Pattern Recognition Lecture
Context-Dependent Classification

Prof. Dr. Marcin Grzegorzek

Research Group for Pattern Recognition
Institute for Vision and Graphics
University of Siegen, Germany

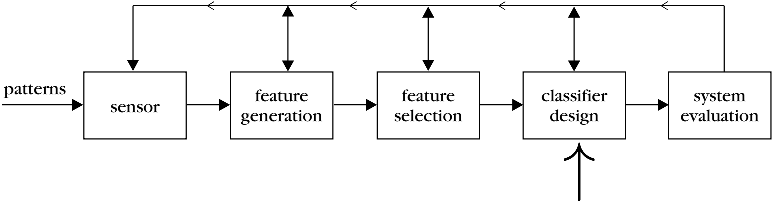
UNIVERSITÄT SIEGEN

IVG

ZESS

Imaging New Modalities
DFG-Graduiertenkolleg 1564

Pattern Recognition Chain



Overview

Overview

Introduction (1)

- The classification tasks considered so far have assumed that no relation exists among the various classes. Having obtained a feature vectors **x** from a class ω_i, the next feature vector could belong to any other class.

- In this lecture we will remove this assumption, and we will assume that the various classes are closely related. That is, successive feature vectors are not independent.

Introduction (2)

- Under such an assumption, classifying each feature vector separately from the others obviously has no meaning.
- The class to which a feature vector is assigned depends on
  - its own value
  - the values of the other feature vectors
  - the existing relation among the various classes.
- The context-dependent classification is performed using all feature vectors simultaneously and in the same sequence in which they occurred from the experiments.
- Therefore, here we will refer to the feature vectors as observations occurring in sequence from $\mathbf{x}_1$ to $\mathbf{x}_N$.

Overview

The Bayes Classifier (1)

- Let $X : \mathbf{x}_1, \ldots, \mathbf{x}_N$ be a sequence of observations (feature vectors) and $\omega_{i=1,\ldots,M}$ the classes in which these vectors are supposed to be classified.
- Let $\Omega_i : \omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_N}$ be one of the possible sequences of these classes corresponding to the observation sequence. The total number of these class sequences is M^N.
- The classification task here is to decide to which class sequence Ω_i a sequence of observations X corresponds. This is equivalent to appending $\mathbf{x}_1$ to class ω_{i_1}, $\mathbf{x}_2$ to ω_{i_2}, and so on.

The Bayes Classifier (2)

- A way to approach the problem is to view each specific sequence X as an extended feature vector and $\Omega_{i=1,\ldots,M^N}$ as the available classes.

- Having observed a specific X, the Bayes rule assigns it to Ω_i for which

$$P(\Omega_i|X) > P(\Omega_j|X), \quad \forall i \neq j$$

- This is equivalent to

$$P(\Omega_i)p(X|\Omega_i) > P(\Omega_j)p(X|\Omega_j), \quad \forall i \neq j \qquad (1)$$

The Bayes Classifier (3)

- The general idea of the context-dependent classification techniques is to use the equation 1 for different class dependence models.

- In the following, some typical class dependence models will be presented.

Overview

Markov Chain Models (1)

- It is one of the most widely used models describing the class dependence.
- If $\omega_{i_1}, \omega_{i_2}, \dots$ is a sequence of classes, then the Markov models assumes that

$$P(\omega_{i_k}|\omega_{i_{k-1}}, \dots, \omega_{i_1}) = P(\omega_{i_k}|\omega_{i_{k-1}})$$

- The meaning of this is that the class dependence is limited only within two successive classes (the so called first-order Markov model).

Markov Chain Models (2)

- In other words, given that the observations $\mathbf{x}_{k-1}, \mathbf{x}_{k-2}, \ldots, \mathbf{x}_1$ belong to classes $\omega_{i_{k-1}}, \omega_{i_{k-2}}, \ldots, \omega_{i_1}$, respectively, the probability of the observation $\mathbf{x}_k$ belonging to class ω_{i_k} depends only on the class from which observation $\mathbf{x}_{k-1}$ has occurred.

