

Pattern Recognition II

Bayesian Decision Theory

Michal Haindl

Faculty of Information Technology, KTI
Czech Technical University in Prague
Institute of Information Theory and Automation
Czech Academy of Sciences Prague, Czechia



<http://ro.utia.cz/>

MI/MIE-ROZ 2023-2024



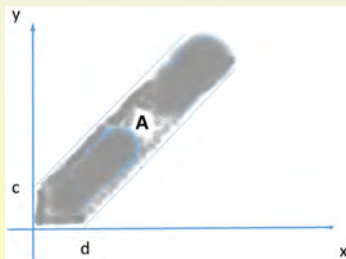
Outline

- 1 Set Theory
- 2 Probability
- 3 Bayesian Decision Theory
 - Posterior Analysis
 - Bayes Minimum Risk Rule
 - Minimax Criterion
 - Rejection Class
 - Bayes Risk Upper Bound
 - Prior Probability Selection
 - Bayes Decision Theory - Discrete



Set Theory

A, B sets e.g. $A = \{\zeta_1, \dots, \zeta_n\}$ $A = \{-c \leq x - y \leq d\}$
 $\mathcal{S}$ space (universe) $A, B \subset \mathcal{S}$, ζ_i object



Set Operations

- equality
 $A = B$ iff $A \subset B \wedge B \subset A$
 $A = \{1, 2, 3\}$ $B = \{3, 2, 1\}$

- sum (union)

$$A \cup B \quad A + B$$

- product (intersection)

$$A \cap B \quad AB$$



Set Operations

• equality

$$A = B \text{ iff } A \subset B \wedge B \subset A$$

• sum (union)

$$A \cup B \quad A + B$$



$$A \cup B = \{1, 2, 3, 4, 5\}$$

$$A = \{1, 2, 3\} \quad B = \{3, 4, 5\}$$

• product (intersection)

$$A \cap B \quad AB$$



Set Operations

• equality

$$A = B \text{ iff } A \subset B \wedge B \subset A$$

• sum (union)

$$A \cup B \quad A + B$$

• product (intersection)

$$A = \{1, 2, 3\} \quad B = \{3, 4, 5\}$$



$$A \cap B = \{3\}$$

$$A \cap B \quad AB$$

A, B appear simultaneously



Set Operations

• equality

$$A = B \text{ iff } A \subset B \wedge B \subset A$$

• sum (union)

$$A \cup B \quad A + B$$

• product (intersection)

$$A \cap B \quad AB$$



• mutually exclusive sets

$$A = \{1, 2, 3\} \quad B = \{4, 5, 6\}$$

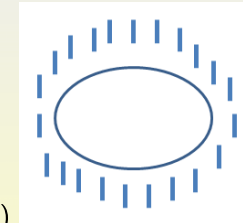
$$\text{iff } A \cap B = \emptyset \quad \text{A, B mutually exclusive}$$

$A_1, \dots, A_n$ are mutually exclusive sets iff

$$A_i \cap A_j = \emptyset \quad \forall i, j, i \neq j$$



Set Operations 2

• complement $A = \{1, 2, 3\} \quad B = \{3, 4, 5\}$ 

$$\bar{A} = \{4, 5\} \quad (S = A \cup B)$$

$$\bar{\bar{A}} = A$$

$$\bar{\emptyset} = S, \quad \bar{S} = \emptyset,$$

• difference

$$A - B, A \setminus B$$

$$A - B = A\bar{B} = A - AB$$



Set Operations 2

- complement $A = \{1, 2, 3\}$ $B = \{3, 4, 5\}$

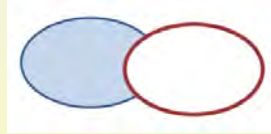
$$\bar{A}$$

$$\bar{\emptyset} = S, \quad \bar{S} = \emptyset,$$

- difference $A \setminus B = \{1, 2\}$

$$A - B, A \setminus B$$

$$A - B = A\bar{B} = A - AB$$



Set Laws

Associative

$$A \cup (B \cap C) = (A \cup B) \cap C$$

$$A \cap (B \cup C) = (A \cap B) \cup C$$

Commutative

$$A \cup B = B \cup A$$

$$A \cap B = B \cap A$$

Distributive

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

Identity

$$A \cup \emptyset = A$$

$$A \cap S = A$$



Complement

Set Laws

Associative

Commutative

Distributive

Identity

Complement

$$A \cup \bar{A} = S$$

$$A \cap \bar{A} = \emptyset$$

Idempotent

$$A \cup A = A$$

$$A \cap A = A$$

Bound

$$A \cup S = S$$

$$A \cap \emptyset = \emptyset$$



Absorption

Set Laws

Associative

Commutative

Distributive

Identity

Complement

Idempotent

Bound

Absorption

$$A \cup (A \cap B) = A$$

$$A \cap (A \cup B) = A$$

De Morgan

$$A = \{1, 2, 3\} \quad B = \{3, 4, 5\}$$

$$\overline{A \cup B} = \bar{A} \cap \bar{B} \quad \emptyset = \{4, 5\} \cap \{1, 2\}$$

$$\overline{A \cap B} = \bar{A} \cup \bar{B} \quad \{1, 2, 4, 5\} = \{4, 5\} \cup \{1, 2\}$$



Probability

definitions:

- Classical (1654 Blaise Pascal, 1601 - 1665 Pierre de Fermat, 1812 Pierre-Simon Laplace)
(till 20th century - A priori definition as a ratio of favourable to total number of alternatives.)
- Axiomatic (measure, 1933 Andrey Nikolaevich Kolmogorov)
- Relative frequency (1936 Richard von Mises)
- Probability as a measure of belief (inductive reasoning)



Probability

definitions:

- Classical (1654 Blaise Pascal, 1601 - 1665 Pierre de Fermat, 1812 Pierre-Simon Laplace)
- Axiomatic (measure, 1933 Andrey Nikolaevich Kolmogorov)
- Relative frequency (1936 Richard von Mises)
- Probability as a measure of belief (inductive reasoning)

Basic rules:

$$A, B \text{ events} \quad 0 \leq P(.) \leq 1$$

$$S \text{ space (certain event)} \quad P(S) = 1,$$

$$\emptyset \text{ impossible event } P(\emptyset) = 0, \quad P(A \setminus B) = P(A) - P(AB)$$

$$A, B \text{ mutually exclusive events iff } A \cap B = \emptyset,$$

$$A \cup \bar{A} = S \quad P(\bar{A}) = 1 - P(A)$$



Classical

$$P(A) = \frac{N_A}{N}$$

N_A no of favourable outcomes

if $A_i \forall i$ are disjoint

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

1654 Blaise Pascal, 1812 Pierre-Simon Laplace Théorie analytique des probabilités



Axiomatic

- 1 $P(A)$ is non-negative: $P(A) \geq 0$
- 2 Probability of certain events equals 1: $P(S) = 1$
- 3 If A and B are mutually exclusive, then:
 $P(A \cup B) = P(A) + P(B)$
 (otherwise $P(A \cup B) = P(A) + P(B) - P(AB)$)

Andrey N. Kolmogorov Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer, Berlin, 1933



Relative Frequency

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}$$

n_A no of event A appearance

1936 Richard von Mises



Relative Frequency

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}$$

n_A no of event A appearance

- ⊖ ∞ no. of experiments
- ⊖ convergence not guaranteed (may fluctuate)
- ⊖ *randomness* what is considered an *identical* experiment

1936 Richard von Mises



Inductive Reasoning

- The probability of future events is based on past events.
- The basis of inference is the Bayes' theorem.



