Foundations of machine learning

Statistical decision theory

Maximilian Kasy

Department of Economics, University of Oxford

Winter 2024

Outline

- Basic definitions
- Optimality criteria
- Relationships between optimality criteria
- Analogies to microeconomics
- Two justifications of the Bayesian approach

Takeaways for this part of class

1. A general framework to think about what makes a “good” estimator, test, etc.
2. How the foundations of statistics relate to those of microeconomic theory.
3. In what sense the set of Bayesian estimators contains most “reasonable” estimators.

Examples of decision problems

- Decide whether or not the hypothesis of no racial discrimination in job interviews is true
- Provide a forecast of the unemployment rate next month
- Provide an estimate of the returns to schooling
- Pick a portfolio of assets to invest in
- Decide whether to reduce class sizes for poor students
- Recommend a level for the top income tax rate

Basic definitions

Optimality criteria

Some relationships between these optimality criteria

Analogies to microeconomics

Two justifications of the Bayesian approach

References

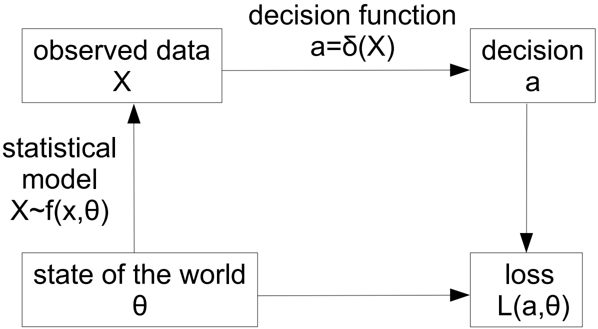
Components of a general statistical decision problem

- Observed data X
- A statistical decision a
- A state of the world θ
- A loss function $L(a, \theta)$ (the negative of utility)
- A statistical model $f(X|\theta)$
- A decision function $a = \delta(X)$

How they relate

- underlying state of the world θ
  $\Rightarrow$ distribution of the observation X.
- decision maker: observes $X \Rightarrow$ picks a decision a
- her goal: pick a decision that minimizes loss $L(a, \theta)$
  (θ unknown state of the world)
- X is useful $\Leftrightarrow$ reveals some information about θ
  $\Leftrightarrow f(X|\theta)$ does depend on θ.
- problem of statistical decision theory:
  find decision functions δ which "make loss small."

Graphical illustration



Examples

- investing in a portfolio of assets:
  - X: past asset prices
  - a: amount of each asset to hold
  - θ: joint distribution of past and future asset prices
  - L: minus expected utility of future income
- decide whether or not to reduce class size:
  - X: data from project STAR experiment
  - a: class size
  - θ: distribution of student outcomes for different class sizes
  - L: average of suitably scaled student outcomes, net of cost

Practice problem

For each of the examples on slide 2, what are

- the data X,
- the possible actions a,
- the relevant states of the world θ, and
- reasonable choices of loss function L?

Loss functions in estimation

- goal: find an a
- which is close to some function μ of θ.
- for instance: $\mu(\theta) = E[X]$
- loss is larger if the difference between our estimate and the true value is larger

Some possible loss functions:

1. **squared error** loss,
$$L(a, \theta) = (a - \mu(\theta))^2$$
2. **absolute error** loss,
$$L(a, \theta) = |a - \mu(\theta)|$$

Loss functions in testing

- goal: decide whether $H_0 : \theta \in \Theta_0$ is true
- decision $a \in \{0,1\}$ (accept / reject)

Possible loss function:

$$L(a,\theta) = \begin{cases} 1 & \text{if } a = 1,\ \theta \in \Theta_0 \\ c & \text{if } a = 0,\ \theta \notin \Theta_0 \\ 0 & \text{else.} \end{cases}$$

	truth	
decision a	$\theta \in \Theta_0$	$\theta \notin \Theta_0$
0	0	c
1	1	0

Risk function

$$R(\delta,\theta) = E_\theta[L(\delta(X),\theta)].$$

- expected loss of a decision function δ
- R is a function of the true state of the world θ.
- crucial intermediate object in evaluating a decision function
- small $R \Leftrightarrow$ good δ
- δ might be good for some θ, bad for other θ.
- Decision theory deals with this trade-off.

