Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

MIN-Fakultät
Fachbereich Informatik
Arbeitsbereich SAV/BV (KOGS)

Image Processing 1 (IP1)
Bildverarbeitung 1

Lecture 16 – Decision Theory

Winter Semester 2015/16

Slides: Prof. Bernd Neumann
Silghtly revised by: Dr. Benjamin Seppke & Prof. Siegfried Stiehl

Statistical Decision Theory

Generating decision functions from a statistical characterization of classes (as opposed to a characterization by prototypes)

Advantages:

1. The classification scheme may be designed to satisfy an objective optimality criterion:
   Optimal decisions minimize the probability of error.
2. Statistical descriptions may be much more compact than a collection of prototypes.
3. Some phenomena may only be adequately described using statistics, e.g. noise.

Example: Medical Screening I

Health test based on some measurement x (e.g. ECG evaluation)

It is known that every 10th person is sick (prior probability):

- ω_1 class of healthy people $P(\omega_1) = 9/10$
- ω_2 class of sick people $P(\omega_2) = 1/10$

Task 1: Classify without taking any measurements (to save money)

- **Decision rule 1a:** Classify every 10th person as sick

$$\begin{aligned} P(error) &= P(decide\ sick\ if\ healthy) + P(decide\ healthy\ if\ sick) \\ &= 1/10 \times 9/10 + 9/10 \times 1/10 = 0.18 \end{aligned}$$

- **Decision rule 1b:** Classify all persons as healthy

$$P(error) = P(decide\ healthy\ if\ sick) = 1/10 = 0.1$$

Decision rule 1b is better because it gives lower probability of error

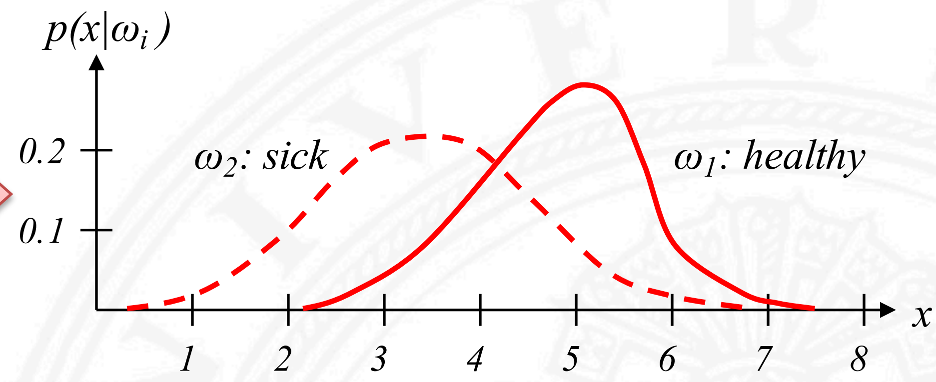
Decision rule 1b is optimal because no other decision rule can give a lower probability of error (try "every n-th" in 1a and minimize over n)

Example: Medical Screening II

Task 2: Classify after taking a measurement x

Assume that the statistics of prototypes are given as $p(x|\omega_i), i = 1, 2$

Person No.	x	indication
•	•	•
•	•	•
134	7.4	neg
135	6.8	neg
136	4.2	pos
137	5.6	neg
138	5.8	pos
139	7.2	neg
•	•	•
•	•	•



$P(e|x) = P(\text{error given } x) = P(\omega \neq \omega'|x) = 1 - P(\omega|x)$

where ω' is the class assigned to x by the decision rule.

$P(e|x)$ is minimized by choosing the class which maximizes $P(\omega|x)$.
Hence $g_i(x) = P(\omega_i|x)$ are discriminant functions.

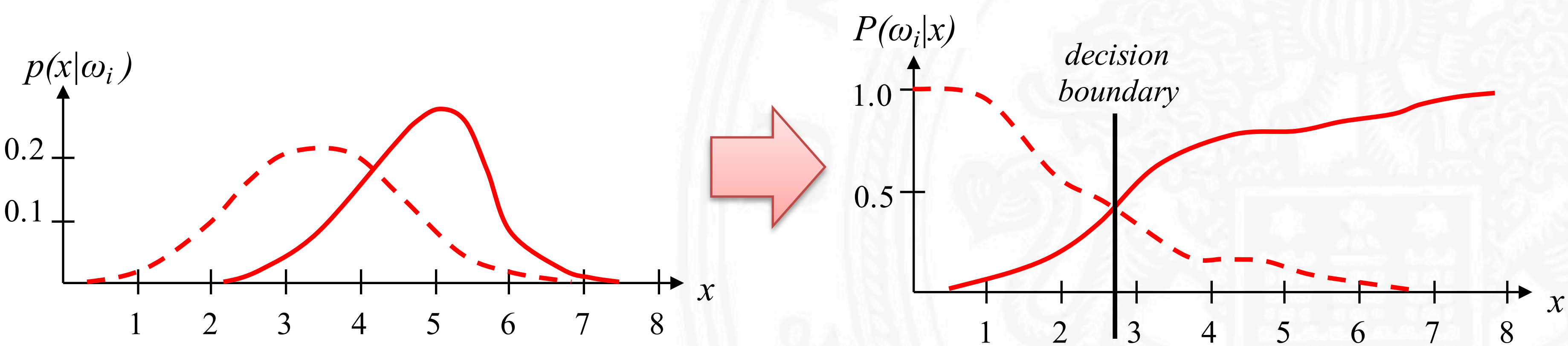
How do we get the "posterior" probabilities $P(\omega_i|x)$?

