Statistical Approaches to
Learning and Discovery

Week 3: Elements of Decision Theory

January 29, 2003

Decision Theory

Statistical decision theory – making decisions in the presence of statistical knowledge.

Example (Berger): A drug company is deciding whether or not to market a new pain reliever. Two important factors:

1. Proportion of people θ_1 for whom the drug will be effective

2. Market share θ_2 the drug will capture

$\theta = (\theta_1, \theta_2)$ are unknown, but the company needs to decide whether to market the drug, the price, etc.

Decision Theory

For each in a set of actions $a \in \mathcal{A}$, if the parameter is θ, a *loss* $L(a, \theta)$ is associated with choosing action a.

The *risk* is the expected loss:

$$R = \int_{\Theta} L(a, \theta) \, dF(\theta)$$

and one chooses the action that minimizes the risk.

Simple Example

Suppose that the company wants to estimate market share θ_2.

The "action" chosen is to use a certain estimate of this in further management decisions.

Suppose

$$L(\theta_2, a) = \begin{cases} 2(\theta_2 - a) & \text{if } \theta_2 - a \geq 0, \\ a - \theta_2 & \text{if } \theta_2 - a \leq 0 \end{cases}$$

An underestimate is penalized more than an overestimate

Simple Example (cont.)

Suppose that the company does a study, interviews n people and finds X people would buy the drug.

Assume $X = \text{Binom}(n, \theta_2)$. Then

$$f(\theta_2 \,|\, x) \;\propto\; \binom{n}{x} \theta_2^x \,(1-\theta_2)^{n-x}\, f(\theta_2)$$

$f(\theta_2)$ might be affected by previous drugs marketed, etc., and is very important in this case.

Example from Information Retrieval

1. Two parts of IR problem: modeling documents and queries

2. Making a decision on what documents to present to the user

Naturally cast in framework of statistical decision theory.

(C. Zhai CMU thesis, 2002).

Some Definitions

$\theta \in \Theta$: "state of nature" — hidden, random

$a \in \mathcal{A}$: possible actions

$X \in \mathcal{X}$: observables, experiments – info about θ

Bayesian expected loss is

$$\rho(\pi, a) = E_\pi[L(\theta, a)] = \int L(\theta, a)\, dF^\pi(\theta)$$

Conditioned on evidence in data X, we average with respect to the posterior:

$$\rho(\pi, a \,|\, X) = E_{\pi(\cdot\,|\,X)}[L(\theta, a)] = \int L(\theta, a)\, p(\theta \,|\, X)$$

Frequentist formulation, $\delta : \mathcal{X} \longrightarrow \mathcal{A}$ a decision rule, *risk function*

$$R(\theta, \delta) = E_X[L(\theta, \delta(X)] = \int_{\mathcal{X}} L(\theta, \delta(X))\, dF^X(x)$$

Bayes Risk

For a prior π, the *Bayes risk* of a decision function is defined by

$$r(\pi, \delta) = E_\pi[R(\theta, \delta)] = E_\pi\left[E_X[L(\theta, \delta(X))]\right]$$

Therefore, the classical and Bayesian approaches define different risks, by averaging:

- Bayesian expected loss: Averages over θ
- Risk function: Averages over X
- Bayes risk: Averages over both X and θ

Admissibility

A decision rule δ_1 is *R-better* than δ_2 in case

$$\begin{aligned} R(\theta, \delta_1) &\leq R(\theta, \delta_2) \quad \text{for all } \theta \in \Theta \\ R(\theta, \delta_1) &< R(\theta, \delta_2) \quad \text{for some } \theta \in \Theta \end{aligned}$$

δ is *admissible* if there exists no R-better decision rule. Otherwise, it's *inadmissible*.

