CHAPTER 2 (PART 2): BAYESIAN DECISION THEORY (SECTIONS 2.3-2.5)

Classifiers, Discriminant Functions and Decision Surfaces

Error Bounds

Missing and Noisy Features

Error Probabilities and Integrals

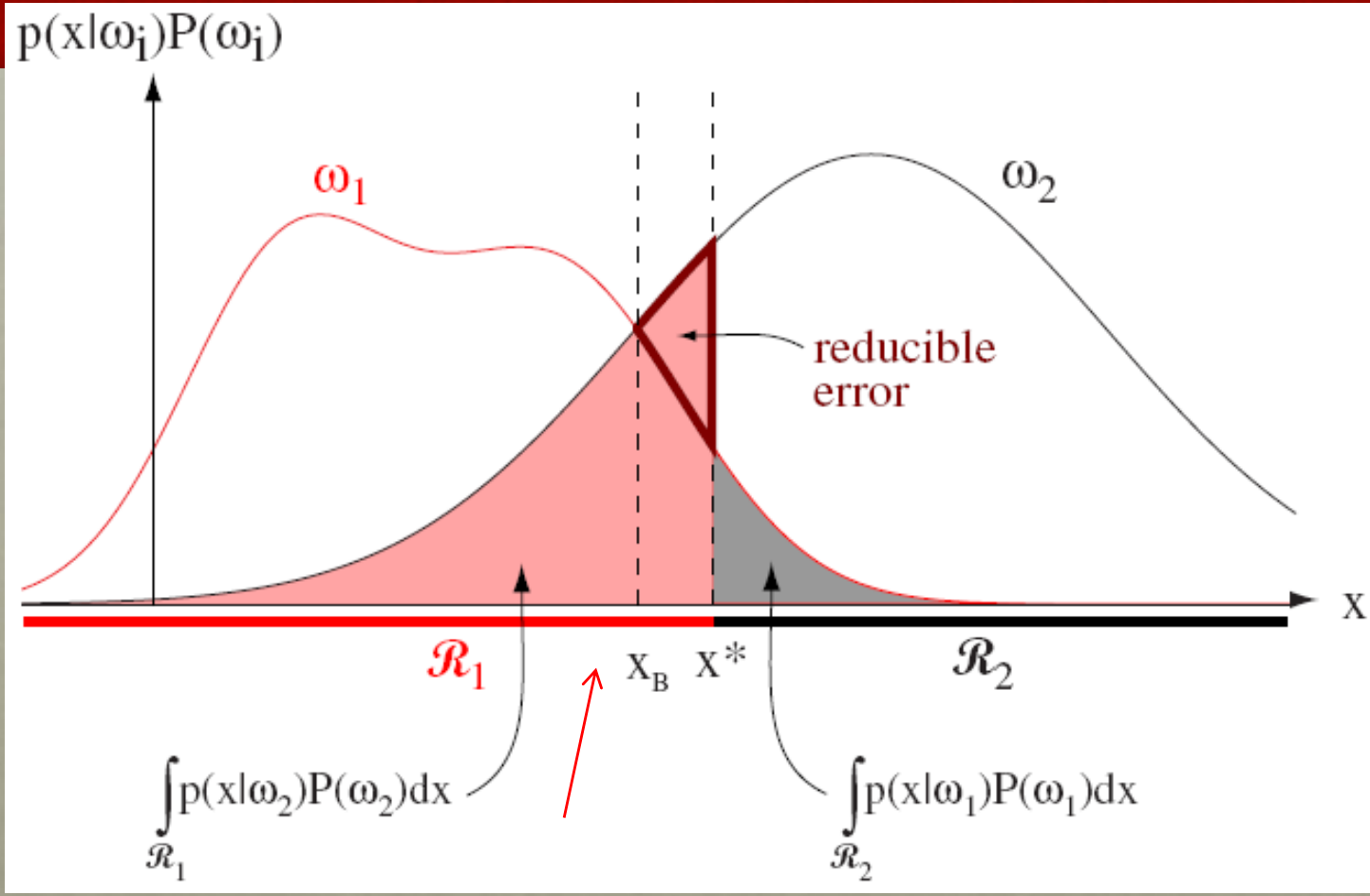


Figure 2.17: Components of the probability of error for equal priors and (non-optimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 68. If the decision boundary is instead at the point of equal posterior probabilities, x_B, then this reducible error is eliminated and the total shaded area is the minimum possible — this is the Bayes decision and gives the Bayes error rate.