Example: estimation of mean

- observe $X \sim N(\mu, 1)$
- want to estimate μ
- $L(a, \theta) = (a - \mu(\theta))^2$
- $\delta(X) = \alpha + \beta \cdot X$

Practice problem (Estimation of means)

Find the risk function for this decision problem.

Variance / Bias trade-off

Solution:

$$\begin{aligned} R(\delta,\mu) &= E[(\delta(X)-\mu)^2] \\ &= \mathrm{Var}(\delta(X)) + \mathrm{Bias}(\delta(X))^2 \\ &= \beta^2\,\mathrm{Var}(X) + (\alpha+\beta E[X]-E[X])^2 \\ &= \beta^2 + (\alpha+(\beta-1)\mu)^2. \end{aligned}$$

- equality 1 and 2: always true for squared error loss
- Choosing β (and α) involves a trade-off of bias and variance,
- this trade-off depends on μ.

Basic definitions

Optimality criteria

Some relationships between these optimality criteria

Analogies to microeconomics

Two justifications of the Bayesian approach

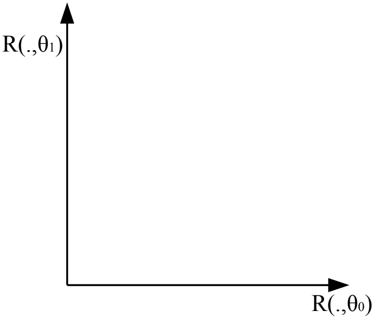
References

Optimality criteria

- Ranking provided by the risk function is multidimensional:
- a ranking of performance between decision functions for every θ
- To get a global comparison of their performance, have to aggregate this ranking into a global ranking.
- preference relationship on space of risk functions
  $\Rightarrow$ preference relationship on space of decision functions

Illustrations for intuition

- Suppose θ can only take two values,
- $\Rightarrow$ risk functions are points in a 2D-graph,
- each axis corresponds to $R(\delta, \theta)$ for $\theta = \theta_0, \theta_1$.



Three approaches to get a global ranking

1. **partial ordering**:
   a decision function is better relative to another
   if it is better for *every* θ

2. complete ordering, **weighted average**:
   a decision function is better relative to another
   if a weighted average of risk across θ is lower
   weights $\sim$ prior distribution

3. complete ordering, **worst case**:
   a decision function is better relative to another
   if it is better under its worst-case scenario.

Approach 1: Admissibility

Dominance:
δ is said to dominate another function δ' if

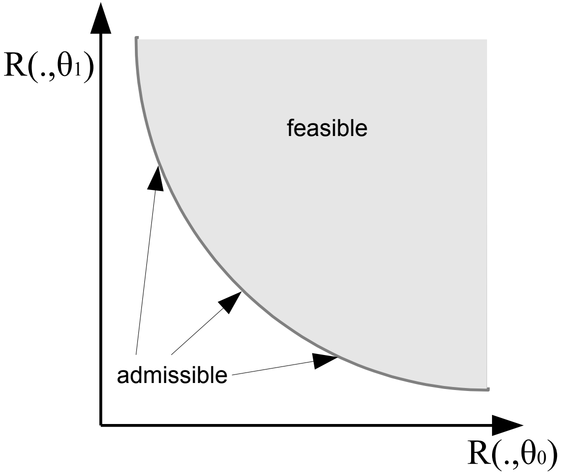
$$R(\delta, \theta) \leq R(\delta', \theta)$$

for all θ, and

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

for at least one θ.

Admissibility:
decisions functions which are not dominated are called admissible,
all other decision functions are inadmissible.