Conditional Probability

given $P(B) > 0$

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Total Probability

$A_i \forall i$ mutually exclusive events $\bigcup_{i=1}^n A_i = S$

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i) = \sum_{i=1}^n P(B A_i)$$

Independent Events

$P(A, B) = P(A)P(B)$ def.

$$P(A|B) = P(A), \quad P(A_1, \dots, A_n) = \prod P(A_i)$$



Conditional Probability

given $P(B) > 0$

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Total Probability

 $A_i \forall i$ mutually exclusive events $\bigcup_{i=1}^n A_i = \mathcal{S}$

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i) = \sum_{i=1}^n P(B A_i)$$

Independent Events

 $P(A, B) = P(A)P(B)$ def.

$$P(A|B) = P(A), \quad P(A_1, \dots, A_n) = \prod P(A_i)$$



Conditional Probability

given $P(B) > 0$

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Total Probability

 $A_i \forall i$ mutually exclusive events $\bigcup_{i=1}^n A_i = \mathcal{S}$

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i) = \sum_{i=1}^n P(B A_i)$$

Independent Events

 $P(A, B) = P(A)P(B)$ def.

$$P(A|B) = P(A), \quad P(A_1, \dots, A_n) = \prod P(A_i)$$



Bayes Theorem - I

$$P(A|B) = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} = \frac{P(B|A)P(A)}{P(B)}$$



1701 - 1761

1742 - elected Fellow of the Royal Society (1 mathematical, 1 theological work)

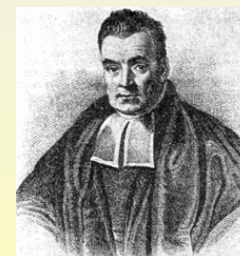
Thomas Bayes: An Essay Toward Solving a Problem in the Doctrine of Chances, 1764 (published posthumously, soon forgotten)



Bayes Theorem - I

Proof:

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(AB)P(A)}{P(B)P(A)} = \frac{P(B|A)P(A)}{P(B)}$$



1701 - 1761

1742 - elected Fellow of the Royal Society (1 mathematical, 1 theological work)

Thomas Bayes: An Essay Toward Solving a Problem in the Doctrine of Chances, 1764 (published posthumously, soon forgotten)



Bayes Theorem - II

$A_i \forall i$ mutually exclusive, positive events $\bigcup_{i=1}^n A_i = \mathcal{S}$

$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{i=1}^n P(B | A_i)P(A_i)}$$



1701 - 1761

1742 - elected Fellow of the Royal Society (1 mathematical, 1 theological work)

Thomas Bayes: An Essay Toward Solving a Problem in the Doctrine of Chances, 1764 (published posthumously, soon forgotten)

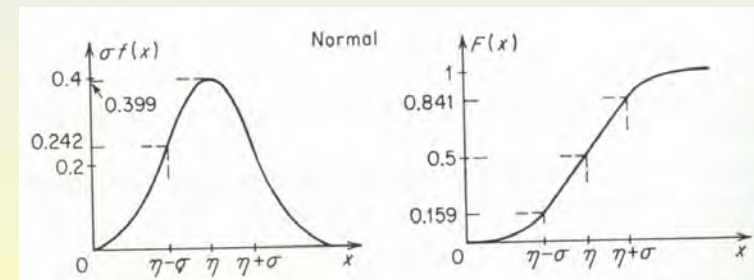


Random Variable

$X : \zeta \rightarrow (\text{real} / \text{complex}) \text{ number}$ ζ experiment outcome

distribution function of the r.v. X

$$F_X(x) = P(X \leq x)$$



properties:

- $F(-\infty) = 0$ $F(+\infty) = 1$
- nondecreasing function of x $F(x_1) \leq F(x_2)$ for $x_1 < x_2$
- continuous from the right $F(x^+) = F(x)$



Random Variable

$X : \zeta \rightarrow (\text{real} / \text{complex}) \text{ number}$ ζ experiment outcome

distribution function of the r.v. X

$$F_X(x) = P(X \leq x)$$

density function

$$f(x) = \frac{dF(x)}{dx}$$

$$F(x) = \int_{-\infty}^x f(t) dt$$

properties:

- nonnegativity $f(x) \geq 0$



Random Variable

$X : \zeta \rightarrow (\text{real} / \text{complex}) \text{ number}$ ζ experiment outcome

distribution function of the r.v. X

$$F_X(x) = P(X \leq x)$$

density function

$$f(x) = \frac{dF(x)}{dx}$$

$$F(x) = \int_{-\infty}^x f(t) dt$$

expected value

$$E\{X\} = \int_{-\infty}^{\infty} xf(x) dx = \int_{-\infty}^{\infty} x dF(x)$$



Random Variable 2

moments

$$\mu_k = E\{X^k\}$$

central moments

$$\mu_k = E\{(X - E\{X\})^k\}$$

 $\mu_2 = \sigma^2$ variance (dispersion) $\sigma = \sqrt{\mu_2}$ standard deviation

median

$$\tilde{x} : F(\tilde{x}) \leq \frac{1}{2} \quad F(\tilde{x} + 0) \geq \frac{1}{2}$$

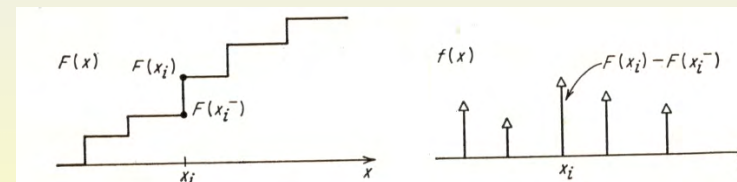
$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd;} \\ (x_{n/2} + x_{n/2+1})/2, & \text{if } n \text{ is even.} \end{cases}$$



