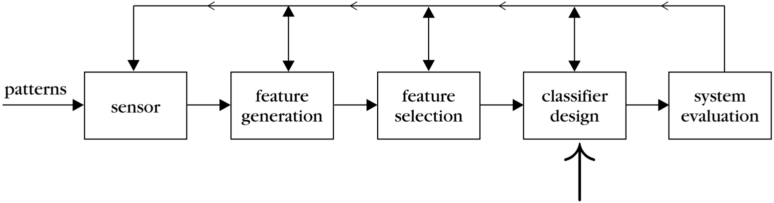
Pattern Recognition Lecture
"Classifiers Based on Bayes Decision Theory"

Prof. Dr. Marcin Grzegorzek

Research Group for Pattern Recognition
www.pr.informatik.uni-siegen.de

Institute for Vision and Graphics
University of Siegen, Germany

Pattern Recognition Chain



Overview

Overview

Statistical Classification - Problem Statement

Classification of an unknown pattern in the most probable of the classes!

- Set of classes: $\{\omega_1, \omega_2, \ldots, \omega_M\}$
- Unknown pattern represented by its feature vector $\mathbf{x}$
- Conditional probabilities: $P(\omega_i|\mathbf{x}), \quad i = 1, 2, \ldots, M$
- Classification result: the class with the maximum conditional probability

But how to compute the conditional probability for a particular class?

Probability P vs. Density p

Probability P

is a real number describing an event belonging to the range $< 0, 1 >$.

Density p

is a value of a function[1] $p(x)$ describing the distribution of the random variable x.

If the random variable takes only discrete values, the densities become probabilities!

[1] This function is often referred as pdf - probability density function.

Overview

A Priori Probability vs. A Posteriori Probability

A priori probability - probability before classification

- How probable is a particular class ω_i for a pattern $\mathbf{x}$ before applying any classification algorithm?
- Answer: $P(\omega_i)$

A posteriori probability - probability after classification

- How probable is a particular class ω_i for a pattern $\mathbf{x}$ after applying a statistical classification algorithm?
- Answer: $P(\omega_i|\mathbf{x})$

Likelihood Density Function

- How feature vectors $\mathbf{x}$ are distributed in a class ω_i?
- Answer: $p(\mathbf{x}|\omega_i)$
- $p(\mathbf{x}|\omega_i)$ is the likelihood function of ω_i with respect to $\mathbf{x}$
- $p(\mathbf{x}|\omega_i)$ can be trained from examples

Bayes Decision Theory for a Two-Class Problem

Known

Classes:	$\{\omega_1, \omega_2\}$
A priori probabilities:	$P(\omega_1)$ and $P(\omega_2)$
Likelihood density functions:	$p(\mathbf{x}\|\omega_1)$ and $p(\mathbf{x}\|\omega_2)$
Pattern to be classified:	$\mathbf{x} = [x_1, x_2, \ldots, x_l]^{\mathrm{T}}$

Assumption

The feature vectors can take any value in the l-dimensional feature space: $\mathbf{x} = [x_1, x_2, \ldots, x_l]^{\mathrm{T}} \in \mathbb{R}^l$

Unknown

A posteriori probabilities: $P(\omega_1|\mathbf{x})$ and $P(\omega_2|\mathbf{x})$

Computation of the A Posteriori Probability

Using the Bayes Rule

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})} \qquad i = 1,2 \qquad (1)$$

$p(\mathbf{x})$ – density function for $\mathbf{x}$

Bayes Classification Rule (1)

Higher a posteriori probability wins

If $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$, $\quad$ $\mathbf{x}$ is classified to ω_1

If $P(\omega_1|\mathbf{x}) < P(\omega_2|\mathbf{x})$, $\quad$ $\mathbf{x}$ is classified to ω_2

Bayes Classification Rule (2)

Considering the Bayes Rule (Eq. 1)

If $\frac{p(\mathbf{x}|\omega_1)P(\omega_1)}{p(\mathbf{x})} > \frac{p(\mathbf{x}|\omega_2)P(\omega_2)}{p(\mathbf{x})}$, $\quad$ $\mathbf{x}$ is classified to ω_1