Error Probabilities and Integrals

- 2-class problem: There are two types of errors

$$\begin{aligned} P(error) &= P(\mathbf{x} \in \mathcal{R}_2, \omega_1) + P(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\ &= P(\mathbf{x} \in \mathcal{R}_2|\omega_1)P(\omega_1) + P(\mathbf{x} \in \mathcal{R}_1|\omega_2)P(\omega_2) \\ &= \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)P(\omega_1)\ d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)P(\omega_2)\ d\mathbf{x}. \end{aligned}$$

- Multi-class problem
  - Simpler to computer the prob. of being correct (more ways to be wrong than to be right)

$$\begin{aligned} P(correct) &= \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i, \omega_i) \\ &= \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i|\omega_i)P(\omega_i) \\ &= \sum_{i=1}^{c} \int_{\mathcal{R}_i} p(\mathbf{x}|\omega_i)P(\omega_i)\ d\mathbf{x}. \end{aligned}$$

Error Bounds for Normal Densities

- Bayes rule guarantees the lowest error rate but it does not tell us the actual error
- In the 2-category case the general error can be approximated analytically to give us an upper bound on the error
- However, the exact calculation of error for the general Gaussian case (case 3) is extremely difficult

CHERNOFF BOUND

- Inequality:

$$\min[a, b] \leq a^{\beta} b^{1-\beta} \quad \text{for } a, b \geq 0 \text{ and } 0 \leq \beta \leq 1.$$

CHERNOFF BOUND

- Inequality:

$$\min[a, b] \leq a^{\beta} b^{1-\beta} \quad \text{for } a, b \geq 0 \text{ and } 0 \leq \beta \leq 1.$$

$$P(error|x) = \min\left[P(\omega_1|x), P(\omega_2|x)\right].$$

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

CHERNOFF BOUND

- To derive a bound for the error,

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

$$P(error|x) = \min\,[P(\omega_1|x), P(\omega_2|x)].$$

$$\min[a, b] \leq a^{\beta} b^{1-\beta} \quad \text{for } a, b \geq 0 \text{ and } 0 \leq \beta \leq 1.$$

$$P(error) \leq P^{\beta}(\omega_1) P^{1-\beta}(\omega_2) \int p^{\beta}(\mathbf{x}|\omega_1) p^{1-\beta}(\mathbf{x}|\omega_2)\, d\mathbf{x} \qquad \text{for } 0 \leq \beta \leq 1.$$

CHERNOFF BOUND

Assuming conditional prob. are normal,

$$\int p^{\beta}(\mathbf{x}|\omega_1)p^{1-\beta}(\mathbf{x}|\omega_2)\ d\mathbf{x} = e^{-k(\beta)}$$