Discrete Random Variable

$$F_X(x) = \sum_i P(X = x_i) \quad \forall i : x_i \leq x \quad (\text{staircase form})$$

$$\text{expected value } E\{X\} = \sum_i x_i P(X = x_i)$$



Mixed Random Variable

 $F_X(x)$ discontinuous but not of a staircase form


Discrete Random Variable

$$F_X(x) = \sum_i P(X = x_i) \quad \forall i : x_i \leq x \quad (\text{staircase form})$$

$$\text{expected value } E\{X\} = \sum_i x_i P(X = x_i)$$

Mixed Random Variable

 $F_X(x)$ discontinuous but not of a staircase form


Conditional Distribution

$$F_X(x|z) = P(X \leq x|z) = \frac{P(X \leq x, Z = z)}{P(z)}$$

$$f(x|z) = \frac{dF(x|z)}{dx}$$

$$\text{total probability } (\sum_i^n P(z_i) = 1)$$

$$F_X(x) = \sum_i^n F_X(x|z_i) P(z_i)$$

conditional expected value

$$E\{X|z\} = \int_{-\infty}^{\infty} x f(x|z) dx$$

iterated law of expectation

$$E\{Y\} = E_X\{E\{Y|X\}\}$$



Joint Distribution

$$Y = \{Y_1, \dots, Y_n\}$$

$$F_Y(y) = P(Y_1 \leq y_1, \dots, Y_n \leq y_n)$$

$$f_Y(y) = \frac{\partial^n F_Y(y)}{\partial y_1 \dots \partial y_n}$$

$$E\{Y\} = (E\{Y_1\}, \dots, E\{Y_n\})$$

$$\text{cov}\{Y_i, Y_j\} = E\{(Y_i - E\{Y_i\})(Y_j - E\{Y_j\})\}$$



Marginal Distribution

$$F_{Y_i}(y_i) = F_Y(\infty, \dots, \infty, y_i, \infty, \dots)$$

$$f_{Y_i}(y_i) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \int_{-\infty}^{y_i} f(y_1, \dots, y_{i-1}, t, y_{i+1}, \dots, y_n) dy_1, \dots, dy_{i-1} dt dy_{i+1}, \dots, dy_n$$

$$f_k(y_1, \dots, y_k) = \int_{R^{n-k}} f(y_1, \dots, y_n) dy_{k+1} \dots dy_n$$



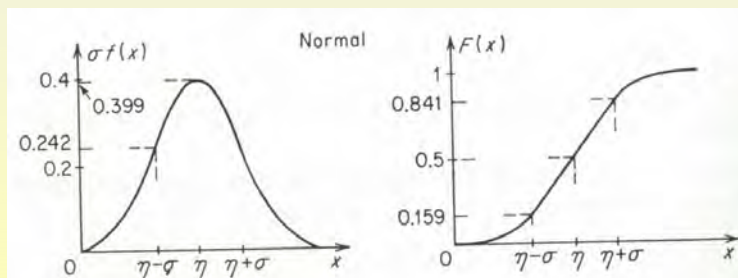
Normal Distribution

Gaussian $\mathcal{N}(\mu, \Sigma)$

$$f(y) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right\}$$

$$E\{y\} = \mu$$

$$E\{(y - \mu)^T (y - \mu)\} = \Sigma$$



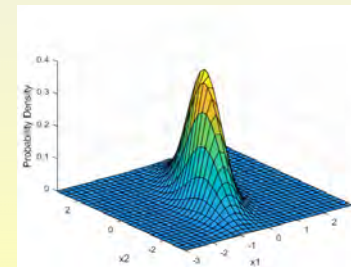
Normal Distribution

Gaussian $\mathcal{N}(\mu, \Sigma)$

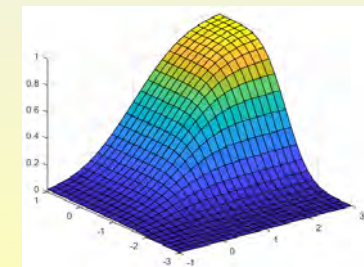
$$f(y) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right\}$$

$$E\{y\} = \mu$$

$$E\{(y - \mu)^T (y - \mu)\} = \Sigma$$



2D density



distribution



Normal Distribution

Gaussian $\mathcal{N}(\mu, \Sigma)$

$$f(y) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y - \mu)^T \Sigma^{-1} (y - \mu) \right\}$$

$$E\{y\} = \mu \quad E\{(y - \mu)^T (y - \mu)\} = \Sigma$$

properties:

- Σ regular, positive definite matrix
- if $\Sigma = \text{diag}\{\Sigma_{1,1}, \dots, \Sigma_{n,n}\}$ then $y_1, \dots, y_n$ independent
- $\forall \tilde{y} \subset y \Rightarrow \tilde{y} \in \mathcal{N}$
- $\forall$ conditional distribution $\mathcal{N}$
- any lin. combination of y_i $\mathcal{N}$



Bayesian Decision - Breast Cancer Example



Breast cancer is the most common type of cancer among middle-aged women in most developed countries.

A woman in her forties has a **positive mammogram**. What is the probability of her breast cancer?



Bayesian Decision - Breast Cancer Example

- the probability that a woman in her forties will have breast cancer

$$p(\text{cancer}) = 0.014$$

- the probability that if a woman has breast cancer, the mammogram will detect it

$$p(\text{mammo} = \oplus | \text{cancer}) = 0.75$$

- the probability that any random woman in her forties will have a positive mammogram

$$p(\text{mammo} = \oplus) = 0.11$$



Bayesian Decision - Breast Cancer Example

- the probability that a woman in her forties will have breast cancer

$$p(\text{cancer}) = 0.014$$

- the probability that if a woman has breast cancer, the mammogram will detect it

$$p(\text{mammo} = \oplus | \text{cancer}) = 0.75$$

- the probability that any random woman in her forties will have a positive mammogram

$$p(\text{mammo} = \oplus) = 0.11$$



Bayesian Decision - Breast Cancer Example

- the probability that a woman in her forties will have breast cancer

$$p(\text{cancer}) = 0.014$$

- the probability that if a woman has breast cancer, the mammogram will detect it

$$p(\text{mammo} = \oplus | \text{cancer}) = 0.75$$

- the probability that any random woman in her forties will have a positive mammogram

$$p(\text{mammo} = \oplus) = 0.11$$



Bayesian Decision - Breast Cancer Example

What is the probability of her breast cancer?