If $\frac{p(\mathbf{x}|\omega_1)P(\omega_1)}{p(\mathbf{x})} < \frac{p(\mathbf{x}|\omega_2)P(\omega_2)}{p(\mathbf{x})}$, $\quad$ $\mathbf{x}$ is classified to ω_2

Bayes Classification Rule (3)

$p(\mathbf{x})$ can be disregarded, because it is the same for all classes

$$\text{If} \quad p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)\ , \qquad \mathbf{x} \text{ is classified to } \omega_1$$

$$\text{If} \quad p(\mathbf{x}|\omega_1)P(\omega_1) < p(\mathbf{x}|\omega_2)P(\omega_2)\ , \qquad \mathbf{x} \text{ is classified to } \omega_2$$

Bayes Classification Rule (4)

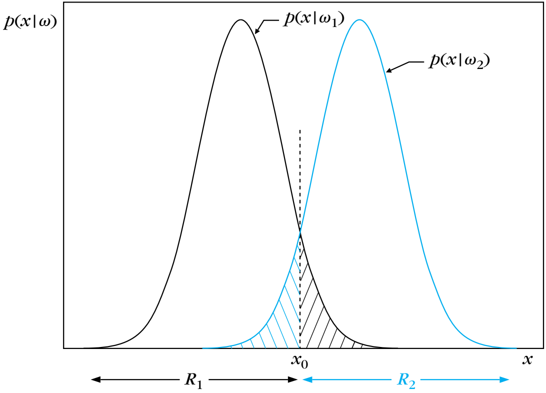
If the a priori probabilities are equal: $P(\omega_1) = P(\omega_2)$

If $p(\mathbf{x}|\omega_1) > p(\mathbf{x}|\omega_2)$, $\quad$ $\mathbf{x}$ is classified to ω_1

If $p(\mathbf{x}|\omega_1) < p(\mathbf{x}|\omega_2)$, $\quad$ $\mathbf{x}$ is classified to ω_2

We are done, since the likelihood density functions $p(\mathbf{x}|\omega_1)$ **and** $p(\mathbf{x}|\omega_2)$ **are assumed to have been trained from examples!**

Classification Error Probability



Error Probability: $$P_e = \frac{1}{2}\int_{-\infty}^{x_0} p(x|\omega_2)dx + \frac{1}{2}\int_{x_0}^{\infty} p(x|\omega_1)dx$$

Classification Error Probability in General

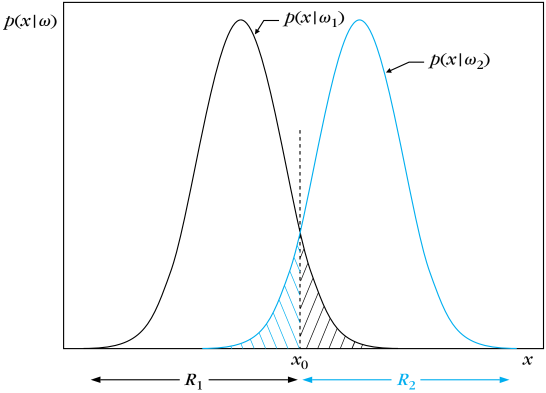
- A priori probabilities are not equal: $P(\omega_1) \neq P(\omega_2)$
- Feature vectors have more than one dimension: $l > 1$

$$\mathbf{x} = [x_1, x_2, \ldots, x_l]^{\mathrm{T}}$$

- General form:

$$P_e = P(\omega_1) \int\limits_{R_2} p(\mathbf{x}|\omega_1) d\mathbf{x} + P(\omega_2) \int\limits_{R_1} p(\mathbf{x}|\omega_2) d\mathbf{x}$$

Classification Error Probability



Bayesian Classifier is OPTIMAL with respect to minimising the classification error probability!

Minimising Average Risk for Two Classes

- Classification error probability assigns the same importance to all errors, which is wrong for many applications (e. g., $\omega_1 \rightarrow$ "malignant tumour", $\omega_2 \rightarrow$ "benign tumour").
- In such cases a penalty term is assigned to weight each error.
- A modified version of the error probability has to be minimised:

$$r = \lambda_{12} P(\omega_1) \int_{R_2} p(\mathbf{x}|\omega_1) d\mathbf{x} + \lambda_{21} P(\omega_2) \int_{R_1} p(\mathbf{x}|\omega_2) d\mathbf{x}$$

- For the tumour example λ_{12} is much greater than λ_{21}.