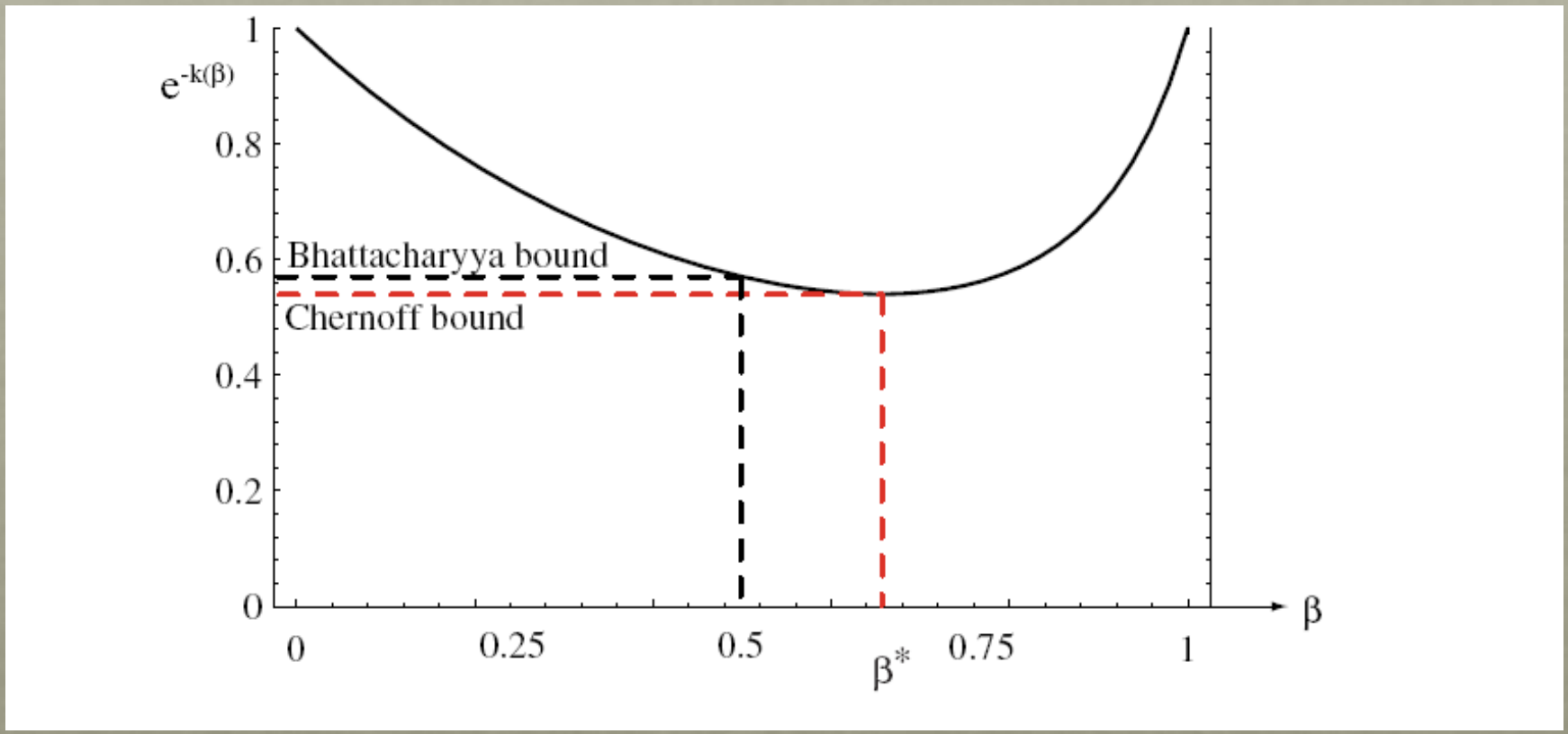
where

$$k(\beta) = \frac{\beta(1-\beta)}{2}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^t[\beta\boldsymbol{\Sigma}_1+(1-\beta)\boldsymbol{\Sigma}_2]^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)+\frac{1}{2}\ln\frac{|\beta\boldsymbol{\Sigma}_1+(1-\beta)\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|^{\beta}|\boldsymbol{\Sigma}_2|^{1-\beta}}.$$

Chernoff bound for P(error) is found by determining the value of β that minimizes exp(-$k(\beta)$)

CHERNOFF BOUND

1D optimization regardless of the dimensionality of the features



BHATTACHARYA BOUND

- Bhattacharyya Bound: Assume $\beta = 1/2$
  - computationally simpler but slightly less tight bound

$$\begin{aligned} P(error) &\leq \sqrt{P(\omega_1)P(\omega_2)} \int \sqrt{p(\mathbf{x}|\omega_1)p(\mathbf{x}|\omega_2)}\, d\mathbf{x} \\ &= \sqrt{P(\omega_1)P(\omega_2)}e^{-k(1/2)}, \end{aligned}$$

$$\begin{aligned} k(1/2) &= 1/8(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t \left[\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right]^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \\ &\quad \frac{1}{2}\ln \frac{\left|\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}}. \end{aligned}$$

When the two covariance matrices are equal, $k(1/2)$ is the same as the Mahalanobis distance between the two means