Example

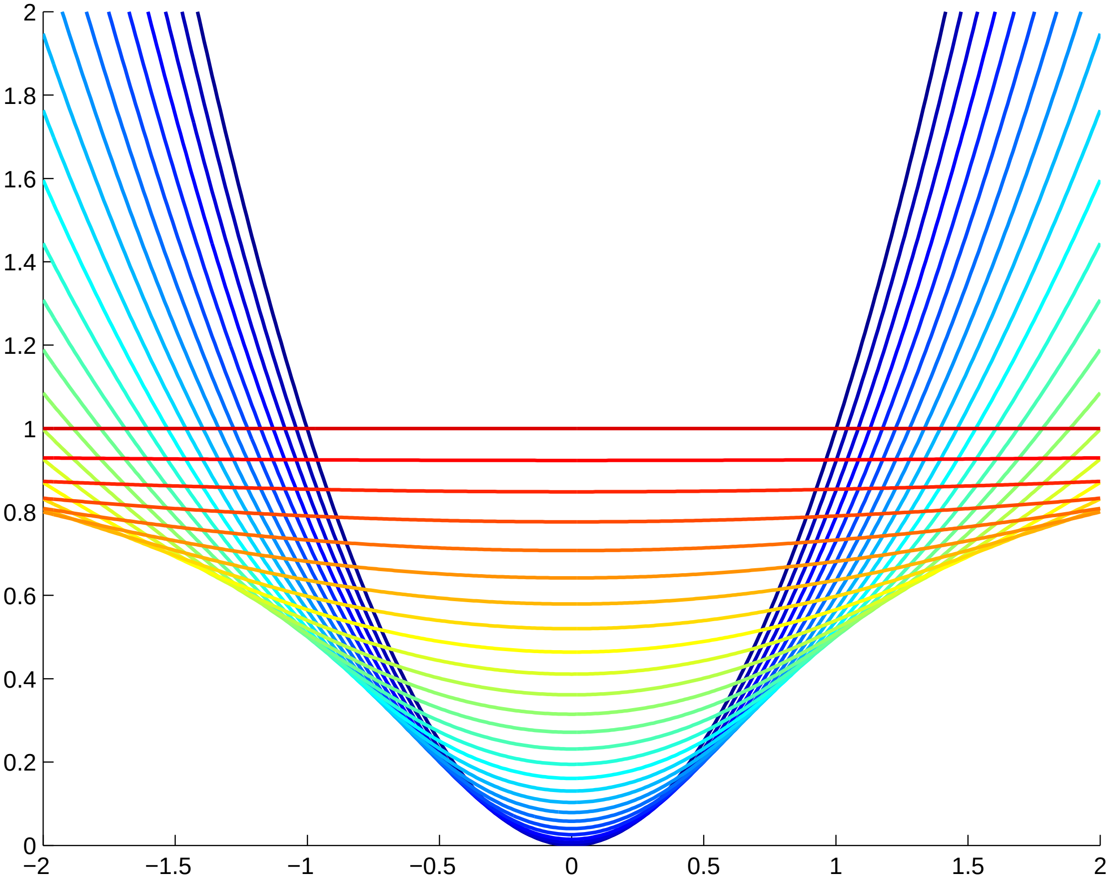
Take $X \sim \mathcal{N}(\theta, 1)$, and problem of estimating θ under square loss $L(\theta, a) = (a - \theta)^2$. Consider decision rules of the form $\delta_c(x) = cx$.

A calculation gives that

$$R(\theta, \delta_c) = c^2 + (1-c)^2\theta^2$$

Then δ_c is inadmissible for $c > 1$, and admissible for $0 \leq c \leq 1$.

Example (cont.)



Risk $R(\theta, \delta_c)$ for admissible decision functions $\delta_c(x) = cx$, $c \leq 1$, as a function of θ. The color corresponds the associated minimum Bayes risk.

Example (cont.)

