Bayesian Statistics
Statistical Decision Theory

Shaobo Jin

Department of Mathematics

Basic Terminology

Let θ be an unknown quantity of interest. Θ is used to denote the set of all possible values of θ.

- If θ is a parameter in a statistical model, then Θ is the parameter space.

We will take a decision (or an action) d based on the observed data x, such as $d = \delta(x)$.

- The set $\mathcal{X}$ of all possible observations is called a sample space.
- The set $\mathcal{D}$ of all possible decisions is called a decision space.
- The function $\delta(x)$ is called a decision rule.

Decision Space: Example

Classification: Consider the problem of predicting $y_i \in \{0, 1\}$.

- The decision space is $\mathcal{D} = \{0, 1\}$ for 0-1 classification.
- The decision space is $\mathcal{D} = [0, 1]$ for probabilistic classification.

Estimation: Let $\theta \in \Theta \subseteq \mathbb{R}^p$ be the parameter vector. We are interested in θ.

- The decision space is $\mathcal{D} = \Theta \subseteq \mathbb{R}^p$.

Prediction: Let $y \in \mathcal{X}$ be a future value that we want to predict.

- The decision space is $\mathcal{D} = \mathcal{X}$.

Loss and Risk

Definition (Loss function)

A loss function $L(\theta, d)$ is any non-negative function $L : \Theta \times \mathcal{D} \to [0, \infty)$.

For example:

$$\begin{aligned} L_2 \text{ loss}: &\qquad L(\theta - d) = (\theta - d)^2 \\ L_1 \text{ loss}: &\qquad L(\theta - d) = |\theta - d| \end{aligned}$$

Once we apply the loss function to the decision rule $\delta(x)$, we should treat $L(\theta, \delta(x))$ as a realization from the random variable $L(\theta, \delta(X))$.

Definition (Risk)

The (frequentist) risk is

$$R(\theta, \delta) = \mathrm{E}[L(\theta, \delta(X)) \mid \theta] = \int L(\theta, \delta(x)) f(x \mid \theta)\, dx.$$

Loss and Risk: Example

Example

Let $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T$ be a vector of iid random variables from the Bernoulli distribution $\text{Bernoulli}(p)$. We are interested in p.

- The sample space is $\mathcal{X} = [0,1]$. The parameter space is $\Theta = [0,1]$.
- The decision space is $\mathcal{D} = [0,1]$.
- If we choose the loss function $L(\theta - d) = (\theta - d)^2$ and decision rule $\delta(X) = \bar{X}$, the risk is

$$R(\theta,\delta) = \mathrm{E}\left[L(p,\delta(X)) \mid p\right] = \mathrm{E}\left[\left(p - \bar{X}\right)^2 \mid p\right] = \frac{p(1-p)}{n},$$

where $\theta = p$ is treated as a fixed quantity here.

Integrated Risk

Definition (Integrated Risk)

The integrated risk is the expectation of the risk with respect to the prior $\pi(\theta)$:

$$\mathrm{E}\left[L(\theta,\delta)\right] = \int R(\theta,\delta)\,\pi(\theta)\,d\theta = \int \mathrm{E}\left[L(\theta,\delta(X)) \mid \theta\right]\pi(\theta)\,d\theta.$$

The decision that minimizes the integrated risk is called the Bayes decision rule (or Bayes estimator). The minimal integrated risk

$$\inf_{\delta} \mathrm{E}\left[L(\theta,\delta)\right]$$

is called the Bayes risk.

Find Bayes Decision

Let the posterior risk be

$$\mathrm{E}\left[L(\theta,\delta) \mid X = x\right] = \int L(\theta,\delta)\,\pi(\theta \mid x)\,d\theta.$$

Theorem (Find Bayes decision rule via posterior risk)

Suppose that

1. *there exists a decision rule with finite risk,*
2. *for almost all x, there exists a $\delta(x)$ minimizing the posterior risk $E\left[L(\theta,\delta) \mid X = x\right]$.*

Then, $\delta(x)$ is a Bayes decision rule.

Take-home Question: does the prior $\pi(\theta)$ have to be proper in order to apply this theorem?

Weighted L_2 Loss

Consider the weighted L_2 loss

$$L_W(\theta, d) = (\theta - d)^T W (\theta - d),$$

where W is a $p \times p$ symmetric and positive definite matrix.