BHATTACHARYA BOUND

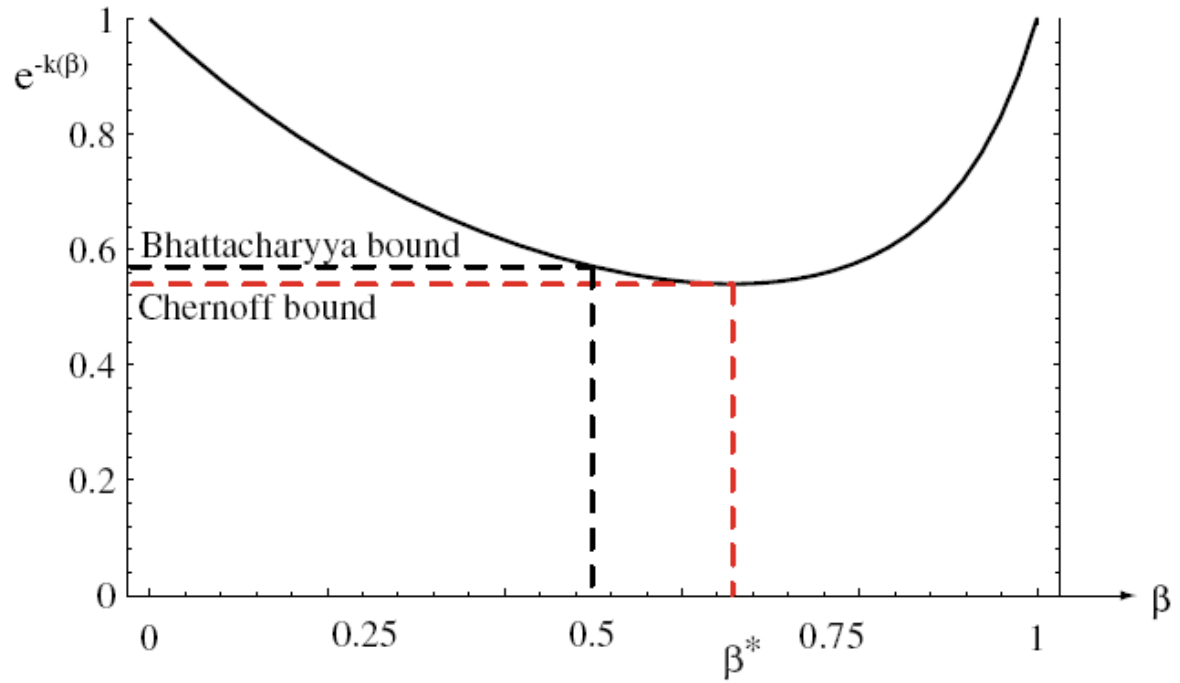


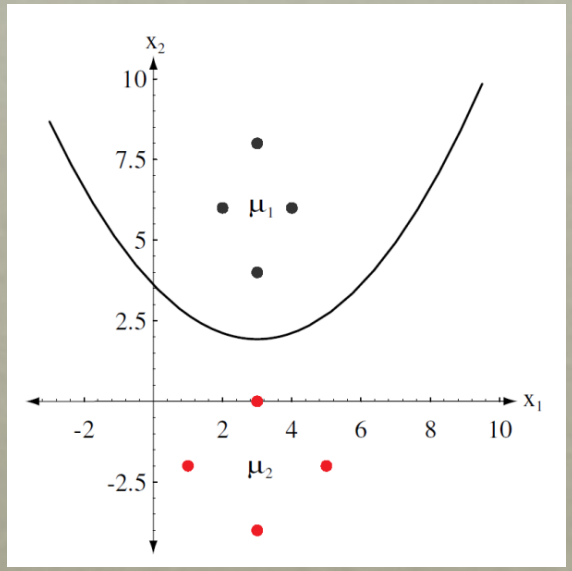
Figure 2.18: The Chernoff error bound is never looser than the Bhattacharyya bound. For this example, the Chernoff bound happens to be at $\beta^* = 0.66$, and is slightly tighter than the Bhattacharyya bound ($\beta = 0.5$).