$$\begin{aligned} p(\text{cancer} | \text{mammo} = \oplus) &= \frac{p(\text{mammo} = \oplus | \text{cancer})p(\text{cancer})}{p(\text{mammo} = \oplus)} \\ &= \frac{0.75 \times 0.014}{0.11} = 0.095 \end{aligned}$$

10% probability



Bayesian Decision Theory

assumptions

- a pattern belongs to class ω_i with a priori probability P_i
 $P_i \geq 0, \sum_{i=1}^K P_i = 1$
- pattern X is a random vector governed by a multivariable pdf $p(X|\omega_i)$ when pattern X belongs to class $\omega_i \quad i = 1, \dots, K$

Bayes Inference

- Posterior Analysis
- Bayes Decision (loss function)



Bayesian Decision Theory

assumptions

- a pattern belongs to class ω_i with a priori probability P_i
 $P_i \geq 0, \sum_{i=1}^K P_i = 1$
- pattern X is a random vector governed by a multivariable pdf $p(X|\omega_i)$ when pattern X belongs to class $\omega_i \quad i = 1, \dots, K$

Bayes Inference

- Posterior Analysis
- Bayes Decision (loss function)



Bayesian Inference

♠ no observations, $P(\omega_i) \quad i = 1, \dots, K$
or noninformative observation $X \quad p(X|\omega_i) = p(X|\omega_j) \quad \forall i, j$

decide ω_i if $\omega_i = \arg \max_j \{P(\omega_j)\}$

⊕ $K = 2 \quad P(\omega_1) \gg P(\omega_2)$

roulette $p(\neg \text{winning a number}) = 0.973$

Sportka $p(\neg \text{winning 6 numbers}) = 1 - 0.00000715 = 0.99999285$



Bayesian Inference

♠ no observations, $P(\omega_i) \quad i = 1, \dots, K$
or noninformative observation $X \quad p(X|\omega_i) = p(X|\omega_j) \quad \forall i, j$

decide ω_i if $\omega_i = \arg \max_j \{P(\omega_j)\}$

⊖ $K = 2 \quad P(\omega_1) \doteq P(\omega_2) \quad P(\text{error}) \approx 50\%$

marry to be happy? $p(\text{divorce}) \approx 0.5$

newborn boy $p(\text{boy}) = p(\text{girl}) \approx 0.5$

tossing a fair coin $p(\text{head}) = p(\text{tail}) = 0.5$

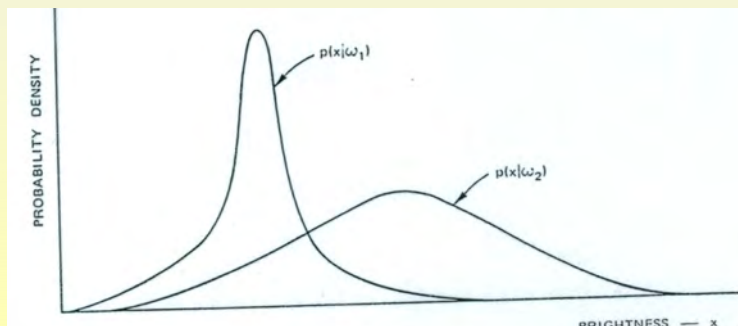


Bayesian Inference 2

♠ equal priors $P(\omega_i) = P(\omega_j) \quad \forall i, j$
decision based entirely on the likelihood of $\omega_i \quad (p(X|\omega_i))$
with respect to X

assign X to ω_i if

$$P(\omega_i|X) = \max_j \{p(X|\omega_j)\}$$



Bayesian Inference 3

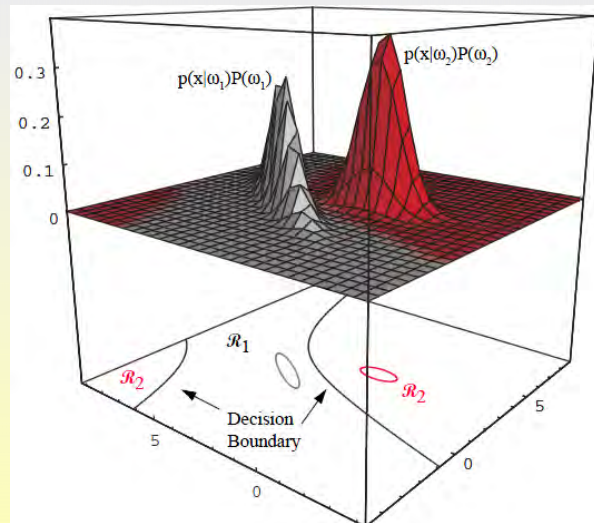
♠ observations $\rightarrow p(X|\omega_i)$ state-conditional p.d. (likelihood)
 $P(\omega_i) \neq P(\omega_j) \quad i = 1, \dots, K$
Bayes rule:

$$P(\omega_i|X) = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}} = \frac{p(X|\omega_i)P(\omega_i)}{p(X)}$$

$$p(X) = \sum_{i=1}^K p(X|\omega_i)P(\omega_i)$$



Bayesian Inference 3



2D

Bayesian Inference 3

$$K = 2$$

$$P(\text{error} | X) = \begin{cases} P(\omega_1 | X), & \text{if chosen } \omega_2; \\ P(\omega_2 | X), & \text{if chosen } \omega_1. \end{cases}$$

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error} | x) p(x) dx$$

$$P^*(\text{error}) = \min\{P(\text{error})\} \rightarrow \forall x \searrow P(\text{error} | x)$$

$$P(\text{error} | x) = \min\{P(\omega_1 | x), P(\omega_2 | x)\}$$



Bayesian Inference 4

assign X to ω_i if

$$P(\omega_i | X) = \max_{\forall j} \{P(\omega_j | X)\} = \max_{\forall j} \{p(X | \omega_j) P(\omega_j)\}$$

$$g_i(X) = \ln p(X | \omega_i) + \ln P(\omega_i)$$



Bayes Minimum Risk Rule

loss function (zero-one / symmetrical / delta)

$$c_{i,j} = c(\omega_i | \omega_j)$$

loss function is the loss incurred for accepting ω_i
when the true state is ω_j

loss matrix

$$\begin{pmatrix} c_{1,1} & c_{1,2} & \dots & c_{1,K} \\ \vdots & \vdots & & \vdots \\ c_{K,1} & c_{K,2} & \dots & c_{K,K} \end{pmatrix} = \begin{pmatrix} 0 & 1 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 0 \end{pmatrix}$$

the conditional risk - the expected loss associated with $\hat{\omega}(X) = \omega_i$

$$R(\omega_i | X) = \sum_{j=1}^K c_{i,j} P(\omega_j | X)$$



Bayes Minimum Risk Rule 2

overall risk - mean loss (average risk)

$$z = \int_S R(\omega_i | X) p(X) dX = \int_S \sum_{j=1}^K c_{i,j} p(X | \omega_j) P(\omega_j) dX$$