Example: Medical Screening (3)

The posterior probabilities $P(\omega_i|x)$ can be computed from the "likelihood" $p(x|\omega_i)$ using **Bayes´ formula**:

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

For the example, using Bayes´ Formula, one could get:



General Framework for Bayes Classification

Statistical decision theory minimizes the probability of error for classifications based on uncertain evidence

$\omega_1 \dots \omega_K$	K classes
$P(\omega_k)$	prior probability that an object of class k will be observed
$\vec{x}^T = (x_1 \dots x_N)$	N-dimensional feature vector of an object
$p(\vec{x} \mid \omega_k)$	conditional probability ("likelihood") of observing $\vec{x}$ given that the object belongs to class ω_K
$P(\omega_k \mid \vec{x})$	conditional probability ("posterior probability") that an object belongs to class ω_K given $\vec{x}$ is observed

Bayes decision rule:

Classify given evidence $\vec{x}$ as class ω' such that ω' minimizes the probability of error $P(\omega \neq \omega' \mid \vec{x})$

→ Choose ω' which maximizes the posterior probability $P(\omega \mid \vec{x})$

$g_i(\vec{x}) = P(\omega_i \mid \vec{x})$ are discriminant functions.

Bayes 2-class Decisions

If the decision is between 2 classes ω_1 and ω_2, the decision rule can be simplified:

Choose ω_1 if $\quad \dfrac{p(\vec{x}\,|\omega_1)}{p(\vec{x}\,|\omega_2)} > \dfrac{P(\omega_2)}{P(\omega_1)}$

$\dfrac{p(\vec{x}\,|\omega_1)}{p(\vec{x}\,|\omega_2)}$ is called the "likelihood ratio"

Several alternative forms are possible for a discriminant function:

$$g(\vec{x}) = P(\omega_1\,|\vec{x}) - P(\omega_2\,|\vec{x}) \qquad g(\vec{x}) = \frac{P(\omega_1\,|\vec{x})}{P(\omega_2\,|\vec{x})}$$

For exponential and Gaussian distributions it is useful to take the logarithm:

$$g(\vec{x}) = \log\left(\frac{P(\omega_1\,|\vec{x})}{P(\omega_2\,|\vec{x})}\right) = \log\left(\frac{p(\vec{x}\,|\omega_1)P(\omega_1)}{p(\vec{x}\,|\omega_2)P(\omega_2)}\right) = \boxed{\log\left(\frac{p(\vec{x}\,|\omega_1)}{p(\vec{x}\,|\omega_2)}\right) - \log\left(\frac{P(\omega_2)}{P(\omega_1)}\right)}$$

Normal Distributions

Gaussian ("normal") multivariate distribution: $p(\vec{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{N}{2}}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}$

with: $\Sigma = E\left[(\vec{x}-\vec{\mu})^T (\vec{x}-\vec{\mu})\right]$ N×N covariance matrix

$\vec{\mu}$ mean vector

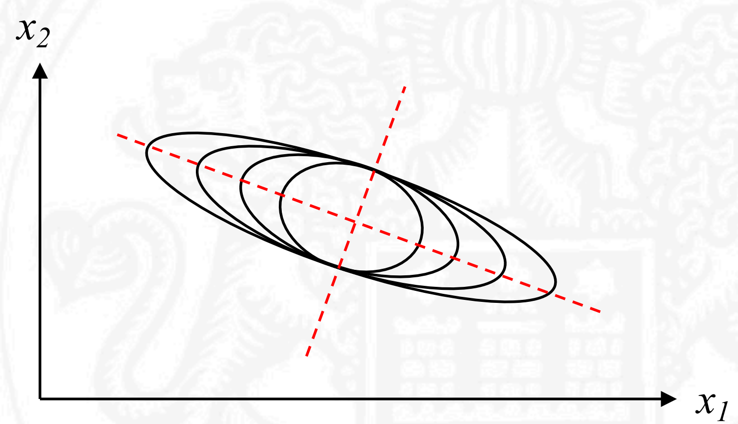
For decision problems, loci of points of constant density are interesting. For Gaussian multivariate distributions, these are hyperellipsoids:

$(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})$ = constant

Eigenvectors of Σ determine directions of principal axes of the ellipsoids,

Eigenvalues determine lengths of the principal axes.