ERROR BOUNDS FOR GAUSSIAN DISTRIBUTION

Best Chernoff error bound is 0.008190

Bhattacharya error bound is 0.008191

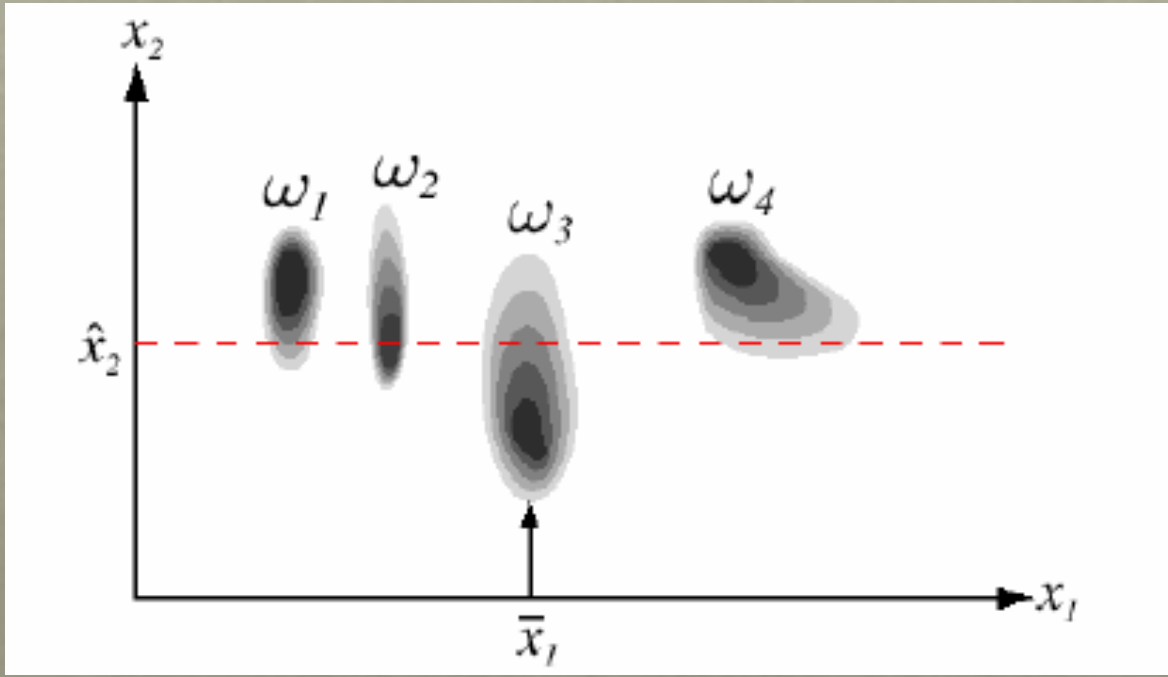
True error using numerical integration = 0.0021



2–category, 2D data

MISSING FEATURES AND NOISY FEATURES

- Features are corrupted by a known noise source
  - Ex: variability of light source may degrade measurement of lightness
- Features are missing
  - Ex: occlusion prevents measurement of length



Four categories with equal priors and class-conditional distributions. If a test point is given for which one features is (*x1*) is missing and the other has value $\overline{x2}$.

We want to classify it as ω*2* since the likelihood p(ω*2*/*x2*) is the highest.

MISSING FEATURES

Good features: *xg*

Bad features (unknown or missing): *xb*

Bayes Discriminant Function:

$$P(\omega_i|\mathbf{x}_g) = \frac{p(\omega_i,\mathbf{x}_g)}{p(\mathbf{x}_g)} = \frac{\int p(\omega_i,\mathbf{x}_g,\mathbf{x}_b)\ d\mathbf{x}_b}{p(\mathbf{x}_g)}$$

$$= \frac{\int P(\omega_i|\mathbf{x}_g,\mathbf{x}_b)p(\mathbf{x}_g,\mathbf{x}_b)\ d\mathbf{x}_b}{p(\mathbf{x}_g)}$$

$$= \frac{\int g_i(\mathbf{x})p(\mathbf{x})\ d\mathbf{x}_b}{\int p(\mathbf{x})\ d\mathbf{x}_b},$$

Marginalize over all values of missing feature

NOISY FEATURES

- Uncorrupted good features: xg
- Bad features (noisy): xb
- Noise model p(xb|xt)
  - xt = True value of the observed value xb
  - Assume if xt were known, xb would be independent of wi and xg

$$P(\omega_i|\mathbf{x}_g,\mathbf{x}_b) = \frac{\int p(\omega_i,\mathbf{x}_g,\mathbf{x}_b,\mathbf{x}_t)\ d\mathbf{x}_t}{p(\mathbf{x}_g,\mathbf{x}_b)}$$

NOISY FEATURES

$$p(\omega_i, \mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) = P(\omega_i|\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)$$

- By the independence assumption, if we know xt, then xb does not provide any additional information about wi

$$P(\omega_i|\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) = P(\omega_i|\mathbf{x}_g, \mathbf{x}_t)$$