Modified Bayes Classification Rule

If the a priori probabilities are equal: $P(\omega_1) = P(\omega_2)$

If $\quad p(\mathbf{x}|\omega_2) > p(\mathbf{x}|\omega_1)\frac{\lambda_{12}}{\lambda_{21}}$, $\qquad$ $\mathbf{x}$ is classified to ω_2

If $\quad p(\mathbf{x}|\omega_2) < p(\mathbf{x}|\omega_1)\frac{\lambda_{12}}{\lambda_{21}}$, $\qquad$ $\mathbf{x}$ is classified to ω_1

Overview

Discriminant Functions

- Sometimes it is more convenient to work with functions of probabilities instead of probabilities

$$g_i(\mathbf{x}) \equiv f(P(\omega_i|\mathbf{x}))$$

- $f(\cdot)$ is a monotonically increasing function
- $g_i(\mathbf{x})$ is known as discriminant function
- The decision test is now stated as

$$\text{classify } \mathbf{x} \text{ into } \omega_i \quad \text{if} \quad g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i$$

- The decision surfaces, separating contiguous regions, are described by

$$g_{ij}(\mathbf{x}) \equiv g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0, \quad i, j = 1, 2, \ldots, M \quad i \neq j$$

Overview

Assumption

- The likelihood density functions describing the data in each of the classes, are multivariate Gaussian (normal) distributions

$$p(\mathbf{x}|\omega_i) = (2\pi)^{-\frac{l}{2}} |\mathbf{\Sigma_i}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{i})^{\mathrm{T}} \mathbf{\Sigma_i}^{-1} (\mathbf{x}-\boldsymbol{\mu}_\mathbf{i})}$$

- This "monster" will be denoted by

$$p(\mathbf{x}|\omega_i) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{i}, \mathbf{\Sigma}_i) \qquad i = 1, 2, \ldots, M$$

Discriminant Function $f(\cdot) = \ln(\cdot)$

- Due to the exponential form of the involved densities, the following discriminant function is applied:

$$g_i(\mathbf{x}) = \ln(p(\mathbf{x}|\omega_i)P(\omega_i)) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$$

$$\Updownarrow \text{ considering the "monster"}$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i) + c_i \quad (2)$$

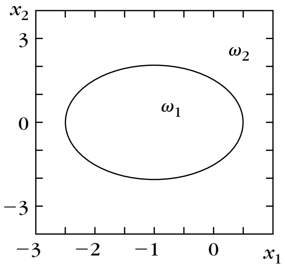
- Where: $c_i = -\frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|$

Quadrics as Decision Curves

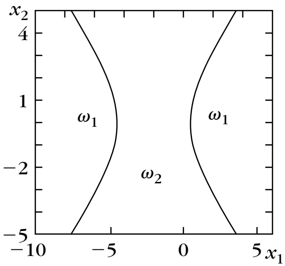
Assuming $l = 2$ and $\sigma_{1,2} = \sigma_{2,1} = 0$, the decision curves

$$g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$$

are quadrics (i. e., ellipsoids, parabolas, hyperbolas, pairs of lines)



(a)



(b)

Decision Hyperplanes

- The only quadric contribution in Equation (2) is $\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{x}$
- Assuming that the covariance matrix is the same for all classes $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ the quadric term will be the same for all discriminant functions
- Thus, the quadric term can be disregarded by decision surface equations. The same is true for the constant c_i
- The simplified version of the discriminant function is just a linear function

$$g_i(\mathbf{x}) = \mathbf{w}_i{}^{\mathrm{T}}\mathbf{x} + w_{i0}$$

where

$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \quad \text{and} \quad w_{i0} = \ln P(\omega_i) - \frac{1}{2}\boldsymbol{\mu}_i{}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$$

Minimum Distance Classifiers

- Assuming equiprobable classes with the same covariance matrix and neglecting the constants Eq. 2 is simplified to:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$$