$d^2 = (\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})$ is called "squared Mahalanobis distance" of $\vec{x}$ from $\vec{\mu}$.



Discriminant Function for Normal Distributions

General form:

$$g_i(\vec{x}) = \log\left(p\left(\vec{x}\,|\,\omega_i\right)\right) - \log\left(P(\omega_i)\right)$$

For $p\left(\vec{x}\,|\,\omega_i\right) \approx N\left(\vec{\mu}_i, \Sigma_i\right)$:

$$g_i(\vec{x}) = -\tfrac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu}) - \tfrac{N}{2}\underbrace{\log(2\pi)}_{\text{irrelevant for decisions}} - \tfrac{1}{2}\log\left(|\Sigma_i|\right) + \log\left(P(\omega_i)\right)$$

We consider the discriminant functions for three interesting special cases:

- univariate distribution $N=1$
- statistically independent, equal variance variables x_i
- equal covariance matrices $\Sigma_i = \Sigma$

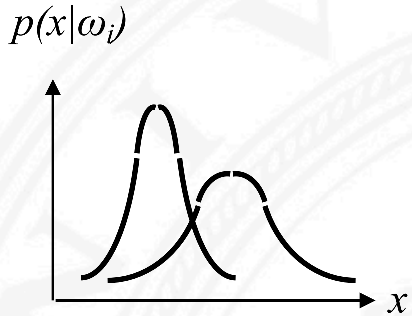
Univariate Distribution

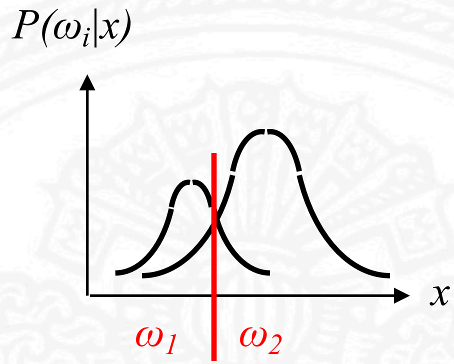
Assumption: $p(x|\omega_i)$ are univariate Gaussian distributions.

<u>Example</u>: 2 classes

$$p(x\,|\,\omega_1) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$$

$$p(x\,|\,\omega_2) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}$$





Decision rule:

$$g_i(x) = \log\left(P\left(\omega_i\,|x\right)\right)$$
$$= -\frac{1}{2\sigma_i^2}(x-\mu)^2 - \tfrac{1}{2}\log\left(\sigma_i\right) + \log\left(P\left(\omega_i\right)\right)$$

Statistically Independent, Equal Variance Variables

In case of insufficient statistical data, variables are sometimes assumed to be statistically independent and of equal variance.

$$\Sigma_i = \sigma^2 I$$

$$g_i(\vec{x}) = -\frac{1}{2\sigma_i^2}\|\vec{x}-\vec{\mu}\|^2 + \log(P(\omega_i))$$

If $P(\omega_i) = 1/N$, then the decision rule is equivalent to the minimum-distance classification rule.

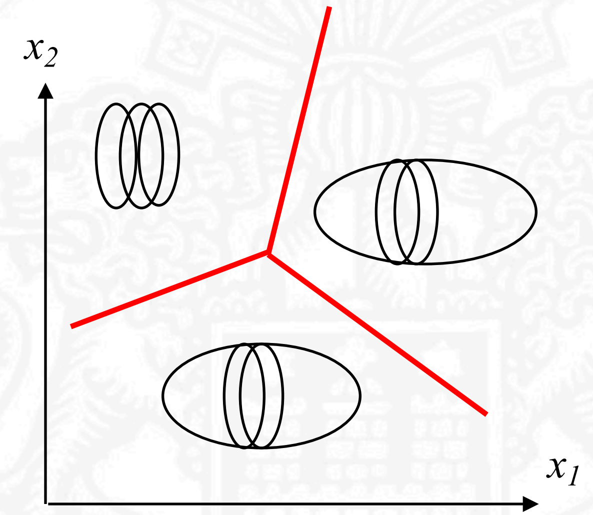
By expanding $g_i(\vec{x})$ and dropping the $\vec{x}^T\vec{x}$ term, one gets the decision rule:

$$g_i(\vec{x}) = -\frac{1}{2\sigma_i^2}\left[-2\vec{\mu}^T\vec{x} + \vec{\mu}^T\vec{\mu}\right] + \log(P(\omega_i))$$

which is linear in $\vec{x}$ and can be written as:

$$g_i(\vec{x}) = (w_i)^T\,\vec{x} + w_{i_0}$$

The decision surface is composed of hyperplanes.