Theorem

Suppose that there exists a decision rule with finite risk. Then, the Bayes decision rule with respect to the weighted L_2 loss is the posterior mean

$$\delta_B(X) = E[\theta \mid X = x],$$

where W does not depend on θ.

Find Bayes Decision: Example

Example

Consider the L_2 loss.

1. Let X_1, ..., X_n be an iid sample from $\text{Bernoulli}(\theta)$. Suppose that $\theta \sim \text{Beta}(a, b)$. Find the Bayes decision rule.
2. Let X_1, ..., X_n be an iid sample from $N(\theta, 1)$. Suppose that $\theta \sim N\left(\mu_0, \sigma_0^2\right)$. Find the Bayes decision rule.

Absolute Error Loss

For $k_1 > 0$ and $k_2 > 0$, define the absolute error loss

$$L_{k_1,k_2}(\theta, d) = \begin{cases} k_2(\theta - d), & \text{if } \theta > d, \\ k_1(d - \theta), & \text{if } \theta \le d. \end{cases}$$

If $k_1 = k_2$, such loss reduces to the L_1 loss.

Theorem

Suppose that there exists a decision rule with finite risk. Then, the Bayes decision rule δ_B with respect to the absolute error loss is the $k_2/(k_1+k_2)$ fractile of the posterior distribution, i.e.,

$$P(\theta \le \delta_B(x) \mid x) = \frac{k_2}{k_1+k_2},$$

where k_1 and k_2 do not depend on θ. In particular, if $k_1 = k_2$, the Bayes rule is the posterior median.

Find Bayes Decision: Example

Example

Consider the L_1 loss.

1. Let X_1, ..., X_n be an iid sample from $\text{Bernoulli}\,(\theta)$. Suppose that $\theta \sim \text{Beta}\,(a, b)$. Find the Bayes decision rule.
2. Let X_1, ..., X_n be an iid sample from $N\,(\theta, 1)$. Suppose that $\theta \sim N\left(\mu_0, \sigma_0^2\right)$. Find the Bayes decision rule.

Prediction

Suppose that we want to predict a future observation, possibly from the conditional distribution $f(z \mid x, \theta)$. Let $L_{\text{pred}}(z, d)$ by the prediction error of predicting z by $d \in \mathcal{D}$.

- We can define the loss function as

$$L(\theta, d) = \int L_{\text{pred}}(z, d) f(z \mid x, \theta)\, dz.$$

- The integrated risk satisfies

$$\begin{aligned} \mathrm{E}\left[L_{\text{pred}}(z, \delta)\right] &= \int\int\int L_{\text{pred}}(z, \delta) f(z \mid x, \theta) \pi(\theta \mid x) m(x)\, dz dx d\theta \\ &= \int \left[\int \underbrace{\int L_{\text{pred}}(z, \delta) f(z \mid x, \theta)\, dz}_{= L(\theta, \delta)} \pi(\theta \mid x)\, d\theta \right] m(x)\, dx. \end{aligned}$$

Bayes Predictor

The Bayes predictor is the Bayesian decision rule that minimizes $\mathrm{E}\left[L_{\text{pred}}(z,\delta)\right]$.

- The posterior risk for prediction is

$$\begin{aligned}\int L(\theta, d)\, \pi(\theta \mid x)\, d\theta &= \int \left[\int L_{\text{pred}}(z, d)\, f(z \mid x, \theta)\, dz\right] \pi(\theta \mid x)\, d\theta \\ &= \int L_{\text{pred}}(z, d)\, f(z \mid x)\, dz,\end{aligned}$$

  where $f(z \mid x)$ is the density of the predictive distribution.
- Thus, $\delta(x)$ minimizing the posterior risk $\mathrm{E}\left[L_{\text{pred}}(z,\delta) \mid X = x\right]$ is the Bayes predictor.

L_2 Loss and L_1 Loss

Applying a previous theorem to the prediction case, we obtain the following Bayes predictors.

Theorem

Suppose that there exists a predictor with finite posterior risk.