- Similarly, $p(\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) = p(\mathbf{x}_b|\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_t)$

$$p(\mathbf{x}_b|\mathbf{x}_g, \mathbf{x}_t) = p(\mathbf{x}_b|\mathbf{x}_t)$$

NOISY FEATURES

- Bayes discriminant function is reduced to:

$$P(\omega_i|\mathbf{x}_g,\mathbf{x}_b) = \frac{\int P(\omega_i|\mathbf{x}_g,\mathbf{x}_t)p(\mathbf{x}_g,\mathbf{x}_t)p(\mathbf{x}_b|\mathbf{x}_t)\ d\mathbf{x}_t}{\int p(\mathbf{x}_g,\mathbf{x}_t)p(\mathbf{x}_b|\mathbf{x}_t)\ d\mathbf{x}_t}$$

$$= \frac{\int g_i(\mathbf{x})p(\mathbf{x})p(\mathbf{x}_b|\mathbf{x}_t)\ d\mathbf{x}_t}{\int p(\mathbf{x})p(\mathbf{x}_b|\mathbf{x}_t)\ d\mathbf{x}_t},$$

Integral is weighted by the noise model

Neyman-Pearson Rule

I. Neyman-Pearson Rule

Let $\Omega = \{\omega_0, \omega_1\}, A = \{a_0, a_1\}$. This rule is related to the following hypothesis testing problem.

H_0 (null hypothesis) : $\mathbf{x}$ comes from a population governed by $p(\mathbf{x} \mid \omega_0)$; ω_0 is the state of nature.

H_1 (alternative hypothesis) : the state of nature is ω_1.

"Classification, Estimation and Pattern recognition" by Young and Calvert

Neyman-Pearson Rule

The N-P decision rule is also called a 'test' which divides the *feature space* (X) into two regions : critical region $C_\delta = \{\mathbf{x} \mid \delta(\mathbf{x}) = a_1\}$ and its complement C_δ^*. A point in C_δ leads to the acceptance of the alternative hypothesis.

There are two types of errors defined below.

1. False alarm : decide H_1 when H_0 is true; also called type I error, denoted by α.

2. False dismissal : decide H_0 when H_1 is true; also called type II error, denoted by β.

This terminology comes from the field of communication theory where we want to detect a message in the presence of noise.

H_0 : a received signal is noise alone.

H_1 : a received signal is message plus noise.

Neyman-Pearson Rule

The false alarm probability $= \int_{C_\delta} p(\mathbf{x} \mid \omega_0) d\mathbf{x}$

The false dismissal probability $= 1 - \int_{C_\delta} p(\mathbf{x} \mid \omega_1) d\mathbf{x}$

These error probabilities can be related to the risk function when a 0-1 loss function is used. That is,

$$L(\omega_k, \alpha_i) = \begin{cases} 0 & \text{if } i = k \\ 1 & \text{if } i \neq k \end{cases}$$

$$\text{Since } R_\delta(\omega_0) = \int L(\omega_0, \delta(\mathbf{x})) p(\mathbf{x} \mid \omega_0) d\mathbf{x}$$

$$\text{Hence } R_\delta(\omega_0) = \int_{C_\delta} p(\mathbf{x} \mid \omega_0) d\mathbf{x} \text{ (false alarm probability)}$$

$$R_\delta(\omega_1) = \int L(\omega_1, \delta(\mathbf{x}) p(\mathbf{x} \mid \omega_1) d\mathbf{x} = \int_{C_{\delta^*}} p(\mathbf{x} \mid \omega_1) d\mathbf{x}$$

$$R_\delta(\omega_1) = 1 - \int_{C_\delta} p(\mathbf{x} \mid \omega_1) d\mathbf{x} \text{ (false dismissal probability)}$$

Assuming that the probability of error of type $I(\alpha)$ is given, then the Neyman-Pearson classifier will minimize β for a fixed α, where

α is called the level or size of the test.

(1-β) is called the power of the test.

Neyman-Pearson Rule

The Neyman-Pearson classifier that minimizes β for a given value of α is a likelihood ratio test with the threshold k_α,

$$Pr\left\{\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_0)} \geq k_\alpha \mid \mathbf{x} \in \omega_0\right\} = \alpha$$

The critical region for the N-P rule is

$$C_\alpha = \left[\mathbf{x} \mid \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_0)} \geq k_\alpha\right].$$

If $k_\alpha = 1$, then the N-P rule is a maximum likelihood rule. The expression of the rule can be simplified by choosing a monotone increasing function g such that $\delta_{NP}(\mathbf{x}) = a_1$ if $g[\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_0)}] \geq g(k_\alpha)$.