Bayes decision rule - minimum error decision rule

$$\begin{aligned} z(\omega_*) &= \min_{\omega_i} \{z\} \\ &= \min_{\omega_i} \left\{ \int_S \sum_{j=1}^K c_{i,j} P(\omega_j | X) p(X) dX \right\} \\ &= \int_S \min_{\omega_i} \left\{ \sum_{j=1}^K c_{i,j} P(\omega_j | X) p(X) \right\} dX \\ &= \int_S R(\omega^* | X) p(X) dX \end{aligned}$$



Bayes Minimum Risk Rule 3

$$\omega_* = \arg \min_{\omega_i} \{R(\omega_i | X)\}$$

optimal decision rule (alternative)

$$R(\omega_i | X) = \sum_{j=1}^K c_{i,j} \frac{p(X | \omega_j) P(\omega_j)}{p(X)}$$

$$\tilde{R}(\omega_i | X) = \sum_{j=1}^K c_{i,j} p(X | \omega_j) P(\omega_j)$$

$$z(\omega_*) = \int_S \tilde{R}(\omega^* | X) dX$$

$$\omega_* = \arg \min_{\omega_i} \{\tilde{R}(\omega_i | X)\}$$



Bayes Minimum Risk Rule 4

- ⊕ minimum conditional risk
- ⊕ minimum average risk → Bayes risk - minimum overall risk $z(\omega_*)$
- ⊕ minimum probability of classification error for zero-one loss function

Bayes risk represents the best possible decision rule beyond which it is impossible to go.

$$g_i(X) = -R(\omega_i | X)$$

$$S_i = \left\{ X \mid R(\omega_i | X) = \min_{j=1, \dots, K} R(\omega_j | X) \right\}$$



Bayes Minimax Criterion

Bayes minimum of the maximum of the error probability $K = 2$

- the task is to design the classifier to perform well over a range of prior probabilities (unpredictable / varies)



Bayes Minimax Criterion

Bayes minimum error:

$$\begin{aligned} e_B &= P(\omega_2) \int_{S_1} p(X|\omega_2) dX + P(\omega_1) \int_{S_2} p(X|\omega_1) dX \\ &= P(\omega_1) \left[\int_{S_2} p(X|\omega_1) dX - \int_{S_1} p(X|\omega_2) dX \right] + \int_{S_1} p(X|\omega_2) dX \end{aligned}$$

for fixed S_1, S_2 ($\tilde{e}_B$) $\rightarrow$ linear function of $P(\omega_1)$

minimise: $\max\{\tilde{e}_B(0), \tilde{e}_B(1)\} \rightarrow$

$$\min \left\{ \max \left\{ \int_{S_1} p(X|\omega_2) dX, \int_{S_2} p(X|\omega_1) dX \right\} \right\}$$



Bayes Minimax Criterion

minimum:

$$\int_{S_1} p(X|\omega_2) dX = \int_{S_2} p(X|\omega_1) dX$$



Bayes Minimax Criterion

$K = 2$

minimize the worst (maximum possible) overall risk for any values of the priors

$$\begin{aligned} z &= \int_{S_1} [c_{1,1}p(X|\omega_1)P(\omega_1) + c_{1,2}p(X|\omega_2)P(\omega_2)] dX \\ &+ \int_{S_2} [c_{2,1}p(X|\omega_1)P(\omega_1) + c_{2,2}p(X|\omega_2)P(\omega_2)] dX \end{aligned}$$



Bayes Minimax Criterion

$K = 2$

$$\begin{aligned} z &= \int_{S_1} [c_{1,1}p(X|\omega_1)P(\omega_1) + c_{1,2}p(X|\omega_2)P(\omega_2)] dX \\ &+ \int_{S_2} [c_{2,1}p(X|\omega_1)P(\omega_1) + c_{2,2}p(X|\omega_2)P(\omega_2)] dX \end{aligned}$$

$$P(\omega_2) = 1 - P(\omega_1)$$

$$1 = \int_{S_1} p(X|\omega_1) dX + \int_{S_2} p(X|\omega_1) dX$$

for fixed $S_1, S_2 \rightarrow z(P(\omega_1))$



Bayes Minimax Criterion

$$K = 2$$

$$\begin{aligned} z(P(\omega_1)) &= c_{2,2} + (c_{1,2} - c_{2,2}) \int_{S_1} p(X|\omega_2) dX \\ &+ P(\omega_1) \left[(c_{1,1} - c_{2,2}) - (c_{2,1} - c_{1,1}) \int_{S_2} p(X|\omega_1) dX \right. \\ &\quad \left. - (c_{1,2} - c_{2,2}) \int_{S_1} p(X|\omega_2) dX \right] \\ 0 &= [\cdot] \quad \text{for minimax solution,} \end{aligned}$$



Bayes Minimax Criterion

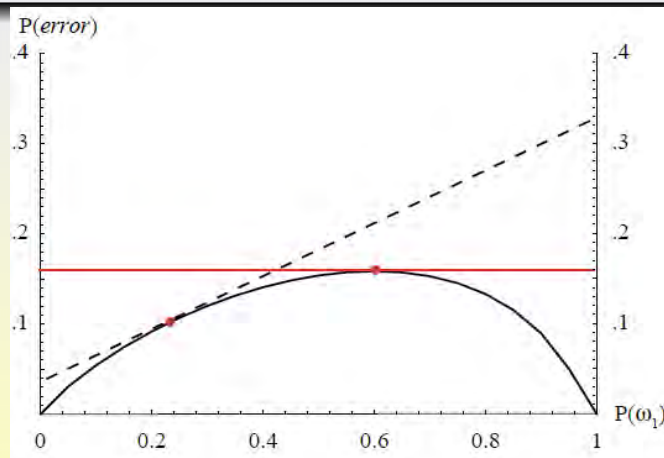
$$K = 2$$

$$\begin{aligned} z(P(\omega_1)) &= c_{2,2} + (c_{1,2} - c_{2,2}) \int_{S_1} p(X|\omega_2) dX \\ &+ P(\omega_1) \left[(c_{1,1} - c_{2,2}) - (c_{2,1} - c_{1,1}) \int_{S_2} p(X|\omega_1) dX \right. \\ &\quad \left. - (c_{1,2} - c_{2,2}) \int_{S_1} p(X|\omega_2) dX \right] \\ 0 &= [\cdot] \quad \text{for minimax solution,} \\ &\quad \text{minimax risk} \end{aligned}$$

$$\begin{aligned} z_{mm} &= c_{2,2} + (c_{1,2} - c_{2,2}) \int_{S_1} p(X|\omega_2) dX \\ &= c_{1,1} + (c_{2,1} - c_{1,1}) \int_{S_2} p(X|\omega_1) dX \end{aligned}$$