R(.,θ1)
feasible
admissible
R(.,θ0)

- admissibility $\sim$ "Pareto frontier"
- Dominance only generates a partial ordering of decision functions.
- in general: many different admissible decision functions.

Practice problem

- you observe $X_i \sim^{iid} N(\mu, 1)$, $i = 1, \ldots, n$ for $n > 1$
- your goal is to estimate μ, with squared error loss
- consider the estimators
  1. $\delta(X) = X_1$
  2. $\delta(X) = \frac{1}{n} \sum_i X_i$
- can you show that one of them is inadmissible?

Approach 2: Bayes optimality

- natural approach for economists:
- trade off risk across different θ
- by assigning weights $\pi(\theta)$ to each θ

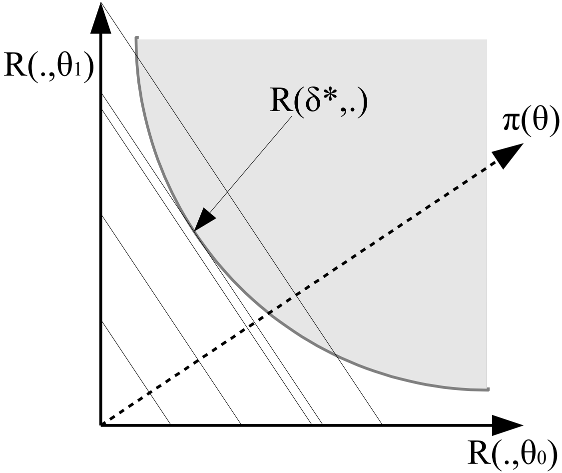
Integrated risk:

$$R(\delta, \pi) = \int R(\delta, \theta)\pi(\theta)d\theta.$$

Bayes decision function:
minimizes integrated risk,

$$\delta^* = \underset{\delta}{\operatorname{argmin}}\ R(\delta, \pi).$$

- Integrated risk $\sim$ linear indifference planes in space of risk functions
- prior $\sim$ normal vector for indifference planes


R(.,θ1)
R(δ*,.)
π(θ)
R(.,θ0)

Decision weights as prior probabilities

- suppose $0 < \int \pi(\theta) d\theta < \infty$
- then wlog $\int \pi(\theta) d\theta = 1$ (normalize)
- if additionally $\pi \geq 0$
- then π is called a prior distribution

Posterior

- suppose π is a prior distribution
- **posterior distribution:**

$$\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{m(X)}$$

- normalizing constant $=$ prior likelihood of X

$$m(X) = \int f(X|\theta)\pi(\theta)d\theta$$

Practice problem

- you observe $X \sim N(\theta, 1)$
- consider the prior
$$\theta \sim N(0, \tau^2)$$
- calculate
  1. $m(X)$
  2. $\pi(\theta|X)$

Posterior expected loss

$$R(\delta, \pi|X) := \int L(\delta(X), \theta)\pi(\theta|X)d\theta$$

Proposition

Any Bayes decision function δ^*
can be obtained by minimizing $R(\delta, \pi|X)$
through choice of $\delta(X)$ for every X.

Practice problem

Show that this is true.

Hint: show first that

$$R(\delta, \pi) = \int R(\delta(X), \pi|X)m(X)dX.$$

Bayes estimator with quadratic loss

- assume quadratic loss, $L(a, \theta) = (a - \mu(\theta))^2$
- posterior expected loss:

$$\begin{aligned} R(\delta, \pi | X) &= E_{\theta|X}\left[L(\delta(X), \theta)|X\right] \\ &= E_{\theta|X}\left[(\delta(X) - \mu(\theta))^2 | X\right] \\ &= \text{Var}(\mu(\theta)|X) + (\delta(X) - E[\mu(\theta)|X])^2 \end{aligned}$$