- If $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ (diagonal matrix) the maximum $g_i(\mathbf{x})$ implies the minimum Euclidean distance $d_\epsilon = ||\mathbf{x} - \boldsymbol{\mu}_\mathbf{i}||$

- Thus, a feature vector $\mathbf{x}$ is assigned to a class $\widehat{i}$ according to its Euclidean distance to the respective mean points $\boldsymbol{\mu}_i$

$$\widehat{i} = \underset{i}{\mathrm{argmax}}(g_i(\mathbf{x})) = \underset{i}{\mathrm{argmin}}(||\mathbf{x} - \boldsymbol{\mu}_i||)$$

Remarks

- In practice, it is quite common to assume the Gaussian distribution of the data. In this case, the Bayesian classifier is either linear or quadratic in nature. These approaches are knows as linear discriminant analysis (LDA) or quadratic discriminant analysis (QDA).

- A major problem associated with LDA and QDA is the large number of parameters to be estimated. Thus, l parameters in each mean vector and approximately $\frac{l^2}{2}$ in each covariance matrix. Moreover, a large number of training points N is needed.

- LDA and QDA perform very good for many different applications. However, in many cases the assumed normal distribution is not the right method to statistically model the data.

Overview

Problem Statement

- So far, we have assumed that the likelihood density functions $p(\mathbf{x}|\omega_i)$ for $i = 1, 2, \ldots, M$ are known.
- This is not the most common case. In many problems, the likelihood density functions have to be estimated from the available training data.
- Here, two estimation methods will be considered, namely
    - Maximum Likelihood Parameter Estimation
    - Maximum a Posteriori Probability Estimation

Maximum Likelihood Parameter Estimation (1)

- Let us consider an M-class problem with feature vectors distributed according to $p(\mathbf{x}|\omega_i)$, $i = 1, 2, \ldots, M$.
- The likelihood functions are assumed to be given in a parametric form. The statistical parameters for the classes ω_i form vectors $\boldsymbol{\theta}_i$ which are unknown

$$p(\mathbf{x}|\omega_i) = p(\mathbf{x}|\omega_i; \boldsymbol{\theta}_i)$$

- Goal: to estimate the unknown parameters using a set of known feature vectors in each class.
- Since the estimation process is the same for all classes, the index i will be skipped for further investigations.

Maximum Likelihood Parameter Estimation (2)

- Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a set of feature vectors describing training samples of a particular class.
- Assuming statistical independence between the different feature vectors, we can form the joint density function

$$p(X; \boldsymbol{\theta}) = p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N; \boldsymbol{\theta}) = \prod_{k=1}^{N} p(\mathbf{x}_k; \boldsymbol{\theta})$$

- The ML method estimates $\boldsymbol{\theta}$ so that the likelihood function takes its maximum value

$$\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{k=1}^{N} p(\mathbf{x}_k; \boldsymbol{\theta})$$

Maximum Likelihood Parameter Estimation (3)

- To find a maximum, the gradient has to be zero

$$\frac{\partial \prod_{k=1}^{N} p(\mathbf{x}_k;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

- Due to the monotonicity of the logarithmic function, we can use also the log-likelihood function

$$L(\boldsymbol{\theta}) = \ln \prod_{k=1}^{N} p(\mathbf{x}_k;\boldsymbol{\theta})$$

- Looking for the maximum here, we have

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{k=1}^{N} \frac{\partial \ln p(\mathbf{x}_k;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{k=1}^{N} \frac{1}{p(\mathbf{x}_k;\boldsymbol{\theta})} \frac{\partial p(\mathbf{x}_k;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

Maximum a Posteriori Probability Estimation

- Set of feature vectors $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$
- $\boldsymbol{\theta}$ is an unknown random vector
- The starting point is the following density function

$$p(\boldsymbol{\theta}|X) = \frac{p(\boldsymbol{\theta})p(X|\boldsymbol{\theta})}{p(X)}$$

- The MAP estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ is defined as a point where $p(\boldsymbol{\theta}|X)$ becomes maximum

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}} : \frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{\theta}|X) = 0 \quad \text{or} \quad \frac{\partial}{\partial \boldsymbol{\theta}} (p(\boldsymbol{\theta})p(X|\boldsymbol{\theta})) = 0$$