1. *The Bayes predictor with respect to the weighted L_2 loss $L_{pred}(z, d) = (z-d)^T W (z-d)$ is the mean of the predictive distribution $E[Z \mid X = x]$, where W does not depend on θ.*
2. *The Bayes predictor with respect to the L_1 loss $L_{pred}(z, d) = |z-d|$ is the median of the predictive distribution.*

Find Bayes Predictor: Example

Example

Let Y_1, ..., Y_n be an iid sample from $N(\theta, 1)$. Suppose that $\theta \sim N\left(\mu_0, \sigma_0^2\right)$. We want to predict an iid future observation $Z = Y_{n+1}$.

1. Find the predictive distribution.
2. Find the Bayes predictor under the L_2 loss.
3. Find the Bayes predictor under the L_1 loss.

$0-1$ Loss

Suppose that we are interested in a testing problem such that

$$\Theta = \Theta_0 \cup \Theta_1.$$

A nonrandomized test for a hypothesis is a statistic $\delta(X)$ taking values in $\{0,1\}$, where X is our data.

- $\delta = 1$ means that we reject H_0 and $\delta = 0$ means that we cannot reject H_0.
- Our decision space is $\mathcal{D} = \{0,1\}$.

We can define the $0-1$ loss by

$$L(\theta, d) = \begin{cases} 0, & \text{if } d = 0 \text{ and } \theta \in \Theta_0, \\ 0, & \text{if } d = 1 \text{ and } \theta \in \Theta_1, \\ 1, & \text{if } d = 0 \text{ and } \theta \in \Theta_1, \\ 1, & \text{if } d = 1 \text{ and } \theta \in \Theta_0, \end{cases} = \begin{cases} d, & \text{if } \theta \in \Theta_0, \\ 1-d, & \text{if } \theta \in \Theta_1. \end{cases}$$

Risk of 0-1 Loss

The frequentist risk is

$$\begin{aligned} R(\theta, \delta) &= \int L(\theta, \delta(x)) f(x \mid \theta) \, dx \\ &= \begin{cases} \mathrm{P}(\delta(X) = 1), & \text{if } \theta \in \Theta_0, \text{ (just Type I Error probablity)} \\ \mathrm{P}(\delta(X) = 0), & \text{if } \theta \in \Theta_1. \text{ (just Type II Error probablity)} \end{cases} \end{aligned}$$

The Bayes decision rule is

$$\delta(x) = \begin{cases} 1, & \text{if } \mathrm{P}(\theta \in \Theta_0 \mid x) < \mathrm{P}(\theta \in \Theta_1 \mid x), \\ 0, & \text{if } \mathrm{P}(\theta \in \Theta_0 \mid x) \geq \mathrm{P}(\theta \in \Theta_1 \mid x), \end{cases}$$

if $\mathrm{P}(\theta \in \Theta_0 \mid x) \in (0, 1)$.

Loss for Distributions

Suppose that we want to find a distribution that fits the data well but we are less interested in the parameters themselves.

- Kullback-Leibler divergence (aka entropy loss):

$$L_{\mathrm{KL}} = \int \log\left(\frac{f(x \mid \theta)}{f(x \mid d)}\right) f(x \mid \theta)\, dx,$$

  where the truth is $f(x \mid \theta)$ and the decision is $f(x \mid d)$.

- Squared Hellinger distance:

$$\begin{aligned} L_{\mathrm{H}} &= \frac{1}{2}\int \left(\sqrt{\frac{f(x \mid d)}{f(x \mid \theta)}} - 1\right)^2 f(x \mid \theta)\, dx \\ &= 1 - \int \sqrt{f(x \mid d)\, f(x \mid \theta)}\, dx. \end{aligned}$$

Admissible Decision

Definition

A decision rule δ_0 is called inadmissible if there exits a decision rule δ_1 such that

$$\begin{aligned} R(\theta, \delta_0) &\geq R(\theta, \delta_1), \text{ for all } \theta \in \Theta, \\ R(\theta, \delta_0) &> R(\theta, \delta_1), \text{ for some } \theta \in \Theta. \end{aligned}$$

We say that δ_1 dominates δ_0. Otherwise, the decision rule δ_0 is called admissible.

- If $R(\theta, \delta_0) \geq R(\theta, \delta_1)$ for all θ, then the decision rule δ_0 is better than δ_1.
- If δ_0 is inadmissible, then δ_0 is uniformly dominated by another decision rule δ_1.

