

Machine Learning

Central Problem of Pattern Recognition: Supervised and Unsupervised Learning

Classification
Bayesian Decision Theory
Perceptrons and SVMs
Clustering
Dimension Reduction

The Problem of Pattern Recognition

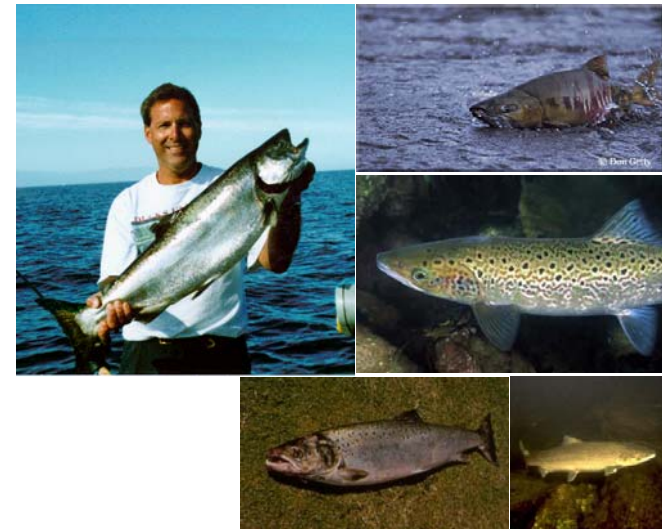
Machine Learning (as statistics) addresses a number of challenging *inference* problems in pattern recognition which span the range from statistical modeling to efficient algorithmics. *Approximative method* which yield *good performance on average* are particularly important.

- **Representation of objects.** ⇒ Data representation
- What is a *pattern*? **Definition/modeling of structure.**
- **Optimization:** Search for preferred structures
- **Validation:** are the structures indeed in the data or are they explained by fluctuations?

Literatur

- Richard O. Duda, Peter E. Hart & David G. Stork, *Pattern Classification*. Wiley & Sons (2001)
- Trevor Hastie, Robert Tibshirani & Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag (2001)
- Luc Devroye, Laslo Györfi & Gabor Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer Verlag (1996)
- Vladimir N. Vapnik, *Estimation of Dependences Based on Empirical Data*. Springer Verlag (1983); *The Nature of Statistical Learning Theory*. Springer Verlag (1995)
- Larry Wasserman, *All of Statistics*. (1st ed. 2004. Corr. 2nd printing, ISBN: 0-387-40272-1) Springer Verlag (2004)

The Classification Problem





Classification as a **Pattern Recognition Problem**

Problem: We look for a partition of the object space \mathcal{O} (fish in the previous example) which corresponds to classification examples.

Distinguish conceptually between “objects” $o \in \mathcal{O}$ and “data” $x \in \mathcal{X}$!

Data: pairs of **feature vectors** and **class labels**

$$\mathcal{Z} = \{(x_i, y_i) : 1 \leq i \leq n, x_i \in \mathbb{R}^d, y_i \in \{1, \dots, k\}\}$$

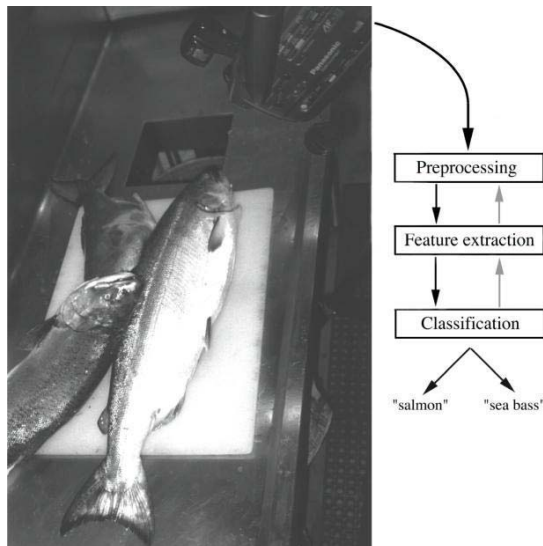
Definitions: **feature space** \mathcal{X} with $x_i \in \mathcal{X} \subset \mathbb{R}^d$