J. Neyman-Pearson Lemma

Consider the class of all decision rules having size α. No rule in this class has power larger than the N-P rule at level α. In other words, for a 0-1 loss function, the N-P rule at level α minimizes $R_\delta(\omega_1)$ among all rules for which $R_\delta(\omega_0) \leq \alpha$.

Neyman-Pearson Rule

Example 2.4

Let the two class-conditional densities be univariate Gaussian. We will now determine the critical region and the power of the N-P rule for a given α.

$$p(x \mid \omega_0) \sim N(\mu_0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu_0)^2}{2\sigma^2}}$$

$$p(x \mid \omega_1) \sim N(\mu_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu_1)^2}{2\sigma^2}}$$

$$\text{Thus, } \log\left[\frac{p(x|\omega_1)}{p(x|\omega_0)}\right] = \frac{1}{\sigma^2}\left[(\mu_1 - \mu_0)x - \frac{1}{2}(\mu_1^2 - \mu_0^2)\right],$$

$$\begin{aligned} C_\alpha &= \left[x \middle| \log[\frac{p(x|\omega_1)}{p(x|\omega_0)}] \geq \log k_\alpha\right] \\ &= \left[x \middle| x \geq x_0, \text{ where } x_0 = \frac{\sigma^2 \log k_\alpha}{\mu_1 - \mu_0} + \frac{1}{2}(\mu_1 + \mu_0)\right], \end{aligned}$$

$$\begin{aligned} \alpha &= \int_{C_\alpha} p(x|\omega_0)dx = \int_{x_0}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu_0)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{\pi}} \int_{\frac{x_0-\mu_0}{\sqrt{2}\sigma}}^{\infty} e^{-t^2} dt \\ &= \frac{-1}{\sqrt{\pi}} \int_0^{\frac{x_0-\mu_0}{\sqrt{2}\sigma}} e^{-t^2} dt + \frac{1}{\sqrt{\pi}} \int_0^{\infty} e^{-t^2} dt. \end{aligned}$$

Neyman-Pearson Rule

Therefore, $\alpha = \frac{1}{2} - \frac{1}{2}\text{erf}(\frac{x_0-\mu_0}{\sqrt{2\sigma}})$.

The threshold value x_0 can be written as

$$x_0 = \mu_0 + \sqrt{2}\sigma\text{erf}^{-1}(1-2\alpha).$$

The power of the test becomes

$$1-\beta = \int_{x_0}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu_1)^2}{2\sigma^2}} dx = \frac{1}{2} - \frac{1}{2}\text{erf}(\frac{x_0-\mu_1}{\sqrt{2\sigma}}).$$

The decision boundary is shown in Figure 2.2.