- Bayes estimator minimizes posterior expected loss $\Rightarrow$

$$\delta^*(X) = E[\mu(\theta)|X].$$

Practice problem

- you observe $X \sim N(\theta, 1)$
- your goal is to estimate θ, with squared error loss
- consider the prior

$$\theta \sim N(0, \tau^2)$$

- for any δ, calculate
  1. $R(\delta(X), \pi | X)$
  2. $R(\delta, \pi)$
  3. the Bayes optimal estimator δ^*

Practice problem

- you observe X_i iid., $X_i \in \{1,2,\dots,k\}$, $P(X_i = j) = \theta_j$
- consider the so called Dirichlet prior, for $\alpha_j > 0$:

$$\pi(\theta) = const. \cdot \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$$

- calculate $\pi(\theta|X)$
- look up the Dirichlet distribution on Wikipedia
- calculate $E[\theta|X]$

Approach 3: Minimaxity

- Don't want to pick a prior?
- Can instead always assume the worst.
- worst $=\theta$ which maximizes risk

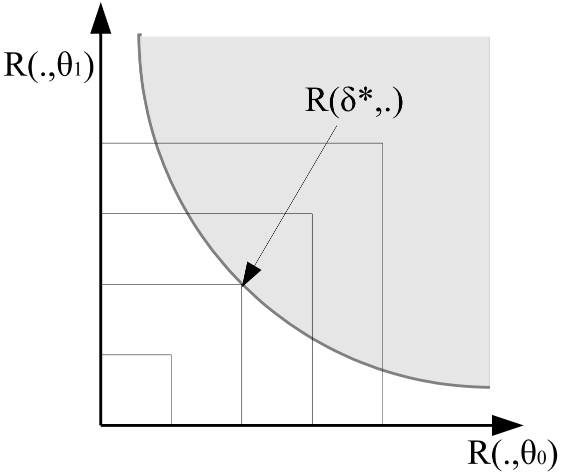
worst-case risk:

$$\overline{R}(\delta)=\sup_{\theta} R(\delta,\theta).$$

minimax decision function:

$$\delta^{*}=\underset{\delta}{\operatorname{argmin}}\ \overline{R}(\delta)=\underset{\delta}{\operatorname{argmin}}\ \sup_{\theta} R(\delta,\theta).$$

(does not always exist!)


R(.,θ1)
R(δ*,.)
R(.,θ0)

Basic definitions

Optimality criteria

Some relationships between these optimality criteria

Analogies to microeconomics

Two justifications of the Bayesian approach

References

Some relationships between these optimality criteria

Proposition (Minimax decision functions)

If δ^* is admissible with constant risk,
then it is a minimax decision function.

Proof:

- picture!
- Suppose that δ' had smaller worst-case risk than δ^*
- Then

$$R(\delta', \theta') \leq \sup_{\theta} R(\delta', \theta) < \sup_{\theta} R(\delta^*, \theta) = R(\delta^*, \theta'),$$

- used constant risk in the last equality
- This contradicts admissibility.

- despite this result,
  minimax decision functions are very hard to find
- Example:
  - if $X \sim N(\mu, I)$, $\dim(X) \geq 3$, then
  - X has constant risk (mean squared error) as estimator for μ
  - but: X is not an admissible estimator for μ
    therefore not minimax
  - We will discuss dominating estimator in the next part of class.

Proposition (Bayes decisions are admissible)

Suppose:

- δ^* is the Bayes decision function
- $\pi(\theta) > 0$ for all θ, $R(\delta^*, \pi) < \infty$
- $R(\delta^*, \theta)$ is continuous in θ

Then δ^* is admissible.

(We will prove the reverse of this statement in the next section.)

Sketch of proof:

- picture!
- Suppose δ^* is not admissible
- $\Rightarrow$ dominated by some δ'
  i.e. $R(\delta', \theta) \leq R(\delta^*, \theta)$ for all θ with strict inequality for some θ
- Therefore

$$R(\delta', \pi) = \int R(\delta', \theta)\pi(\theta)d\theta < \int R(\delta^*, \theta)\pi(\theta)d\theta = R(\delta^*, \pi)$$

- This contradicts δ^* being a Bayes decision function.