class labels $y_i \in \{1, \dots, k\}$

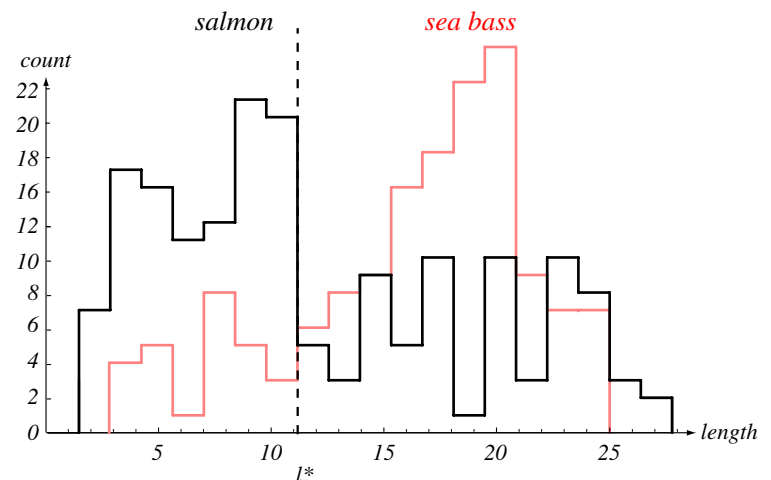
Classifier: mapping $c : \mathcal{X} \rightarrow \{1, \dots, k\}$

k class problem: What is $y_{n+1} \in \{1, \dots, k\}$ for $x_{n+1} \in \mathbb{R}^d$?