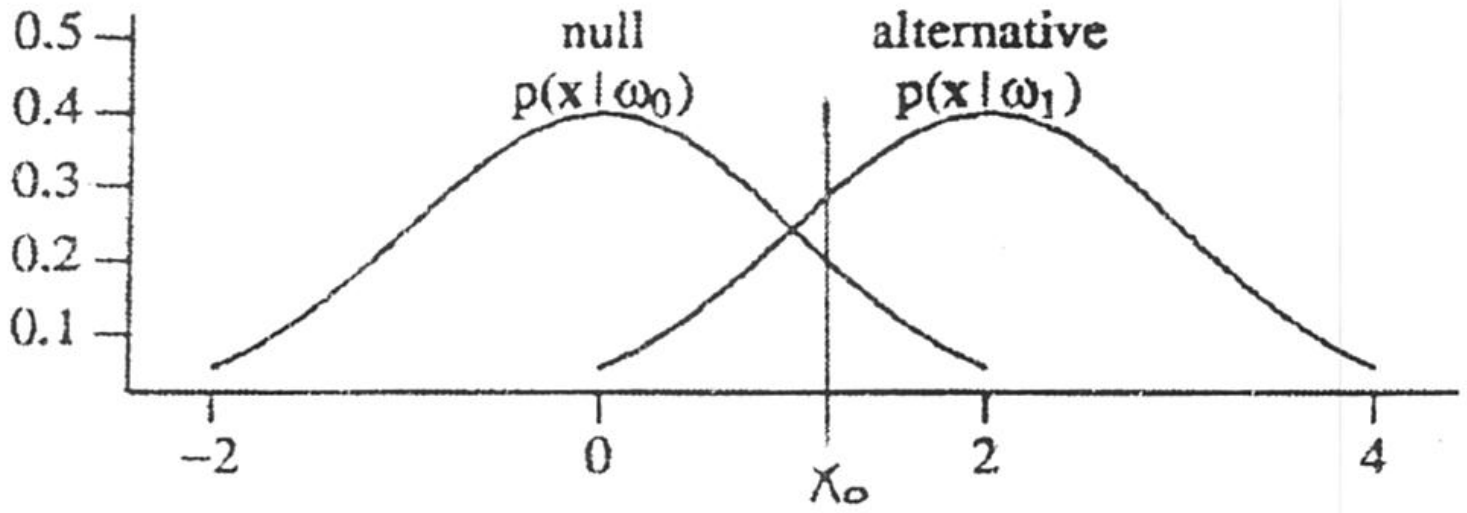
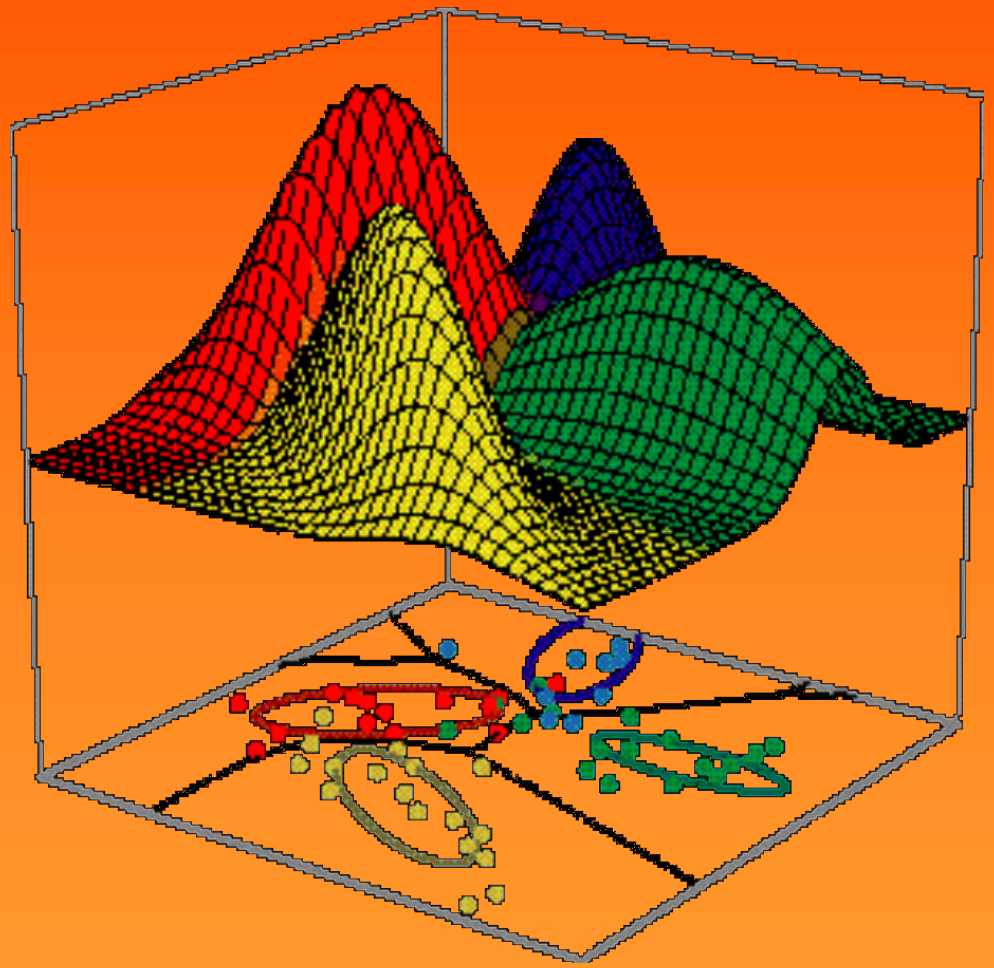


Figure 2.2: N-P rule



Pattern Classification

All materials in these slides were taken from
Pattern Classification (2nd ed) *by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000*
with the permission of the authors and the publisher