Consider now $\pi = \mathcal{N}(0, \tau^2)$. Then the Bayes risk is

$$r(\pi, \delta_c) = c^2 + (1-c)^2\tau^2$$

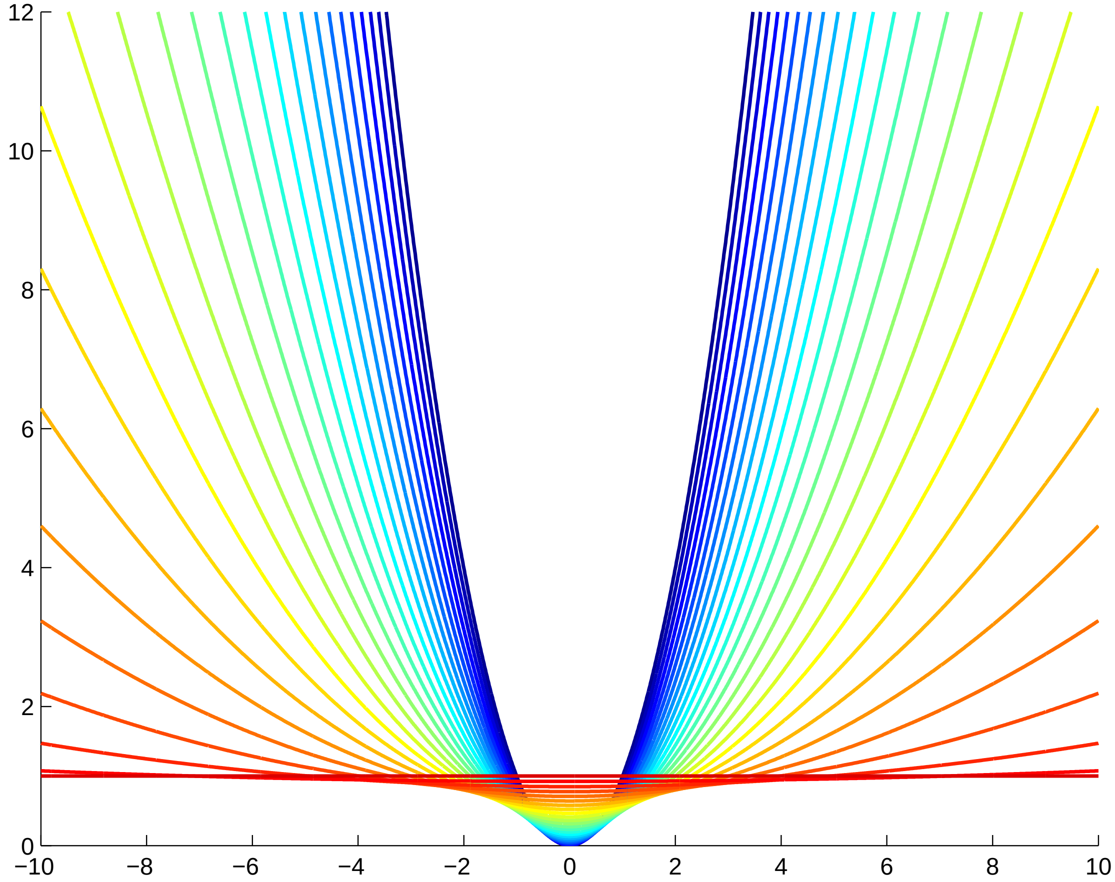
Thus, the best Bayes risk is obtained by the Bayes estimator δ_{c^*} with

$$c^* = \frac{\tau^2}{1+\tau^2}$$

and this is the same value of the Bayes risk of π. That is, each δ_c is Bayes for the conjugate $\mathcal{N}(0, \tau_c^2)$ prior with

$$\tau_c = \sqrt{\frac{c}{1-c}}$$

Example (cont.)



At a larger scale, it becomes clearer that the decision function with $c = 1$ is minimax. It corresponds to the (improper) conjugate prior $\mathcal{N}(0, \tau^2)$ with $\tau \to \infty$.

Simplifying

Basic fact: When loss is convex and there is a sufficient statistic T for θ, only non-randomized decision rules based on T need be considered.

See Berger, Chap. 1 for details and examples.

Bayes Actions

$\delta^\pi(x)$ is a *posterior Bayes action for x* if it minimizes

$$\int_\Theta L(\theta, a)\, p(\theta \,|\, x)\, d\theta$$

Equivalently, it minimizes

$$\int_\Theta L(\theta, a)\, f(x \,|\, \theta)\, \pi(\theta)\, d\theta$$

Need not be unique.

Equivalence of Bayes actions and Bayes decision rules

A decision rule δ^π minimizing the Bayes risk $r(\pi, \delta)$ can be found "pointwise," by minimizing

$$\int_\Theta L(\theta, a)\, p(x \mid \theta)\, \pi(\theta)\, d\theta$$

for each x. So, the two problems are equivalent.

Special Case: Squared Loss

For $L(\theta, a) = (\theta - a)^2$, the Bayes rule is the posterior mean

$$\delta^\pi(x) = E[\theta \,|\, x]$$

For weighted squared loss, $L(\theta, a) = w(\theta)(\theta - a)^2$, the Bayes rule is weighted posterior mean:

$$\delta^\pi(x) = \frac{\int_\Theta \theta\, w(\theta)\, f(x \,|\, \theta)\, \pi(\theta)\, d\theta}{\int_\Theta \theta\, w(\theta)\, f(x \,|\, \theta)\, \pi(\theta)\, d\theta}$$

Note: w acts like a prior here

We will see later how L^2 case—posterior mean—applies to some classification problems, in particular learning with labeled/unlabeled data.

Special Case: L^1 Loss

For $L(\theta, a) = |\theta - a|$, the Bayes rule is a posterior median.

More generally, for

$$L(\theta, a) = \begin{cases} c_0(\theta - a) & \theta - a \geq 0 \\ c_1(a - \theta) & \theta - a < 0 \end{cases}$$

a $\frac{c_0}{c_0+c_1}$-fractile of posterior $p(\theta \,|\, x)$ is a Bayes estimate.