Proposition (Bayes risk and minimax risk)

The Bayes risk
$R(\pi) := \inf_\delta R(\delta, \pi)$
is never larger than the minimax risk
$\overline{R} := \inf_\delta \sup_\theta R(\delta, \theta)$.

Proof:

$$
\begin{aligned}
R(\pi) &= \inf_\delta R(\delta, \pi) \\
&\leq \sup_\pi \inf_\delta R(\delta, \pi) \\
&\leq \inf_\delta \sup_\pi R(\delta, \pi) \\
&\leq \inf_\delta \sup_\theta R(\delta, \theta) = \overline{R}.
\end{aligned}
$$

If there exists a prior π^* such that $R(\pi) = \overline{R}$, it is called the least favorable distribution.

Analogies to microeconomics

1) Welfare economics

statistical decision theory	social welfare analysis
different parameter values θ	different people i
risk $R(.,\theta)$	individuals' utility $u_i(.)$
dominance	Pareto dominance
admissibility	Pareto efficiency
Bayes risk	social welfare function
prior	welfare weights (distributional preferences)
minimaxity	Rawlsian inequality aversion

2) choice under uncertainty / choice in strategic interactions

statistical decision theory	**strategic interactions**
dominance of decision functions	dominance of strategies
Bayes risk	expected utility
Bayes optimality	expected utility maximization
minimaxity	(extreme) ambiguity aversion

Basic definitions

Optimality criteria

Some relationships between these optimality criteria

Analogies to microeconomics

Two justifications of the Bayesian approach

References

Two justifications of the Bayesian approach
justification 1 – the complete class theorem

- last section: every Bayes decision function is admissible
  (under some conditions)
- the reverse also holds true (under some conditions):
  every admissible decision function is Bayes,
  or the limit of Bayes decision functions
- can interpret this as:
  all reasonable estimators are Bayes estimators
- will state a simple version of this result

Preliminaries

- set of risk functions that correspond to some δ is the **risk set**,

$$\mathscr{R} := \{r(.) = R(., \delta) \text{ for some } \delta\}$$

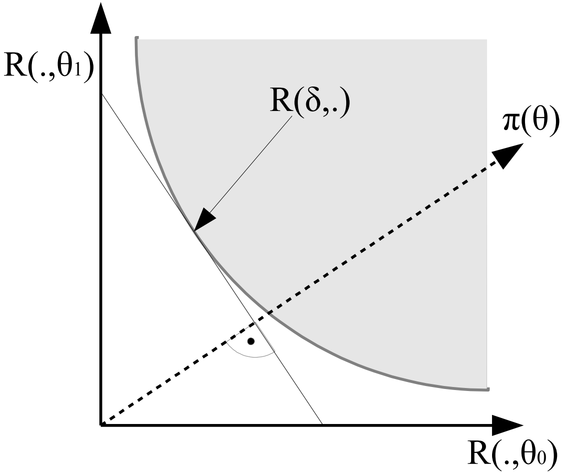
- will assume **convexity** of $\mathscr{R}$
  – no big restriction, since we can always randomly "mix" decision functions
- a class of decision functions δ is a **complete class** if it contains every admissible decision function δ^*

Theorem (Complete class theorem)

Suppose

- the set Θ of possible values for θ is compact
- the risk set $\mathscr{R}$ is convex
- all decision functions have continuous risk

Then the Bayes decision functions constitute a complete class:
For every admissible decision function δ^*, there exists a prior distribution π such that δ^* is a Bayes decision function for π.


