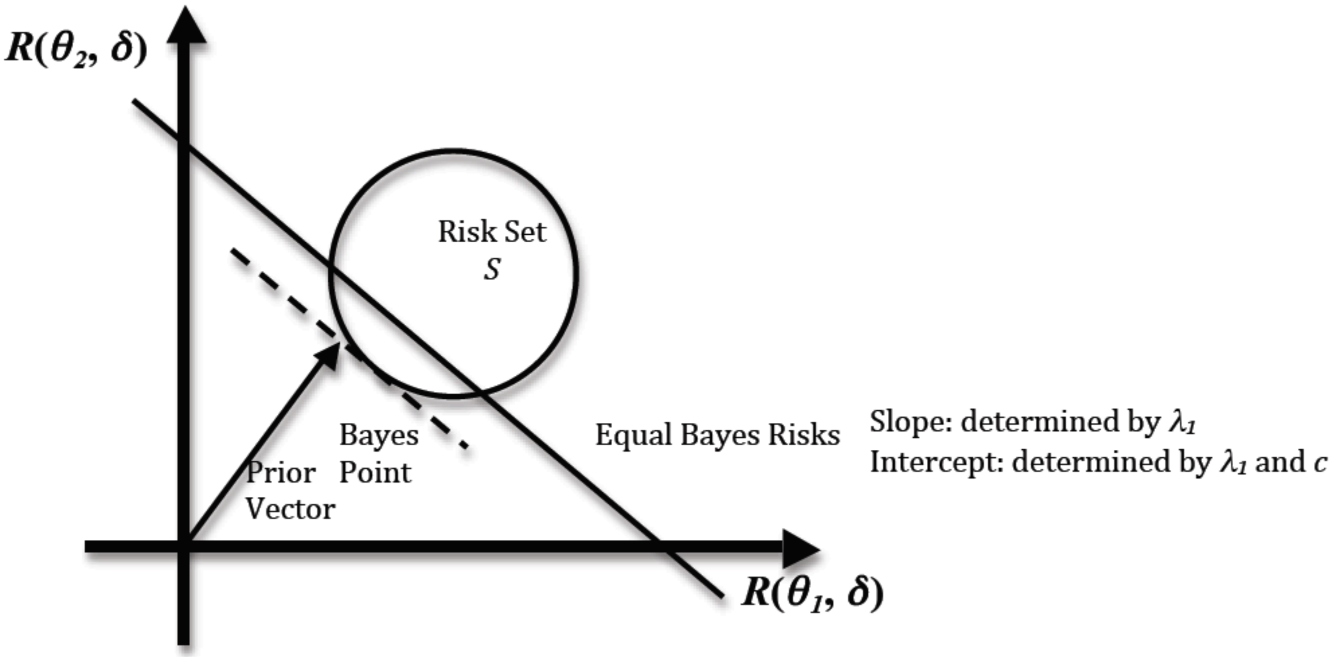
$$\lambda(\theta) = (\lambda_1, \cdots, \lambda_k) = (\lambda(\theta_1), \cdots, \lambda(\theta_k)),$$

with $\sum_{i=1}^{k} \lambda_i = 1$ and $\lambda \geqslant 0$. The Bayes risk is

$$r(\Lambda, \delta) = \sum_{i=1}^{k} \lambda_i R(\theta_i, \delta) = (\lambda_1, \cdots, \lambda_k) \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix}$$