Conjugacy

Note that if $X \sim$ exponential family under square loss, restricting to linear estimators can turn out to be equivalent to using a conjugate prior – by Diaconis and Ylvisaker.

See Berger, §4.7.9 for discussion and examples

Problem 1: Channel Capacity

$$1010010001 \longrightarrow \boxed{Q(y \mid x)} \longrightarrow 1011010101$$

What is the maximum rate at which information can be sent with arbitrarily small probability of error?

For a code $\mathbb{C}$ with M codewords of length n bits,

$$\text{Rate}(\mathbb{C}) = \frac{\log_2 M}{n}$$

Problem 2: Minimax Risk

I choose model θ, generate *iid* examples $\mathbf{y} = y_1, \ldots, y_n$ according to $Q(\cdot \mid \theta)$. You predict using estimate $\hat{P}(y_t \mid y^{t-1})$.

Risk (expected loss) after n steps:

$$
\begin{aligned}
R_{n,\hat{P}}(\theta^*) &\overset{\text{def}}{=} \sum_{k=1}^{n} \int_{\mathcal{Y}^k} Q^k(y^k \mid \theta^*) \log \frac{Q(y_k \mid \theta^*)}{\hat{P}(y_k \mid y^{k-1})} \, dy^k \\
&= D(Q^n_{\theta^*} \,\|\, \hat{P})
\end{aligned}
$$

Minimax risk:

$$
R_n^{\textit{minimax}} \overset{\text{def}}{=} \inf_{\hat{P}} \sup_{\theta^* \in \Theta} R_{n,\hat{P}}(\theta^*)
$$

Problem 3: Non-informative Priors

- In Bayesian statistics, a “non-informative” prior is one that is “most objective,” encoding the least amount of prior knowledge.

- With a non-informative prior, even moderate amounts of data should dominate the prior information.

- Many contend there is no truly “objective” prior that represents ignorance.

Connections Between These Problems

Shannon showed that the engineering notion of channel capacity is the same as the *information capacity*:

$$C(Q) \; = \; \sup_{P} I(X, Y)$$

Where $\sup$ is over all distributions $P(X)$ on the input to the channel.

Connections Between These Problems (cont)

Theorem (Haussler, 1997). The minimax risk is equal to the information capacity:

$$R_n^{\textit{minimax}} = \sup_P R_{n,P}^{\textit{Bayes}} = \sup_P I(\Theta, Y^n)$$

Moreover, the minimax risk can be written as a minimax with respect to Bayes strategies:

$$R_n^{\textit{minimax}} = \inf_P \sup_{\theta^* \in \Theta} R_{n, P_{\textit{Bayes}}}(\theta^*)$$

where $P_{\textit{Bayes}}$ denotes the predictive distribution (Bayes strategy) for $P \in \Delta_\Theta$.

Connections Between These Problems (cont)

Can use information-theoretic measures to define *reference priors* (Bernardo *et al.*)

For a parametric family $\{Q(y \,|\, \theta)\}_{\theta \in \Theta}$, define

$$\pi_k \;=\; \text{argmax}_P \, I(\Theta, Y^k)$$

where

$$I(\Theta, Y^k) = \int_\Theta \int_{\mathcal{Y}^k} P(\theta)\, Q^k(y^k \,|\, \theta)\; \log \frac{Q^k(y^k \,|\, \theta)}{M(y^k)}\, dy^k \, d\theta$$

Connections Between These Problems (cont)

Bernardo (1979) proposed *reference priors* defined by

$$\pi(\theta) \;=\; \lim_{k \to \infty} \pi_k(\theta)$$

when this exists.

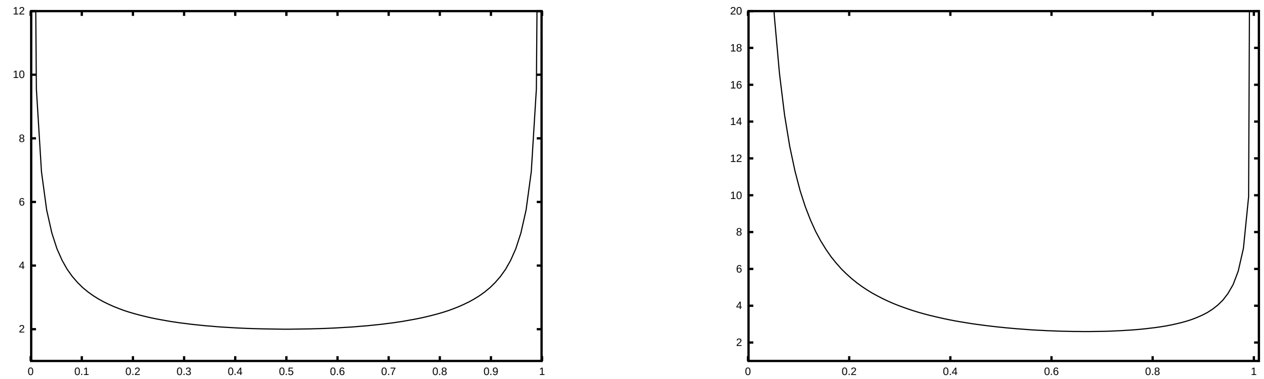
Thus, channel capacity, minimax risk, and reference priors all given by maximizing mutual information.

