Decision Theory Basics

Andrew Nobel

August 2024

Decision Theory Overview

Decision theory provides a framework for specifying and formalizing inference problems. Components include

- Statistical model: Family of distributions governing observations
- Observations: access to information about unknown quantities
- Inference target: Parameter indexing distributions in the statistical model
- Decision rules: take action based on observations
- Loss and risk: assessments of performance, comparison of decision rules

Statistical Inference

Setting: Random experiment yields measurement vector $X = (X_1, \ldots, X_n)$ taking values in a *sample space* $\mathcal{X}$

- Typically $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{X} \subseteq \mathbb{N}^n$, or $\mathcal{X} \subseteq \{0, 1\}^n$
- Random measurement X called *observation(s)*
- Realized value $x = (x_1, \ldots, x_n)$ of X called *data*

Statistical model: An indexed family $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$ of pdf's or pmf's on $\mathcal{X}$

- Index set Θ called *parameter space*. Typically $\Theta \subseteq \mathbb{R}^d$
- Elements $\theta \in \Theta$ called *parameters*

Statistical Inference, cont.

Key Assumption: Distribution of X is an (unknown) element $f(\cdot|\theta)$ of model $\mathcal{P}$

- Model $\mathcal{P}$ is properly specified (often an optimistic assumption)

Inference: Reason about true parameter θ based on observation X

- Reasoning involves actions, formalized by decision rules
- Quality of actions measured by a loss function
- Quality of decision rules measured by risk function

Ingredients

Decision rule. A map $d : \mathcal{X} \to \mathcal{A}$ from observations to actions

- Set $\mathcal{A}$ is the *action space*, depends on the inference problem under study
- Each decision rule d represents an inference procedure

Allowable rules. Family $\mathcal{D}$ of decision rules $d : \mathcal{X} \to \mathcal{A}$ under study. Depends on

- Nature of inference problem
- Criteria such as invariance and unbiasedness

Loss function. Function $\ell : \Theta \times \mathcal{A} \to \mathbb{R}$.

- $\ell(\theta, a) =$ cost if we make decision a when true parameter is θ

Example: Point Estimation

Set-Up

- Observation X in sample space $\mathcal{X} = \mathbb{R}^n$ or $\mathcal{X} = \mathbb{N}^n$
- Model $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^d$
- Assume $X \sim f(\cdot|\theta)$ where $\theta \in \Theta$ is unknown

Inference: Estimate θ based on observed value x of X

- Action space $\mathcal{A} = \Theta$
- Decision rule $d : \mathcal{X} \to \Theta$ is an *estimator*. Common to write $d(x)$ as $\hat{\theta}(x)$
- Squared loss $\ell(\theta, \theta') = (\theta - \theta')^2$ or absolute loss $\ell(\theta, \theta') = |\theta - \theta'|$

Example: Hypothesis Testing

Set-Up

- Observation X in sample space $\mathcal{X} = \mathbb{R}^n$ or $\mathcal{X} = \mathbb{N}^n$
- Model $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$
- Partition $\Theta = \Theta_0 \cup \Theta_1$ of parameter space
- Observation $X \sim f(\cdot|\theta)$ where $\theta \in \Theta$ is unknown

Inference: Decide if $\theta \in \Theta_0$ or $\theta \in \Theta_1 = \Theta_0^c$ based on value x of X

- Action space $\mathcal{A} = \{0, 1\}$
- Decision rule $d : \mathcal{X} \to \{0, 1\}$ is an *hypothesis test*
- Zero-one loss $\ell(\theta, a) = \mathbb{I}(\theta \notin \Theta_a)$ (1 if decision is incorrect, 0 otherwise)

Example: Interval Estimatioin

Set-Up

- Observation X in sample space $\mathcal{X} = \mathbb{R}^n$ or $\mathcal{X} = \mathbb{N}^n$
- Model $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}$
- Assume $X \sim f(\cdot|\theta)$ where $\theta \in \Theta$ is unknown

Inference: Find *confidence set* $C \subseteq \Theta$ likely to contain θ based on value x of X

- Action space $\mathcal{A} =$ family of subsets of Θ (for example, intervals)
- Decision rule $d : \mathcal{X} \to \mathcal{A}$ is an interval estimate
- Weighted 0-1 loss $\ell(\theta, C) = \mathbb{I}(\theta \notin C) + \lambda \,\text{length}(C)$, some $\lambda > 0$

The Risk Function

The *risk function* of a decision rule tells us how well that rule performs for each possible parameter $\theta \in \Theta$. It is the basis for comparing decision rules

Definition: The *risk function* of a decision rule $d : \mathcal{X} \to \mathcal{A}$ is defined by

$$R(\theta, d) \;=\; \mathbb{E}_\theta \ell(\theta, d(X)) \quad \theta \in \Theta$$

- Notation: $\mathbb{E}_\theta h(X)$ is the expectation of $h(X)$ when $X \sim f(\cdot|\theta)$
- $R(\theta, d) =$ expected loss of rule d when applied to observation $X \sim f(\cdot|\theta)$
- Continuous case $R(\theta, d) = \int \ell(\theta, d(x)) f(x|\theta)\, dx$
- Discrete case $R(\theta, d) = \sum_x \ell(\theta, d(x))\, p(x|\theta)$