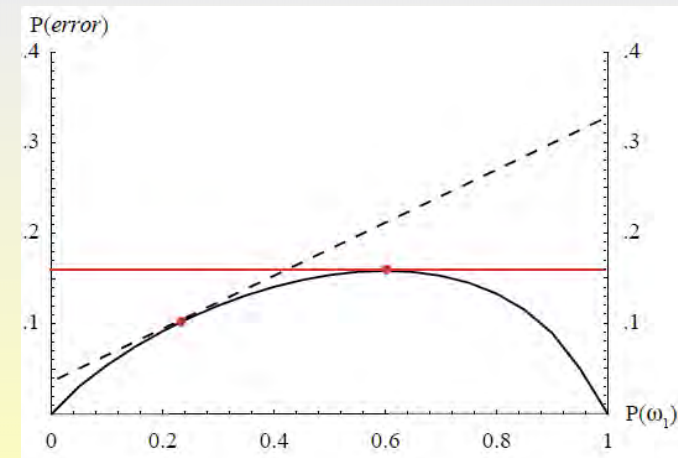
Bayes Minimax Criterion



- solid line minimum Bayes error $P(\omega_1)$
- dashed line for each fixed optimal boundary $P(\omega_1)$ value if the priors are then changed, the probability of error will change as a linear function of $P(\text{error}) = aP(\omega_1) + b$
- solid red line decision boundary for the maximum Bayes error $P(\text{error}) = c$



Bayes Minimax Criterion



optimal Bayesian usually performs better than minimax classifier



Rejection Class

rejection class ω_0
 reject region S_0



Rejection Class

rejection class ω_0
 reject region S_0

assume a constant loss for rejection $c_{0,j} = c_r \quad \forall j = 1, \dots, K$
 c_r rejection threshold

$$c_{i,j} = \begin{cases} c_{i,j} & i \neq 0 \\ c_r & \text{otherwise} \end{cases}$$



Rejection Class

rejection class ω_0
 reject region S_0

$$S_i = \left\{ X \mid R(\omega_i | X) = \min_{j=1, \dots, K} R(\omega_j | X), \quad R(\omega_i | X) \leq c_r \right\}$$

$$S_0 = \left\{ X \mid \min_{j=1, \dots, K} R(\omega_j | X) > c_r \right\}$$

$$R(\omega_0 | X) = \sum_{j=1}^K c_r P(\omega_j | X) = c_r$$

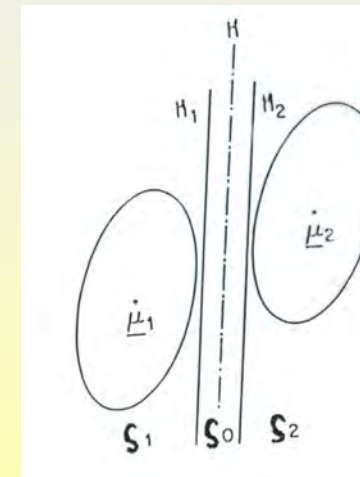


Example

$K = 2$

$X \in \mathcal{N}(\mu_i, \Sigma)$

$c_{i,j} = 1 - \delta_{ij}$



Example

$$K = 2$$

$$X \in \mathcal{N}(\mu_i, \Sigma)$$

$$c_{i,j} = 1 - \delta_{ij}$$

$$\omega^* = \begin{cases} \omega_1 & \text{if } \frac{p(X|\omega_1)}{p(X|\omega_2)} \geq \frac{P(\omega_2)}{P(\omega_1)} \frac{1-c_r}{c_r} \\ \omega_2 & \text{if } \frac{p(X|\omega_1)}{p(X|\omega_2)} \leq \frac{P(\omega_2)}{P(\omega_1)} \frac{c_r}{1-c_r} \\ \omega_0 & \text{otherwise} \end{cases}$$

$$\ln \frac{p(X|\omega_1)}{p(X|\omega_2)} = \left(X - \frac{1}{2}(\mu_1 + \mu_2) \right)^T \Sigma^{-1}(\mu_1 - \mu_2)$$

$$H_1 = \left\{ X : X^T \Sigma^{-1}(\mu_1 - \mu_2) = \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) + \ln \left(\frac{P(\omega_2)}{P(\omega_1)} \frac{1-c_r}{c_r} \right) \right\}$$

Bayes Minimum Risk Decision Rules $K = 2$

- minimum-risk decision rule $K = 2$
assign X to ω_1 if

$$R(\omega_1|X) < R(\omega_2|X)$$

i.e.

$$(c_{2,1} - c_{1,1}) P(\omega_1|X) > (c_{1,2} - c_{2,2}) P(\omega_2|X)$$

$(c_{2,1} - c_{1,1}) > 0, (c_{1,2} - c_{2,2}) > 0 \rightarrow$ the more likely class

- likelihood ratio ($c_{2,1} > c_{1,1}$ assumed) $K = 2$
assign X to ω_1 if

$$\frac{p(X|\omega_1)}{p(X|\omega_2)} > \frac{(c_{1,2} - c_{2,2})P(\omega_2)}{(c_{2,1} - c_{1,1})P(\omega_1)}$$

Bayes Minimum Risk Decision Rules $K > 2$

- minimum probability of error
zero-one (symmetrical) loss function, mean loss $z \rightarrow P(\text{error})$

$$c_{i,j} = \begin{cases} 0 & \text{if } i=j \\ 1 & \text{otherwise} \end{cases} \quad \begin{pmatrix} 0 & 1 & \dots & 1 \\ \vdots & & & \vdots \\ 1 & 1 & \dots & 0 \end{pmatrix}$$

$$R(\omega_i|X) = \sum_{j \neq i} P(\omega_j|X) = 1 - P(\omega_i|X) \quad i = 1, \dots, K$$

conditional risk reduces to the conditional probability of classification error

minimum conditional error probability

$$e^*(X) = \min_{i=1, \dots, K} \{R(\omega_i|X)\} = 1 - \max_{i=1, \dots, K} \{P(\omega_i|X)\}$$