Admissible Decision: Example

Let X_1, ..., X_n be independent random variables where $X_i \sim N(\theta_i, 1)$. The parameter is $\theta = \begin{bmatrix} \theta_1 & \cdots & \theta_n \end{bmatrix}^T \in \mathbb{R}^n$.

- An unbiased estimator of θ is $\delta_0(X) = X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T$.
- The James-Stein estimator is

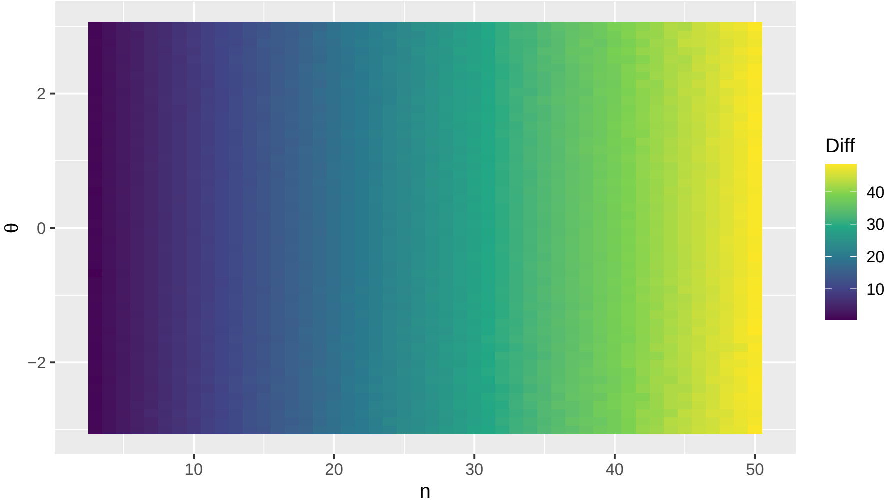
$$\delta_1(x) = \left(1 - \frac{n-2}{x^T x}\right) x.$$

If we consider the L_2 loss, then the difference in the risk satisfies

$$\mathrm{E}\left[L(\theta, \delta_0(X)) \mid \theta\right] - \mathrm{E}\left[L(\theta, \delta_1(X)) \mid \theta\right] \geq \frac{(n-2)^2}{n-2+\theta^T\theta} > 0,$$

for all θ.

Admissible Decision: Example



Minimax Decision Rule

Definition

A decision rule is minimax if it minimizes the maximum risk as

$$\inf_{d \in \mathcal{D}} \left[\sup_{\theta \in \Theta} R\left(\theta, d\right) \right] = \inf_{d \in \mathcal{D}} \left[\sup_{\theta \in \Theta} \mathrm{E}\left[L\left(\theta, d\left(X\right)\right) \mid \theta \right] \right].$$

Example

Suppose $X \mid \theta$ follows a 5-category multinomial distribution and $\theta \in \Theta = \{1, 2, 3\}$ indicates which distribution it is. The candidate distributions are

	x				
θ	1	2	3	4	5
1	0	0.05	0.05	0.8	0.1
2	0.05	0.05	0.8	0.1	0
3	0.9	0.05	0.05	0	0

Find Minimax Decision Rule: Example (Contd.)

Example

Suppose that our decision space $\mathcal{D} = \Theta$. Consider

Our decision rule					
	Observed x				
δ	1	2	3	4	5
δ_1	$d = 3$	3	2	2	1
δ_2	3	2	2	1	1
δ_3	1	1	1	1	1

Loss function			
	Decision d		
θ	1	2	3
1	$L(\theta, d) = 0$	0.8	1
2	0.3	0	0.8
3	0.3	0.1	0

Find the minimax decision rule.

Minimax and Admissible

Theorem (Relation between minimax rule and admissible rule)

1. *If there exists a unique minimax decision rule, then it is also admissible.*
2. *If δ is admissible and has constant risk, then δ is minimax.*
3. *Suppose that $\mathcal{D}$ is convex and, for all $\theta \in \Theta$, the loss function $L(\theta, \cdot)$ is strictly convex. If δ_0 is admissible and has constant risk, then δ_0 is unique minimax.*

Why Bayesian? 1: Bayes is Admissible

Theorem

The Bayes decision rule is admissible if either set of the following conditions hold.