Point Estimation Under Squared Loss

Given family $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$, and an estimator $\hat{\theta} : \mathcal{X} \to \Theta$

- The *bias* of $\hat{\theta}$ at θ is $\text{bias}_\theta(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}(X)] - \theta$
- The *variance* of $\hat{\theta}$ at θ is $\text{Var}_\theta(\hat{\theta}) = \mathbb{E}_\theta \left[\hat{\theta}(X) - \mathbb{E}_\theta\, \hat{\theta}(X)\right]^2$
- Say $\hat{\theta}$ is *unbiased* if $\text{bias}_\theta(\hat{\theta}) = 0$ for all θ

Bias-Variance Decomposition: Under the squared loss $\ell(\theta, a) = (\theta - a)^2$

$$R(\theta, \hat{\theta}) = \text{Var}_\theta(\hat{\theta}) + (\text{bias}_\theta(\hat{\theta}))^2$$

Upshot: For $\hat{\theta}$ to perform well when $X \sim f(\cdot|\theta)$ it should

- Be centered near the true parameter (small bias)
- Not be too spread out (small variance)

Example: Estimation of a Normal Mean

Setting

- Single observation $X \in \mathbb{R}$
- $\mathcal{P} = \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$
- Estimate θ from X under the squared error loss

Two estimators

- $\hat{\theta}_1(x) = x$, risk function $R(\theta, \hat{\theta}_1) = 1$
- $\hat{\theta}_2(x) = 3$, risk function $R(\theta, \hat{\theta}_2) = (\theta - 3)^2$

Neither risk function dominates the other

Example: Probability of Success in Bernoulli Trial

Setting

- Observations $X = (X_1, \ldots, X_n) \in \{0,1\}^n$
- $\mathcal{P} = \{\text{Bern}(\theta)^n : \theta \in (0,1)\}$
- Estimate θ from X under the squared error loss

Two estimators

$$\hat{\theta}_1(x) = \overline{x} \qquad R(\theta, \hat{\theta}_1) = \frac{\theta(1-\theta)}{n}$$

$$\hat{\theta}_2(x) = \frac{n\overline{x} + \sqrt{n}/2}{n + \sqrt{n}} \qquad R(\theta, \hat{\theta}_2) = \frac{1}{4(1+\sqrt{n})^2} \quad \text{(constant)}$$

Neither risk function dominates the other

Admissibility

Setting: General inference problem with family $\mathcal{D}$ of candidate decision rules

Definition: A decision rule $d \in \mathcal{D}$ is *inadmissible* if there is some $d' \in \mathcal{D}$ such that

(i) $R(\theta, d') \leq R(\theta, d)$ for all $\theta \in \Theta$

(ii) $R(\theta, d') < R(\theta, d)$ for some $\theta \in \Theta$

If no such d' exists, then d is said to be *admissible*

- Admissibility depends on the family $\mathcal{D}$ and the loss function ℓ
- A rule d is either admissible or inadmissible
- Admissible rules are candidates for good/reasonable rules
- There may be many admissible rules
- Admissibility is a weak criterion. Obviously silly rules can be admissible.

Example

Observations: $X_1, \ldots, X_n$ i.i.d. Bern(θ) with $\theta \in (0, 1)$

Goal: Estimate of θ under squared loss. Candidate estimators

- $\hat{\theta}_1(x) = \overline{x}$ with $R(\theta, \hat{\theta}_1) = \theta(1-\theta)/n$
- $\hat{\theta}_2(x) = x_1$ with $R(\theta, \hat{\theta}_2) = \theta(1-\theta)$
- $\hat{\theta}_3(x) = \frac{1}{2}$ with $R(\theta, \hat{\theta}_3) = (\theta - \frac{1}{2})^2$

Fact

1. $\hat{\theta}_1$ is admissible
2. $\hat{\theta}_2$ is inadmissible (bettered by $\hat{\theta}_1$)
3. $\hat{\theta}_3$ is *admissible* (lazy, but unbeatable when $\theta = \frac{1}{2}$)

Frequentist and Bayesian Perspectives on Inference

Different approaches stemming in part from different interpretations of probability

Frequentist

- Probability defined through repetitions of a random experiment
- True parameter θ is a fixed element of Θ, but otherwise unknown
- Analysis and interpretation of inference based on (potentially unrealized) replications of basic experiment

Bayesian

- Probability understood as a (potentially subjective) measure of belief
- Belief about true parameter before/after an experiment represented by prior/posterior distributions on the parameter space Θ
- Experiment regarded as unique. Inference based on updating prior based on data, without reference to other experiments or repetition

Overview of Bayesian Inference

Basic ingredients

- Family $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$ of sampling densities on $\mathcal{X}$
- Prior density $\pi(\theta)$ on parameter space Θ
- Joint densities $f(x, \theta) = f(x|\theta)\pi(\theta)$, marginal density $m(x) = \int f(x, \theta)d\theta$
- Observation model: First θ drawn from π, then X drawn from $f(x|\theta)$

Idea: Prior density $\pi(\theta)$ reflects belief/information about parameters before experiment is conducted. Given data x, we update prior using Bayes formula to obtain

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)} \quad \textit{posterior density}$$

Key point: *All inferences about* θ, (point estimates, hypothesis tests, interval estimates) *are based on the posterior density*

Comparing Decision Rules: Different Perspectives

Recall: Risk of decision rule $d : \mathcal{X} \to \Theta$ under loss ℓ summarized by risk *function*

$$R(\theta, d) \;=\; \mathbb{E}_\theta \ell(\theta, d(X))$$

Question: How should we compare two decision rules d_1 and d_2 based on their risk functions $R(\theta, d_1)$ and $R(\theta, d_2)$?