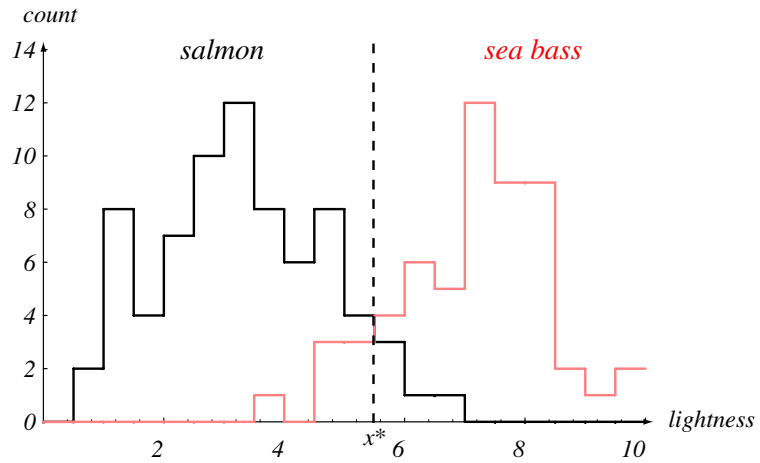
Example of Classification



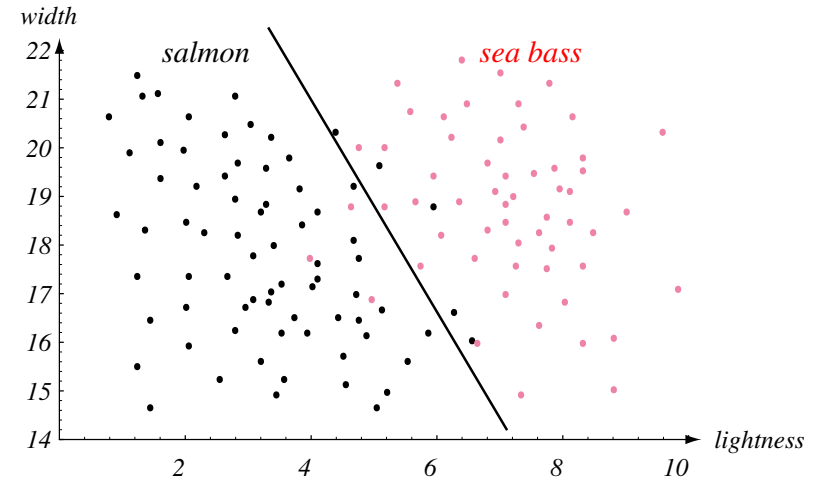
Histograms of Length Values



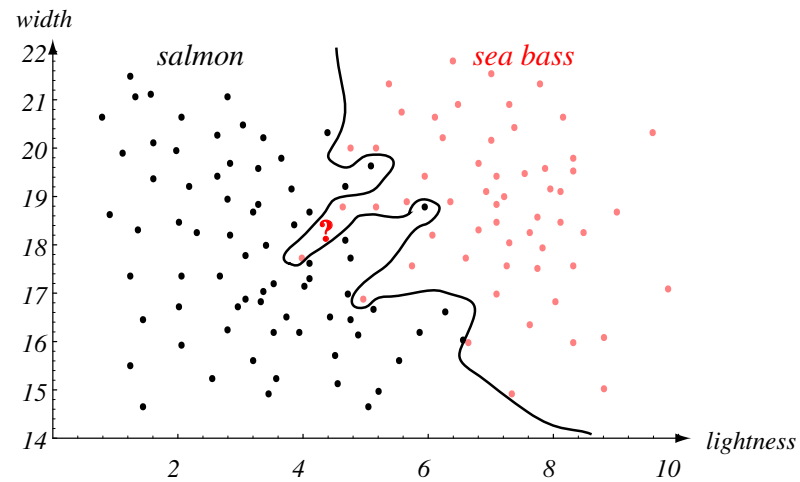
Histograms of Skin Brightness Values



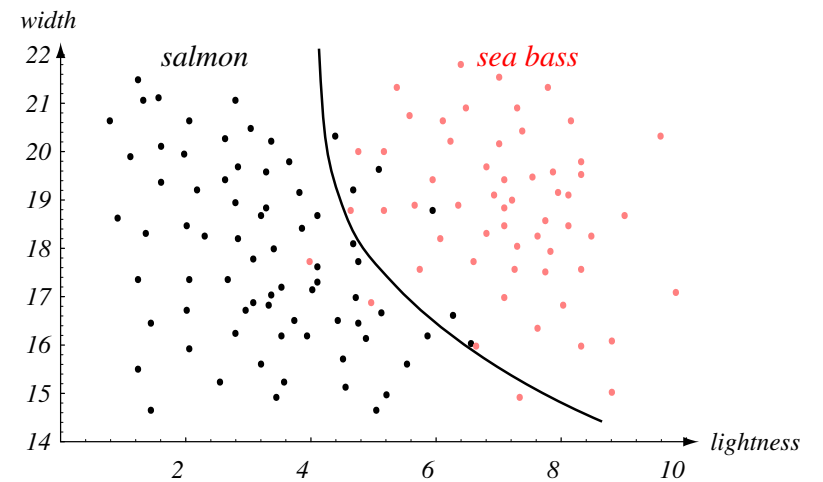
Linear Classification



Overfitting



Optimized Non-Linear Classification

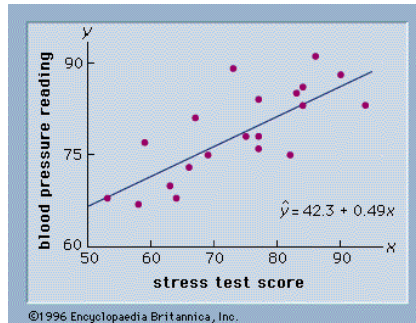


Occam's razor argument: *Entia non sunt multiplicanda praeter necessitatem!*

Regression

(see Introduction to Machine Learning)

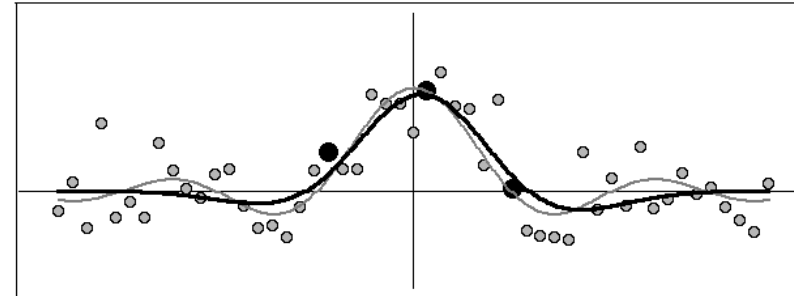
Question: Given a feature (vector) x_i and a corresponding noisy measurement of a function value ($f(x_i) + \text{noise}$) what is the unknown function $f(\cdot) \in$ hypothesis class?