Bayes Minimum Risk Decision Rules $K > 2$

Bayes rule

$$\omega^*(X) = \begin{cases} \omega_i & \text{if } e^*(X) = R(\omega_i^*|X) \leq c_r \\ \omega_0 & \text{if } e^*(X) > c_r \end{cases}$$

or

$$\omega^*(X) = \begin{cases} \omega_i & \text{if } P(\omega_i|X) = \max_{j=1, \dots, K} P(\omega_j|X) \geq 1 - c_r \\ \omega_0 & \text{if } \max_{j=1, \dots, K} P(\omega_j|X) < 1 - c_r \end{cases}$$

$$0 \leq 1 - \max_{j=1, \dots, K} P(\omega_j|X) \leq \frac{K-1}{K}$$

thus

$$0 \leq c_r \leq \frac{K-1}{K}$$



Bayes Minimum Risk Decision Rules - Acceptance / Reject Rate

acceptance region $S_a = \bigcup_{j=1}^K S_j \quad j \neq 0$

reject region S_0

acceptance probability for a pattern X for the decision $\omega^*(X) = \omega_i$

$$A^i = \int_{S_i} p(X) dX$$

the overall acceptance rate (acceptance probability)

$$A = \sum_{j=1}^K A^j$$

reject rate

$$R = \int_{S_0} p(X) dX$$



Bayes Minimum Risk Decision Rules - Classification Error

the classification error probability for $\omega^*(X) = \omega_i$

$$P^i(\text{error}) = \int_{S_i} e^*(X) p(X) dX = 1 - \int_{S_i} \max_j p(X | \omega_j) P(\omega_j) dX$$

the overall probability of error (error rate)

$$P(\text{error}) = \sum_{j=1}^K P^j(\text{error})$$

Bayes rule for minimum error rate

assign X to ω_i if

$$P(\omega_i | X) = \max_j \{P(\omega_j | X)\}$$

$$g_i(X) = P(\omega_i | X)$$

Bayes Minimum Risk Decision Rules - $K = 2$

$$g(X) = P(\omega_1 | X) - P(\omega_2 | X) \quad K = 2$$

$$g(X) = \log \frac{p(X | \omega_1)}{p(X | \omega_2)} + \log \frac{P(\omega_1)}{P(\omega_2)} \quad K = 2$$

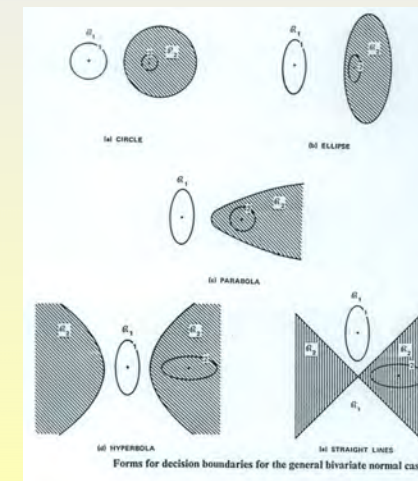
$$\hat{\omega}(X) = \begin{cases} \omega_1 & \text{if } P(\omega_1 | X) \geq 1 - c_r \\ \omega_2 & \text{if } P(\omega_2 | X) \geq 1 - c_r \\ \omega_0 & \text{otherwise} \end{cases}$$

$$0 \leq c_r \leq \frac{1}{2}$$



Bayes Minimum Risk Decision Rules - Bivariate Normal Data

$$g(X) = P(\omega_1 | X) - P(\omega_2 | X) \quad K = 2$$

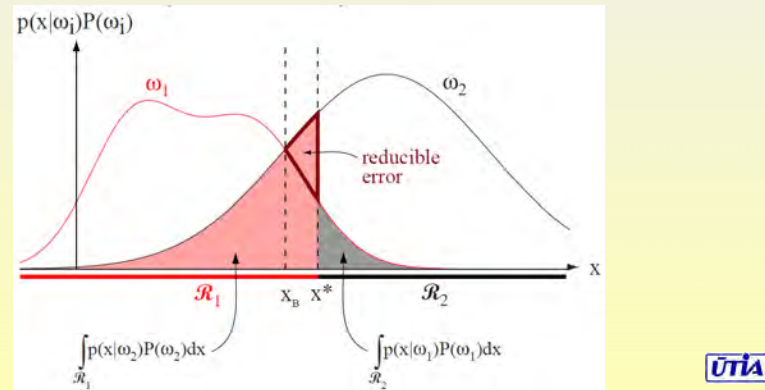


Bayes Minimum Risk Decision Rules - Probability of Error

probability of error

- $K = 2$

$$P(\text{error}|X) = \begin{cases} P(\omega_1|X) & \text{if we decide } \omega_2 \\ P(\omega_2|X) & \text{if we decide } \omega_1 \end{cases}$$



Bayes Minimum Risk Decision Rules - Probability of Error

average probability of error

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, X) dX = \int_{-\infty}^{\infty} P(\text{error}|X) p(X) dX$$

 $K = 2$

$$P(\text{error}) = P(X \in S_2, \omega_1) + P(X \in S_1, \omega_2) = P(X \in S_2|\omega_1)P(\omega_1) +$$

$$P(X \in S_1|\omega_2)P(\omega_2) = \int_{S_2} p(X|\omega_1)P(\omega_1) dX + \int_{S_1} p(X|\omega_2)P(\omega_2) dX$$

$$P(\text{error}) = P(\omega_1) - \int_{S_1} (P(\omega_1|X) - P(\omega_2|X)) p(X) dX$$

minimized error iff S_1, S_2 chosen to be:

$$S_1 : P(\omega_1|X) > P(\omega_2|X)$$

$$S_2 : P(\omega_2|X) > P(\omega_1|X)$$

Bayes Minimum Risk Decision Rules - Probability of Error

average probability of error

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, X) dX = \int_{-\infty}^{\infty} P(\text{error}|X) p(X) dX$$

 $K = 2$

$$P(\text{error}) = P(X \in S_2, \omega_1) + P(X \in S_1, \omega_2) = P(X \in S_2|\omega_1)P(\omega_1) +$$

$$P(X \in S_1|\omega_2)P(\omega_2) = \int_{S_2} p(X|\omega_1)P(\omega_1) dX + \int_{S_1} p(X|\omega_2)P(\omega_2) dX$$

 $K > 2$

$$P(\text{correct}) = \sum_{i=1}^K P(X \in S_i, \omega_i) = \sum_{i=1}^K \int_{S_i} p(X|\omega_i)P(\omega_i) dX$$

Bayes Minimum Risk Decision Rules - Probability of Error

average probability of error

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, X) dX = \int_{-\infty}^{\infty} P(\text{error}|X) p(X) dX$$

