7. Basics of Pattern Classification
Definition of the Task
Informal Definition: Telling things apart
Pattern Recognition: Automatic transformation of data $x_i$ (observations, features) into a set of symbols $\omega_i$ (classes).
Flow of Data in Pattern Recognition

Test Data \( x_i \)

Feature Extraction

Classifier

\( \omega_1 \)

\( \omega_2 \)

\( \ldots \)

\( \omega_n \)

Model

Training Data

Feature Extraction

Training Algorithm
Supervised Learning

- Labeled data exist
- Task: label new data
- Question: Where to put the decision boundary

=? = new sample
Unsupervised Learning

• True labels are unknown
• Task: put similar objects into the same class
• How to cluster the data?
• Do you know the number of classes?
Types of Tasks and Signals
Types of Data (Patterns): Discrete

e.g. text classification

Speech Recognition
Information Retrieval
Computer Linguistics
Everything else
Types of Data (Patterns):
One Dimensional time series

Speech or other kinds of sounds

ECG or other bio signal
LMT Applikationsbeispiel: Tanksensor

- Schnelle Unterscheidung von Benzin- und Dieseldämpfen zur Vermeidung von Fehlbetankungen (insbesondere für TDI Motoren)

- Alternativ einsetzbar in der Zapfpistole oder im Fahrzeug
Types of Data (Patterns):
Multi-dimensional time series

Controlling a car

Stock prices

From: Schuckat-Talamzzini
Types of Data (Patterns):

two dimensional data
Types of Data (Patterns): two dimensional time series

Video
(Informedia project TRECVID)
Types of Data (Patterns):
three dimensional

3d reconstruction of CT images
A Simple Classifier:
K Nearest Neighbours
measure distance to training data

- Consider two class problem

k-nearest neighbor classifier
measure distance to training data

• Consider two class problem

k-nearest neighbor classifier
Voronoi Cells in 2 Dimensions

From: Duda+Hart: Pattern Classification
Voronoi Cells in 3 Dimensions
Decision Boundary for a nearest-neighbour classifier in a Simulation

(Probability Distribution given)

From: Hastie et al.: Statistical Learning
k-Nearest-Neighbour-Classifier

From: Duda+Hart: Pattern Classification
Error of k-Nearest-Neighbour-Classifier

From: Duda+Hart: Pattern Classification
Decision Boundary for a $k$-nearest-neighbour classifier in a Simulation
(Probability Distribution given)

From: Hastie et al.: Statistical Learning
Missclassification vs. Number of Neighbours
Other Classifiers

- Bayes Classifier
- Decision Trees
- Gaussian Mixture Models
- Support Vector Machines
- Neural Networks
- Hidden Markov Models
- Conditional Random Fields
- Boosting
- ...

-> See winter semester: Pattern and Speech Recognition
Welcome to the UC Irvine Machine Learning Repository!

We currently maintain 171 data sets as a service to the machine learning community. You may view all data sets through our searchable interface. Our old web site is still available, for those who prefer the old format. For a general overview of the Repository, please visit our About page. For information about citing data sets in publications, please read our citation policy. If you wish to donate a data set, please consult our donation policy. For any other questions, feel free to contact the Repository librarians.

Supported By: In Collaboration With:
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Task</th>
<th>Data Type</th>
<th>Instances</th>
<th>Attributes</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>Multivariate</td>
<td>Classification</td>
<td>Categorical</td>
<td>286</td>
<td>9</td>
<td>1988</td>
</tr>
<tr>
<td>Breast Cancer Wisconsin (Diagnostic)</td>
<td>Multivariate</td>
<td>Classification</td>
<td>Real</td>
<td>569</td>
<td>32</td>
<td>1995</td>
</tr>
<tr>
<td>Breast Cancer Wisconsin (Original)</td>
<td>Multivariate</td>
<td>Classification</td>
<td>Integer</td>
<td>699</td>
<td>10</td>
<td>1992</td>
</tr>
<tr>
<td>Breast Cancer Wisconsin (Prognostic)</td>
<td>Multivariate</td>
<td>Classification, Regression</td>
<td>Real</td>
<td>198</td>
<td>34</td>
<td>1995</td>
</tr>
<tr>
<td>CalIt2 Building People Counts</td>
<td>Multivariate, Time-Series</td>
<td>Classification</td>
<td>Categorical, Integer</td>
<td>10080</td>
<td>4</td>
<td>2006</td>
</tr>
<tr>
<td>Car Evaluation</td>
<td>Multivariate</td>
<td>Classification</td>
<td>Categorical</td>
<td>1728</td>
<td>6</td>
<td>1997</td>
</tr>
<tr>
<td>Census Income</td>
<td>Multivariate</td>
<td>Classification</td>
<td>Categorical, Integer</td>
<td>48842</td>
<td>14</td>
<td>1996</td>
</tr>
<tr>
<td>Census-Income (KDD)</td>
<td>Multivariate</td>
<td>Classification</td>
<td>Categorical, Integer</td>
<td>299285</td>
<td>40</td>
<td>2000</td>
</tr>
<tr>
<td>Challenger USA Space Shuttle O Ring</td>
<td>Multivariate</td>
<td>Regression</td>
<td>Integer</td>
<td>23</td>
<td>4</td>
<td>1993</td>
</tr>
</tbody>
</table>
Letter Recognition Data Set

**Abstract:** Database of character image features; try to identify the letter

<table>
<thead>
<tr>
<th>Data Set Characteristics:</th>
<th>Multivariate</th>
<th>Number of Instances:</th>
<th>20000</th>
<th>Area:</th>
<th>Computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute Characteristics:</td>
<td>Integer</td>
<td>Number of Attributes:</td>
<td>16</td>
<td>Date Donated</td>
<td>1991-01-01</td>
</tr>
<tr>
<td>Associated Tasks:</td>
<td>Classification</td>
<td>Missing Values?:</td>
<td>No</td>
<td>Number of Web Hits:</td>
<td>3821</td>
</tr>
</tbody>
</table>

**Source:**

Creator:

David J. Slate  
Odesta Corporation; 1890 Maple Ave; Suite 115; Evanston, IL 60201

Donor:

David J. Slate (dave '@' math.nwu.edu) (708) 491-3867
Summary

• Task of machine learning
• Example applications
• Nearest neighbor classifier