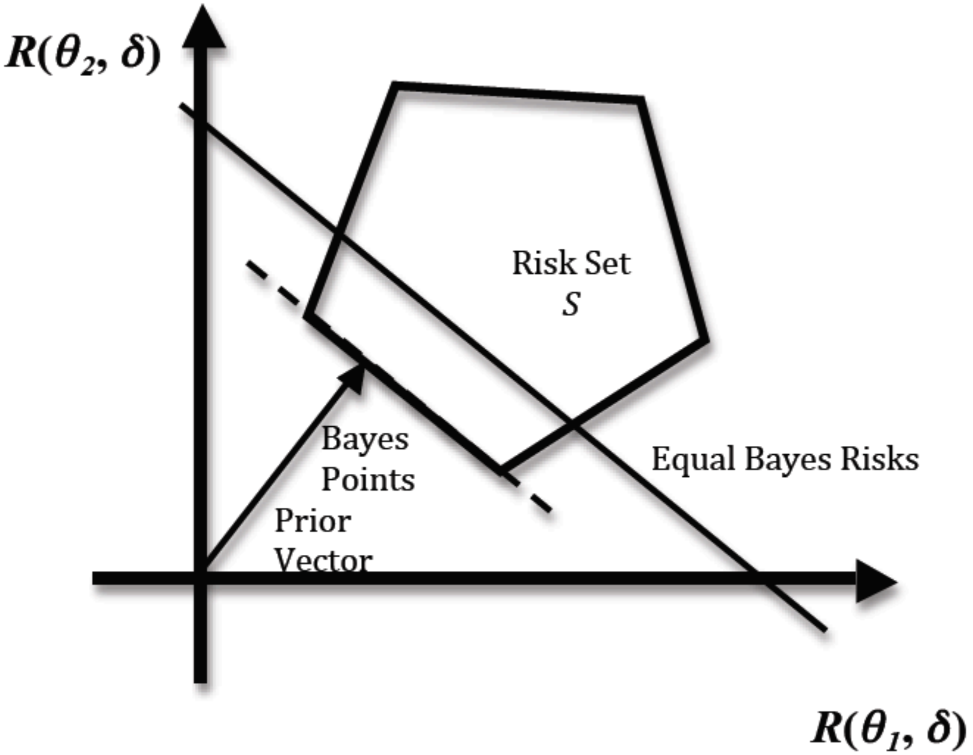
Any given prior vector $(\lambda_1, \cdots, \lambda_k)$ is normal to hyperplanes of constant Bayes risks in $\mathbb{R}^k$:

$$\text{Hyperplane:}\{(y_1, \cdots, y_k) : (\lambda_1, \cdots, \lambda_k) \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix} = c \text{ for } \delta \in \mathcal{A}\}$$



Geometry of a Bayes Point for $k = 2$.

For any given Λ, the Bayes points should be on the hyperplane that is tangent to the risk set (ie. gives the smallest c). But Bayes points may not be unique.

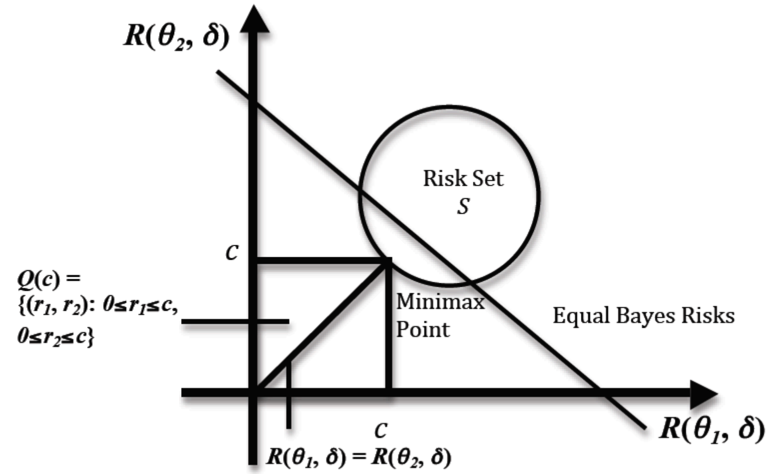


Where is the minimax point?

Where is the minimax point?

$$\sup_{\theta_1,\theta_2}(\theta, \delta^*) = \inf_{\delta \in \mathcal{A}} \sup_{\theta_1,\theta_2} R(\theta, \delta)$$

Consider the points on the vertical and horizontal segments: the points corresponding to $\sup_{\theta\in\Omega} R(\theta,\delta) = c$. The points which give the smallest c in that risk set are the minimax points. That should be the first points of contact between the squares $Q(c)$ and the risk set S.