1. $\pi(\theta) > 0$ *for all* $\theta \in \Theta$, $R(\theta, \delta)$ *is continuous in* θ *for all* δ, *and*

$$\inf_{\delta \in \mathcal{D}} \int R(\theta, \delta)\, \pi(\theta)\, d\theta \quad < \quad \infty.$$

2. *The Bayes decision rule is unique.*
3. $\mathcal{D}$ *is convex, the loss function* $L(\theta, \cdot)$ *is strictly convex for all* $\theta \in \Theta$, *and*

$$\inf_{\delta \in \mathcal{D}} \int R(\theta, \delta)\, \pi(\theta)\, d\theta \quad < \quad \infty.$$

Why Bayesian? 1: Bayes is Admissible

We can use the previous theorem to show an estimator is admissible.

Example

Let $X \sim N(\mu, 1)$ and the prior $\pi(\mu) = 1$. The parameter of interest is

$$\theta = 1(\mu \leq 0).$$

Consider a L_2 loss. Find the Bayes estimator of θ.

Blyth Theorem

Theorem

Let Θ be an open set. Suppose that the set of decision rules with continuous $R(\theta, d)$ in θ forms a class $\mathcal{C}$ such that for any $d' \notin \mathcal{C}$ we can find a $d \in \mathcal{C}$ such that d dominates d'. Let δ be an estimator such that $R(\theta, \delta)$ is continuous of θ. Let $\{\pi_n\}$ be a sequence of priors such that

1. $\int R(\theta, \delta)\, \pi_n(\theta)\, d\theta < \infty$ *for all* n,
2. *for every nonemptry open set* $\Theta_0 \in \Theta$, *there exist constants* $B > 0$ *and* N *such that*

$$\int_{\Theta_0} \pi_n(\theta)\, d\theta \geq B, \text{ for all } n \geq N,$$

3. $\int R(\theta, \delta)\, \pi_n(\theta)\, d\theta - \int R(\theta, \delta_n)\, \pi_n(\theta)\, d\theta \to 0$ *as* $n \to \infty$, *where* δ_n *is the Bayes rule under the prior* π_n.

Then, δ is admissible.

Limit of Bayes Rules

We have shown that the Bayes decision rule is admissible under some assumption. The Blyth theorem says that the admissible decision can be obtained such that

$$\lim_{n\to\infty} \int R(\theta,\delta)\,\pi_n(\theta)\,d\theta - \int R(\theta,\delta_n)\,\pi_n(\theta)\,d\theta \quad = \quad 0.$$

We can in fact claim that every admissible estimator is either a Bayes estimator or a limit of Bayes estimators as

$$\lim_{n\to\infty} \delta_n(x) \quad = \quad \delta_B(x)\,, \text{ almost everywhere,}$$

under quite mild assumptions (e.g., $f(x \mid \theta) > 0$ for any $(x,\theta) \in \mathcal{X} \times \Theta$, $L(\theta,d)$ is continuous and strictly convex in d for every θ, among others).

Why Bayesian? 2: Bayes is Minimax

Definition

A prior distribution π is least favorable if

$$\int R(\theta, \delta)\, \pi(\theta)\, d\theta \quad \geq \quad \int R(\theta, \delta)\, \pi'(\theta)\, d\theta$$

for all prior distributions π'.

Theorem

Let δ_B be the Bayes decision rule with respect to the prior $\pi(\theta)$. Suppose that

$$\int R(\theta, \delta_B)\, \pi(\theta)\, d\theta \quad = \quad \sup_{\theta} R(\theta, \delta_B).$$

Then, δ_B is minimax and $\pi(\theta)$ is least favorable. Further, if δ_B is the unique Bayes decision rule with respect to the prior $\pi(\theta)$, then it is the unique minimax estimator.

Bayes is Minimax: A Corollary

Corollary

Let δ_B be the Bayes decision rule with respect to the proper prior $\pi(\theta)$. If δ_B has constant (frequentist) risk, then it is minimax.

Example

Let X_1, ..., X_n be an iid sample from $\text{Bernoulli}(\theta)$. Suppose that $\theta \sim \text{Beta}(a, b)$. Find the minimax estimator of θ.

Bayes is Minimax: Another Corollary

Theorem

Suppose that δ_B is a Bayes decision rule with respect to a proper prior $\pi(\theta)$. If

$$R(\theta, \delta_B) \leq \int R(\theta, \delta_B) \pi(\theta) \, d\theta$$

for every $\theta \in \Theta$, then δ_B is minimax.