- Frequentist perspective: Consider *maximum risk* of each rule over all $\theta \in \Theta$
- Bayesian perspective: Consider *average risk* of each rule relative to prior π

Maximum Risk and Bayes Risk

Idea: Single number summaries of overall risk obtained from risk function $R(\theta, d)$

Definition: Given family $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$ and loss function $\ell : \Theta \times \mathcal{A} \to \mathbb{R}$

(i) The *maximum risk* of a decision rule $d : \mathcal{X} \to \mathcal{A}$ is

$$R_m(d) = \max_{\theta \in \Theta} R(\theta, d)$$

(ii) The *Bayes risk* of a decision rule $d : \mathcal{X} \to \mathcal{A}$ under prior density π is

$$R_\pi(d) = \int R(\theta, d)\pi(\theta)d\theta$$

Example: Probability of Success in Bernoulli Trial

Recall: Observe $X_1, \ldots, X_n \sim \text{Bern}(\theta)$. Estimators $\hat{\theta}_1, \hat{\theta}_2$ for θ with

$$R(\theta, \hat{\theta}_1) = \frac{\theta(1-\theta)}{n} \qquad R(\theta, \hat{\theta}_2) = \frac{n}{4(n+\sqrt{n})^2}$$

A. *Maximum risk:* Prefer estimator $\hat{\theta}_2(x)$ as

$$R_m(\hat{\theta}_1) = \frac{1}{4n} > \frac{1}{4(1+\sqrt{n})^2} = R_m(\hat{\theta}_2)$$

B. *Bayes risk:* Under uniform prior $\pi(\theta) = 1$, prefer estimator $\hat{\theta}_1$ for $n \geq 20$ as

$$R_\pi(\hat{\theta}_1) = \frac{1}{6n} < \frac{1}{4(1+\sqrt{n})^2} = R_\pi(\hat{\theta}_2)$$

Minimax and Bayes Rules for a Family $\mathcal{D}$

Definition: The *minimax risk* for a family of decision rules $\mathcal{D}$ is

$$R_m^* = \min_{d \in \mathcal{D}} R_m(d) = \min_{d \in \mathcal{D}} \max_{\theta \in \Theta} R(\theta, d)$$

A rule $d \in \mathcal{D}$ is said to be *minimax* if $R_m(d) = R_m^*$.

Definition: The optimal Bayes risk for a family of decision rules $\mathcal{D}$ under a prior π is

$$R_\pi^* = \min_{d \in \mathcal{D}} R_\pi(d) = \min_{d \in \mathcal{D}} \int R(\theta, d)\pi(\theta)\, d\theta$$

A rule $d \in \mathcal{D}$ is called a *Bayes rule* for π if $R_\pi(d) = R_\pi^*$. Note that R_π^* depends on π

Fact: Minimax risk is always bounded below by the optimal Bayes risk: for every prior distribution π on Θ one has $R_m^* \geq R_\pi^*$

Bayes Rules with Constant Risk are Minimax

Theorem: Let d_π be the Bayes rule for a family $\mathcal{D}$ under a prior π. If the risk function $R(\theta, d_\pi)$ is constant then d_π is minimax for $\mathcal{D}$.

Terminology: If d_π is minimax then π is said to be a *least favorable* prior

Example: Let $X_1, \ldots, X_n \sim \text{Bern}(\theta)$. Consider family $\mathcal{D}$ of all point estimators of θ under the squared loss. Consider point estimator

$$\hat{\theta}(x) = \frac{n\overline{x}_n + \sqrt{n}/2}{n + \sqrt{n}}$$

Can show that

- $R(\theta, \hat{\theta}) = 1/4(1 + \sqrt{n})^2$ is constant
- $\hat{\theta}$ is the Bayes rule for $\mathcal{D}$ under prior $\pi_0 = \text{Beta}(\sqrt{n}/2, \sqrt{n}/2)$ prior

By Theorem, $\hat{\theta}$ is minimax among all estimators of θ, and π_0 is least favorable prior

Admissibility of Bayes Rules

Fact: Consider a Bayesian decision problem in which

- $\Theta \subseteq \mathbb{R}^p$ is open
- $\pi(\theta) > 0$ for every $\theta \in \Theta$
- R^*_π is finite
- $R(\theta, d)$ is a continuous function of θ for each $d \in \mathcal{D}$

Then the Bayes rule for π is admissible.