Equal Covariance Matrices

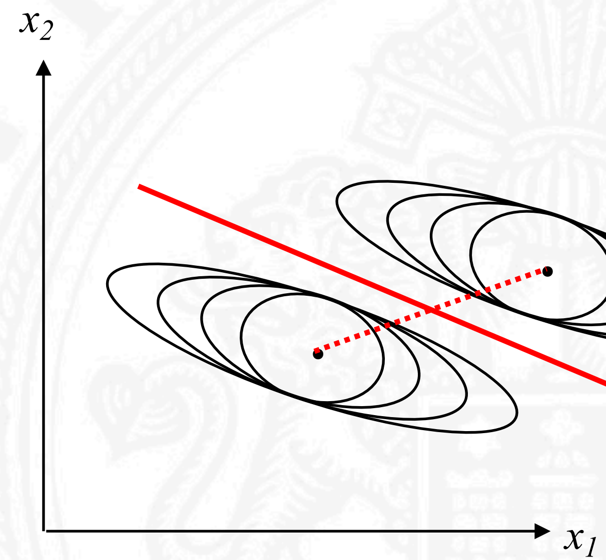
If $\Sigma_i = \Sigma$, the decision rule can be simplified:

$$g_i(\vec{x}) = -\frac{1}{2\sigma_i^2}(\vec{x}-\vec{\mu})^T \Sigma^{-1}(\vec{x}-\vec{\mu}) + \log(P(\omega_i))$$

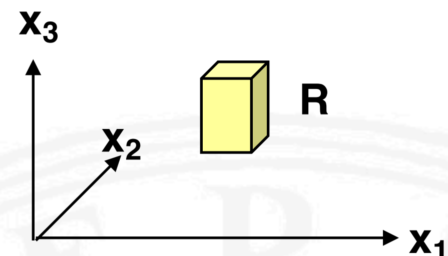
By expanding the quadratic form and dropping $\vec{x}^T \Sigma^{-1} \vec{x}$ one gets another linear decision rule which can (again) be written as:

$$g_i(\vec{x}) = (w_i)^T \vec{x} + w_{i_0}$$

If the a-priori probabilities are equal, the decision rule assigns $\vec{x}$ to the class where the Mahalanobis distance to the mean $\vec{\mu}_i$ is minimal.



Estimating Probability Densities



Let R be a region in feature space with volume V.

Let k out of N samples lie in R.

$$\int_R p(\vec{x}')d\vec{x}' \approx \frac{k}{N} \approx p(\vec{x})V$$

$$p(\vec{x}) \approx \frac{\frac{k}{N}}{V}$$ **relative frequency of samples per volume**

A sequence of approximations $p_n(\vec{x})$ may be obtained by changing the volume V_n as the number of samples n increases.

Examples:

$V_n \sim 1/\sqrt{n}$	Parzen Windows
$k_n \sim \sqrt{n}$	adjust volume for k nearest neighbours

Conditions for a <u>converging</u> sequence of estimates $p_n(\underline{x})$:

1. $\lim_{n\to\infty} V_n = 0$
2. $\lim_{n\to\infty} k_n = \infty$
3. $\lim_{n\to\infty} \frac{k_n}{n} = 0$

Estimating the Mean in a Univariate Normal Density

Given:

$p(x|\mu) = N(\mu, \sigma^2)$ known normal probability density for x except of unknown mean μ

$p(\mu) = N(\mu_0, \sigma_0)$ prior knowledge about μ: a normal density with known μ_0 and σ_0

$X = \{x_1 \dots x_n\}$ samples drawn from $p(x)$

Estimation using Bayes Rule:

$$p(\mu \mid X) = \frac{p(X \mid \mu)p(\mu)}{\int p(X \mid \mu)p(\mu)\,d\mu} = \alpha \prod_{k=1}^{n} p(x_k \mid \mu)p(\mu)$$

α is scale factor independent of μ

$$= \alpha \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2} = \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2}$$

with

$$\boxed{\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \quad \text{and} \quad \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}}$$

Best estimate of mean μ after observing n samples