R(.,θ1)
R(δ,.)
π(θ)
R(.,θ0)

Intuition for the complete class theorem

- any choice of decision procedure has to trade off risk across θ
- slope of feasible risk set
  = relative "marginal cost" of decreasing risk at different θ
- pick a risk function on the admissible frontier
- can rationalize it with a prior
  = "marginal benefit" of decreasing risk at different θ
- for example, minimax decision rule:
  rationalizable by least favorable prior
  slope of feasible set at constant risk admissible point
- analogy to social welfare: any policy choice or allocation corresponds to distributional preferences / welfare weights

Proof of complete class theorem:

- application of the separating hyperplane theorem, to the space of functions of θ, with the inner product

$$\langle f,g\rangle = \int f(\theta)g(\theta)d\theta.$$

- for intuition: focus on binary θ, $\theta \in \{0,1\}$, and $\langle f,g\rangle = \sum_\theta f(\theta)g(\theta)$
- Let δ^* be admissible. Then $R(.,\delta^*)$ belongs to the lower boundary of $\mathscr{R}$.
- convexity of $\mathscr{R}$, separating hyperplane theorem separating $\mathscr{R}$ from (infeasible) risk functions dominating δ^*

- $\Rightarrow$ there exists a function $\tilde{\pi}$ (with finite integral) such that for all δ

$$\langle R(.,\delta^*),\tilde{\pi}\rangle \leq \langle R(.,\delta),\tilde{\pi}\rangle.$$

- by construction $\tilde{\pi} \geq 0$
- thus $\pi := \tilde{\pi}/\int\tilde{\pi}$ defines a prior distribution.
- δ^* minimizes

$$\langle R(.,\delta^*),\pi\rangle = R(\delta^*,\pi)$$

  among the set of feasible decision functions
- and is therefore the optimal Bayesian decision function for the prior π.

justification 2 – subjective probability theory

- going back to Savage (1954) and Anscombe and Aumann (1963).
- discussed in chapter 6 of
  Mas-Colell, A., Whinston, M., and Green, J. (1995), *Microeconomic theory*, Oxford University Press
- and maybe in Econ 2010 / Econ 2059.

- Suppose a decision maker ranks risk functions $R(., \delta)$ by a **preference relationship** $\succeq$
- properties $\succeq$ might have:
  1. **completeness**: any pair of risk functions can be ranked
  2. **monotonicity**: if the risk function R is (weakly) lower than R' for all θ, than R is (weakly) preferred
  3. **independence**:

$$R^1 \succeq R^2 \Leftrightarrow \alpha R^1 + (1-\alpha)R^3 \succeq \alpha R^2 + (1-\alpha)R^3$$

     for all R^1, R^2, R^3 and $\alpha \in [0,1]$

- Important: this independence has nothing to do with statistical independence

Theorem

If $\succeq$ is complete, monotonic, and satisfies independence, then there exists a prior π such that

$$R(.,\delta^1) \succeq R(.,\delta^2) \Leftrightarrow R(\pi,\delta^1) \leq R(\pi,\delta^2).$$

Intuition of proof:

- Independence and completeness imply linear, parallel indifference sets
- monotonicity makes sure prior is non-negative

Sketch of proof:
Using independence repeatedly, we can show that for all $R^1, R^2, R^3 \in \mathbb{R}^{\mathscr{X}}$, and all $\alpha > 0$,

1. $R^1 \succeq R^2$ iff $\alpha R^1 \succeq \alpha R^2$,
2. $R^1 \succeq R^2$ iff $R^1 + R^3 \succeq R^2 + R^3$,
3. $\{R : R \succeq R^1\} = \{R : R \succeq 0\} + R^1$,
4. $\{R : R \succeq 0\}$ is a convex cone.
5. $\{R : R \succeq 0\}$ is a half space.

The last claim requires completeness. It immediately implies the existence of π. Monotonicity implies that π is not negative.

Remark

- personally, I’m more convinced by the complete class theorem than by normative subjective utility theory
- admissibility seems a very sensible requirement
- whereas “independence” of the preference relationship seems more up for debate

References

Robert, C. (2007). The Bayesian choice: from decision-theoretic foundations to computational implementation. *Springer Verlag, chapter 2.*

Casella, G. and Berger, R. L. (2001). Statistical inference. *Duxbury Press, chapter 7.3.4.*