$R(\theta_2, \delta)$
Risk Set
S
c
$Q(c) = \{(r_1, r_2): 0 \le r_1 \le c, 0 \le r_2 \le c\}$
Minimax Point
Equal Bayes Risks
c
$R(\theta_1, \delta)$
$R(\theta_1, \delta) = R(\theta_2, \delta)$

In general it can be difficult to find minimax rules when the parameter space is infinite. One connection to Bayes rules is the following:

Suppose that $\hat{\theta}$ is the Bayes rule with respect to some prior f. Suppose further that $\hat{\theta}$ has constant risk: $R(\theta, \hat{\theta}) = c$ for some c. Then $\hat{\theta}$ is minimax.

Proof: Let $R(\theta, \delta)$ denote the risk of a decision rule δ at parameter value θ. We need to prove that, for all δ,

$$\sup_{\theta} R(\theta, \hat{\theta}) \leq \inf_{\delta} \sup_{\theta} R(\theta, \delta).$$

Note that the Bayes risk at $\hat{\theta}$ is $r(f, \hat{\theta}) = \inf_{\delta} r(f, \delta)$, where $r(f, \delta) = E_{\theta \sim f}[R(\theta, \delta)]$. Therefore we have

$$\sup_{\theta} R(\theta, \hat{\theta}) = c = E_{\theta \sim f}[R(\theta, \hat{\theta})] = r(f, \hat{\theta}) \leq E_{\theta \sim f}[R(\theta, \delta)] \leq \sup_{\theta} R(\theta, \delta). \quad \square$$

Example: Suppose $X|p \sim Bin(n,p)$ and the loss is squared error.

- Show $\hat{p} = X/n$ is not minimax. *Hint: Consider the randomized estimator*

$$\tilde{p} = \left\{ \begin{array}{ll} X/n & \text{with probability } 1 - \frac{1}{n+1} \\ 1/2 & \text{with probability } \frac{1}{n+1} \end{array} \right\}$$

- Consider the Bayes estimator when $p \sim Beta(a,b)$. Find a and b so that the Bayes estimator has constant frequentist risk. This estimator is then minimax.

Summary: We are considering taking possible actions $a \in \mathcal{A}$, and unknown quantities affecting our decision are represented by $\theta \in \Theta$.

In the estimation context, the action is just an estimate of θ, $\hat{\theta}(x)$.

The loss function describes the consequences of taking action a when the true state of nature is θ. We write it $L(\theta, a)$ or $L(\theta, \hat{\theta}(x))$.

Ultimately, we want to *choose* an action a or an estimate $\hat{\theta}(x)$. Our choice is driven by looking at a particular *risk function.*

So far we have seen two strategy:
(1) the Bayes rule, which chooses $\hat{\theta}(x)$ to minimize the Bayes risk; (2) the minimax rule, which minimizes the maximum frequentist risk.

Recall that for an estimation problem in a parametric framework, we also learned MOM MLE.

Example: Consider a decision problem with possible states of nature θ_1 and θ_2 . Let X be a random variable with probability function $p(x|\theta)$:
$P(X=0|\theta_1)=0.2, P(X=1|\theta_1)=0.8;$
$P(X=0|\theta_2)=0.4, P(X=1|\theta_2)=0.6.$

Two non-randomized actions a_1 and a_2 are considered with the following loss function:
$L(\theta_1,a_1(0))=1, L(\theta_1,a_1(1))=2, L(\theta_1,a_2(0))=4, L(\theta_1,a_2(1))=0;$
$L(\theta_2,a_1(0))=3, L(\theta_2,a_1(1))=1, L(\theta_2,a_2(0))=1, L(\theta_2,a_2(1))=4.$

1. Give and plot the risk set $S=\{(r_1,r_2): r_1=\lambda R(\theta_1,a_1)+(1-\lambda)R(\theta_1,a_2), r_2=\lambda R(\theta_2,a_1)+(1-\lambda)R(\theta_2,a_2), \lambda\in[0,1]\}$.

2. Suppose θ has the prior distribution $\Lambda(\theta)$ defined by $P(\theta=\theta_1)=0.9, P(\theta=\theta_2)=0.1$. What is the Bayes rule with respect to $\Lambda(\theta)$?

3. Find the minimax rule(s).