 $K = 2$

$$P(\text{error}) = P(X \in S_2, \omega_1) + P(X \in S_1, \omega_2) = P(X \in S_2|\omega_1)P(\omega_1) +$$

$$P(X \in S_1|\omega_2)P(\omega_2) = \int_{S_2} p(X|\omega_1)P(\omega_1) dX + \int_{S_1} p(X|\omega_2)P(\omega_2) dX$$

 $K > 2$

$$P(\text{correct}) = \sum_{i=1}^K P(X \in S_i, \omega_i) = \sum_{i=1}^K \int_{S_i} p(X|\omega_i)P(\omega_i) dX$$

$$P(\text{error}) = 1 - P(\text{correct})$$

Bayes Minimum Risk Decision Rules Probability of Error

average probability of error

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, X) dX = \int_{-\infty}^{\infty} P(\text{error}|X) p(X) dX$$

$K = 2$

$$P(\text{error}) = P(X \in S_2, \omega_1) + P(X \in S_1, \omega_2) = P(X \in S_2 | \omega_1) P(\omega_1) +$$

$$P(X \in S_1 | \omega_2) P(\omega_2) = \int_{S_2} p(X | \omega_1) P(\omega_1) dX + \int_{S_1} p(X | \omega_2) P(\omega_2) dX$$

$K > 2$

$$P(\text{correct}) = \sum_{i=1}^K P(X \in S_i, \omega_i) = \sum_{i=1}^K \int_{S_i} p(X | \omega_i) P(\omega_i) dX$$

Bayes decision rule **minimises the probability of error**, no other partitioning can have a smaller $P(\text{error})$



Bayes Risk Upper Bound

$$e^*(X) \leq c_r$$

$$P^i(\text{error}) = \int_{S_i} e^*(X) p(X) dX \leq c_r \int_{S_i} p(X) dX = c_r A^i$$

$$P(\text{error}) = \sum_{i=1}^K P^i(\text{error}) \leq \sum_{i=1}^K c_r A^i = c_r A(c_r) \leq c_r$$

$$P(\text{error}) \leq c_r$$

⊙ the rejection threshold is the upper bound for the error rate



Bayes Risk Upper Bound 2

$$\begin{aligned} z(\omega^*, c_r) &= \int_S R(\omega^* | X) p(X) dX \\ &= \int_{S_a} R(\omega^* | X) p(X) dX + \int_{S_0} c_r p(X) dX \\ &= P^*(\text{error}, c_r) + c_r R^*(c_r) \end{aligned}$$

$$z(\omega^*, c_r) \leq c_r A(c_r) + c_r R^*(c_r) = c_r$$

$$z(\omega^*, c_r) \leq c_r$$

⊙ c_r is the upper bound for the Bayes risk



Prior Probability Selection

Bayes rule prior dependence - possibly very complex

① for fixed S_i the Bayes risk is a linear function of $P(\omega_i)$

$$z = \sum_{i=1}^K P(\omega_i) \int_{S_i} p(X | \omega_i) dX$$

② z dep. on $P(\omega_i)$ implicitly through the acceptance regions S_i

Prior selection

- missing prior information - uniform priors

$$P(\omega_i) = \frac{1}{K} \quad \forall i$$

Bayesian classifier → ML classifier

- prior information

- another classification result
- type of data (e.g. agricultural scene)
- additional information (maps, time, weather reports, norms, etc.)



Prior Probability Selection

Bayes rule prior dependence - possibly very complex

- 1 for fixed S_i the Bayes risk is a linear function of $P(\omega_i)$

$$z = \sum_{i=1}^K P(\omega_i) \int_{S_i} p(X|\omega_i) dX$$

- 2 z dep. on $P(\omega_i)$ implicitly through the acceptance regions S_i

Prior selection

- missing prior information - uniform priors

$$P(\omega_i) = \frac{1}{K} \quad \forall i$$

Bayesian classifier $\rightarrow$ ML classifier

- prior information
 - another classification result
 - type of data (e.g. agricultural scene)
 - additional information (maps, time, weather reports, norms, etc.)



Bayes Decision Theory - Discrete

X discrete, i.e. $X \in \{v_1, \dots, v_n\}$, $\int \rightarrow \sum$

Bayes rule:

$$P(\omega_i|X) = \frac{p(X|\omega_i)P(\omega_i)}{p(X)}$$

$$p(X) = \sum_{i=1}^K p(X|\omega_i)P(\omega_i)$$

$$g_i(X) = \ln p(X|\omega_i) + \ln P(\omega_i)$$

decision rule is unchanged



Bayes Decision Theory - Discrete

$X = [x_1, \dots, x_I]$ independent binary features $x_i \in \{0, 1\}$, $K = 2$

$$p_i = P(x_i = 1|\omega_1)$$

$$q_i = P(x_i = 1|\omega_2)$$

assuming conditional independence:

$$P(X|\omega_1) = \prod_{j=1}^I p_j^{x_j} (1 - p_j)^{1-x_j}$$

$$P(X|\omega_2) = \prod_{j=1}^I q_j^{x_j} (1 - q_j)^{1-x_j}$$



Bayes Decision Theory - Discrete

$X = [x_1, \dots, x_I]$ independent binary features $x_i \in \{0, 1\}$, $K = 2$

$$p_i = P(x_i = 1|\omega_1)$$

$$q_i = P(x_i = 1|\omega_2)$$

$$\frac{P(X|\omega_1)}{P(X|\omega_2)} = \prod_{i=1}^I \left(\frac{p_i}{q_i} \right)^{x_i} \left(\frac{1-p_i}{1-q_i} \right)^{1-x_i}$$



Bayes Decision Theory - Discrete

$X = [x_1, \dots, x_I]$ independent binary features $x_i \in \{0, 1\}$, $K = 2$

$$p_i = P(x_i = 1 | \omega_1) \quad q_i = P(x_i = 1 | \omega_2)$$

linear discriminant function (ω_1 if $g(X) > 0$), decision surface hyperplane

$$g(X) = \sum_{j=1}^I \left[x_j \log \frac{p_j}{q_j} + (1 - x_j) \log \frac{1 - p_j}{1 - q_j} \right] + \log \frac{P(\omega_1)}{P(\omega_2)}$$



Bayes Decision Theory - Discrete 3

$$g(X) = \sum_{j=1}^I w_j x_j + w_0$$

- ① $p_i = q_i \Rightarrow w_i = 0$ x_i gives no information
- ② $p_i > q_i \Rightarrow w_i > 0$ $x_i = 1$ contributes w_i votes for ω_1
- ③ $p_i < q_i \Rightarrow w_i < 0$ $x_i = 1$ contributes w_i votes for ω_2
- ④ $P(\omega_1) \propto$ bias to ω_1



Information

<http://mosaic.utia.cas.cz/>
<http://ro.utia.cz>