Jeffreys Priors

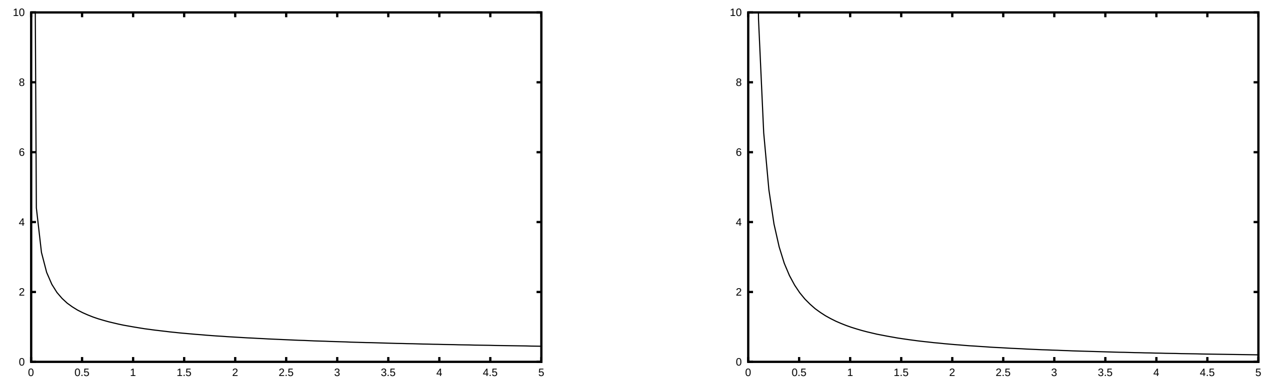
For $\Theta \subset \mathbb{R}$, if the posterior is asymptotically normal, the limiting reference prior is given by Jeffreys' rule:

$$
\begin{aligned}
\pi(\theta) &\propto h(\theta)^{1/2} \\
h(\theta) &= \int_{\mathcal{X}} Q(x \,|\, \theta) \left(-\frac{\partial^2}{\partial \theta^2} \log Q(x \,|\, \theta) \right) dx
\end{aligned}
$$

Jeffreys Priors



binomial, $\pi(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$ neg. binom., $\pi(\theta) \propto \theta^{-1}(1-\theta)^{-\frac{1}{2}}$



Poisson, $\pi(\theta) \propto \theta^{-\frac{1}{2}}$ exponential, etc. $\pi(\theta) \propto \theta^{-1}$

Finite Sample Sizes

For finite k, little is known about the reference prior π_k.

If $Q(\cdot \mid \theta)$ is from the exponential family, then π_k is a finite discrete measure.

(Berger, Bernardo, and Mendoza, 1989)

"*Solving for π_k explicitly is not easy....Numerical solution is needed*."

Blahut-Arimoto Algorithm

- In information theory, input to channel is typically discrete.
- Convex optimization problem
- Simple iterative algorithm discovered independently in early 1970s by Blahut and Arimoto.
- Allows easy calculation of capacity for arbitrary channels (even with constraints).

Blahut-Arimoto Algorithm

Initialize: Let $P^{(0)}$ be arbitrary, $t = 0$.

Iterate until convergence:

1. $M^{(t)}(y) = \sum_x P^{(t)}(x)\, Q(y \,|\, x)$

2. $P^{(t+1)}(x) = \frac{P^{(t)}(x)\, C^{(t)}(x)}{\sum_x P^{(t)}(x)\, C^{(t)}(x)}$

where $C^{(t)}(x) = \exp\left(\sum_{y \in \mathcal{Y}} Q(y \,|\, x) \log \frac{Q(y \,|\, x)}{M^{(t)}(y)}\right)$

3. $t \leftarrow t + 1$

MCMC version developed in (L. and Wasserman, 2001)