Minimax From Limit of Bayes Decision Rules

Theorem

Let $\{\pi_m\}$ be a sequence of proper prior distributions, and δ_m be the Bayes decision rule corresponding to the prior π_m. If δ is an estimator such that

$$\sup_\theta R(\theta, \delta) = \lim_{m \to \infty} \int R(\theta, \delta_m) \pi_m(\theta) \, d\theta.$$

Then δ is minimax.

Example

Let X_1, ..., X_n be iid observations from $N(\theta, \sigma^2)$, where σ^2 is known. Consider the L_2 loss $L(\theta, d) = (\theta - d)^2$. Find the minimax estimator.

Mutual Information

Let $m(x;\pi)$ be the marginal likelihood of x under the prior $\pi(\theta)$. We define the frequentist risk between $f(x \mid \theta)$ and $m(x;\pi)$ as

$$R_n(\theta,\pi) = \mathrm{KL}\left(f(x \mid \theta), m(x;\pi)\right) = \int f(x \mid \theta) \log\left[\frac{f(x \mid \theta)}{m(x;\pi)}\right] dx.$$

The integrated risk is then

$$\begin{aligned} R_n(\pi) &= \int R_n(\theta,\pi)\,\pi(\theta)\,d\theta = \int\int f(x,\theta)\log\left[\frac{f(x,\theta)}{m(x;\pi)\,\pi(\theta)}\right] dx d\theta \\ &= \mathrm{E}\left[\mathrm{KL}\left(\pi(\theta \mid x), \pi(\theta)\right)\right], \end{aligned}$$

which is the same as the mutual information of X and θ, and the expected Kullback-Leiber divergence.

Jeffreys Prior and Minimax

Suppose that some regularity conditions are satisfied, including Θ is a compact set, the Fisher information equals to the negative expected Hessian, among others.

- It has been proved that, among all positive and continuous priors,

$$\sup_{\pi} R_n(\pi) - \inf_{p(x)} \sup_{\theta \in \Theta} \mathrm{KL}\left(f(x \mid \theta), p(x)\right) \quad \to \quad 0.$$

- It has also been proved that the Jeffreys prior $\pi^*(\theta)$ is the unique continuous and positive prior such that

$$\sup_{\pi} R_n(\pi) - R_n(\pi^*) \quad \to \quad 0.$$

Hence, asymptotically, Jeffreys prior maximizes the mutual information, is the least favorable prior, and the integrated risk equals to the minimax risk.

Randomized Decision Rule

For simplicity, all results in our slides are formulated in terms of non-randomized decision rules. For completeness, we need to consider the randomized decision rules such that the action is generated according to some distribution once x has been observed.

Example

The Neyman-Pearson test is a randomized decision

$$\phi(x) = \begin{cases} 1, & \text{if } f_0(x) < kf_1(x), \\ r, & \text{if } f_0(x) = kf_1(x), \\ 0, & \text{if } f_0(x) > kf_1(x). \end{cases}$$

If $f_0(x) = kf_1(x)$, we let $\phi(x) = 1$ with probability r and $\phi(x) = 0$ with probability $1 - r$.

Loss and Risk For Randomized Decision

Since the decision is random, even though x is fixed, we need to take such extra randomness into account. That is, $\delta^*(x,\cdot)$ should be viewed as a density over $\mathcal{D}$ for fixed x.

1. The loss function of a randomized decision rule should be defined as an expected loss

$$L(\theta,\delta^*) = \int_{\mathcal{D}} L(\theta,a)\,\delta^*(x,a)\,da.$$

2. The (frequentist) risk is

$$\begin{aligned} R(\theta,\delta^*) &= \int L(\theta,\delta^*)\,f(x\mid\theta)\,dx \\ &= \int \left[\int_{\mathcal{D}} L(\theta,\delta^*(x,a))\,\delta^*(x,a)\,da\right] f(x\mid\theta)\,dx. \end{aligned}$$

Equivalence

The nonrandomized decision is a special case of the randomized decision rule, where we consider a dirac distribution $\delta^* (x, a) = 1$ on one action a. However, the inclusion of randomized decision rule does not affect the Bayes risk.

Theorem

For every prior π on Θ, the Bayes risk on the set of randomized decision rules is the same as the Bayes risk on the set of nonrandomized decision rules.