- Using the probability chain rule

$$P(\Omega_i) \equiv P(\omega_{i_1}, \ldots, \omega_{i_N}) =$$

$$P(\omega_{i_N} | \omega_{i_{N-1}}, \ldots, \omega_{i_1}) P(\omega_{i_{N-1}} | \omega_{i_{N-2}}, \ldots, \omega_{i_1}) \ldots P(\omega_{i_1})$$

Markov Chain Models (3)

- From the previous two equations we obtain

$$P(\Omega_i) = P(\omega_{i_1}) \prod_{k=2}^{N} P(\omega_{i_k} | \omega_{i_{k-1}})$$

where $P(\omega_{i_1})$ is the prior probability for class ω_{i_1} to occur.

- Furthermore, two commonly adopted assumptions are:
  a) Given the sequence of classes, the observations are statistically independent, and
  b) the probability density function in one class does not depend on the other classes. This implies:

$$p(X | \Omega_i) = \prod_{k=1}^{N} p(\mathbf{x}_k | \omega_{i_k})$$

Markov Chain Models - Statement

- Having observed the sequence of feature vectors $X : \mathbf{x}_1, \ldots, \mathbf{x}_N$, classify them in the respective sequence of classes $\Omega_i : \omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_N}$ so that the quantity

$$p(X|\Omega_i)P(\Omega_i) = P(\omega_i)p(\mathbf{x}_1|\omega_{i_1}) \prod_{k=2}^{N} P(\omega_{i_k}|\omega_{i_{k-1}})p(\mathbf{x}_k|\omega_{i_k})$$

  becomes maximum.

- Searching for this maximum requires computation of this equation for each of the $\Omega_{i=1,\ldots,M^N}$. This amounts to a very high number of $O(NM^N)$ multiplications. However, in the next lecture we will learn how to implement the maximisation algorithm in a efficient way.

Overview

Introductory Example (1)

- Let us consider the coin-tossing problem as an example.
- Assume that the coin tossing takes place behind a curtain and all we known is just an outcome of each experiment.
- That is, each time an experiment is performed we cannot know the specific coin (in the case of multiple coins) whose tossing resulted in the current observation (head or tails).
- Thus, a crucial part of the probabilistic process is hidden to us.

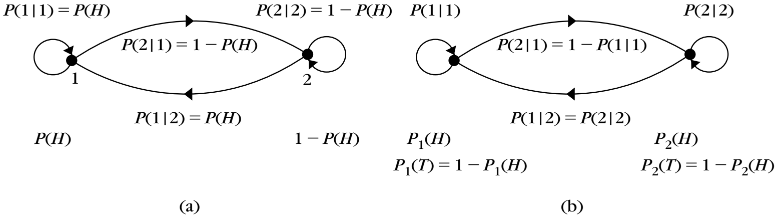
Introductory Example (2)

- In the first experiment, a single coin is tossed to produce a sequence of heads (H) and tails (T).
- This experiment is characterised by a single parameter indicating the propensity of the coin for landing heads and is quantified by the probability $P(H) = 1 - P(T)$.
- A straightforward modelling of this statistical process is to associate one state with the outcome H and one with the outcome T.
- Hence, this is an example of a process with observable states.

Introductory Example (3)

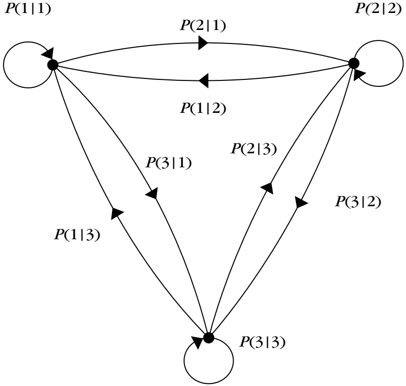
- In the second experiment, two coins are used behind the curtain.
- This model cannot be sufficiently modelled by a single parameter.
- To model the process we assume two states, corresponding to the two coins.
- The states of this process are not observable, since we have no access to the information related to which coin is tossed each time.

Markov Models for One and Two Coins



(a) single coin, (b) two coins. $P(i|j)$ denotes the transition probability from state s_j to state s_i once the coin has been tossed and an observation has been made available to us.

Markov Model for Three Hidden Coins



$P_1(H)$ $P_2(H)$ $P_3(H)$

$P_1(T) = 1 - P_1(H)$ $P_2(T) = 1 - P_2(H)$ $P_3(T) = 1 - P_3(H)$

HMM in General

- A HMM generates an observation string, that is, the sequence of observation vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$.
- Thus, an HMM model consists of a number of, say, K states and the observation string is produced as a result of successive transitions from one state i to another j.
- On the next slide, a typical diagram of an HMM of three states is shown, where arrows indicate transitions.

Three-State Hidden Markov Model