Data: $\mathcal{Z} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : 1 \leq i \leq n\}$

Modeling choice: What is an adequate hypothesis class?
Fitting with linear or nonlinear functions?

Nonlinear Regression



50 noisy data from a sinc function $\text{sinc}(x) := \sin(x)/x$ (gray) with a regression fit (black).

The Regression Function

Question: What is the optimal estimate of a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ based on noisy data?

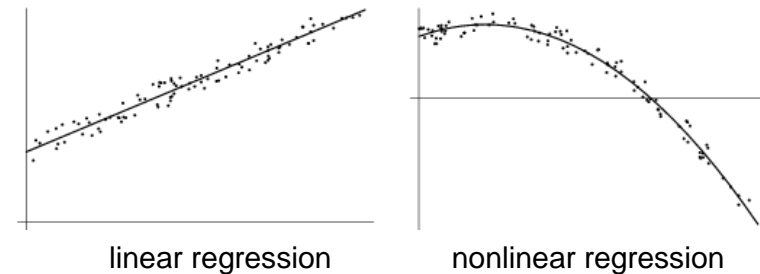
$$y_i = f(x_i) + \eta_i \quad \leftarrow \quad \text{noise ?}$$

Solution: the regression function

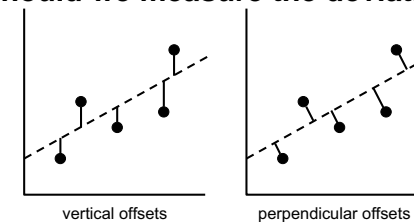
$$y(x) = \mathbb{E} \{y|X = x\} = \int_{\Omega} y p(y|X = x) dy$$

Model selection: How many data are required to estimate a model or a regression function?

Examples of linear and nonlinear regression



How should we measure the deviations?



Core Questions of Pattern Recognition: Unsupervised Learning

1. **data clustering, vector quantization**
2. hierarchical data analysis; search for tree structures in data
3. visualisation, **dimension reduction**
 - (a) *PCA* or principal component analysis
 - (b) *ICA*, independent component analysis
 - (c) *overcomplete basis*
 - (d) *projection pursuit*
 - (e) multidimensional scaling, representation of proximity data as Euclidean distances in \mathbb{R}^d

Modes of Learning

Reinforcement Learning: weakly supervised learning

Action chains are evaluated at the end.

Backgammon; the neural network *TD-Gammon* gained the world championship! Quite popular in **Robotics**

Active Learning: Data are selected according to their expected information gain.

Information Filtering

Inductive Learning: the learning algorithm extracts logical rules from the data.

Inductive Logic Programming is a popular sub area of **Artificial Intelligence**

Taxonomy of Data with some Examples

Data are representations of measurements / observations!

a) monadic data: $\mathcal{O} = \mathcal{O}^{(1)}$

water depth, temperature, pressure, intensity, ...

b) dyadic data: $\mathcal{O} = \mathcal{O}^{(1)} \times \mathcal{O}^{(2)}$

{users} \times {websites}

{gene expression levels} \times {diseases}

{word counts} \times {documents}

pairwise data: $\mathcal{O} = \mathcal{O}^{(1)} \times \mathcal{O}^{(2)}$ with $\mathcal{O}^{(1)} = \mathcal{O}^{(2)}$

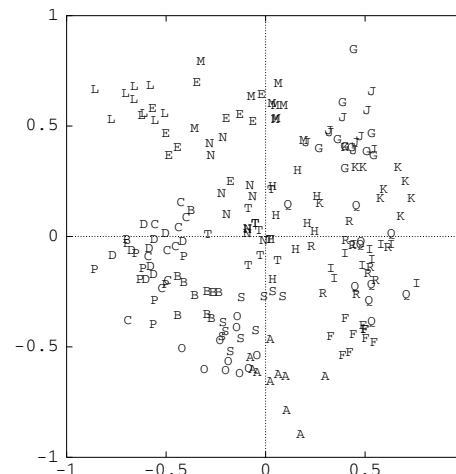
{proteins} \times {proteins}

{image patches} \times {image patches}

c) polyadic data: $R \geq 3$

{test persons} \times {behaviors} \times {traits}

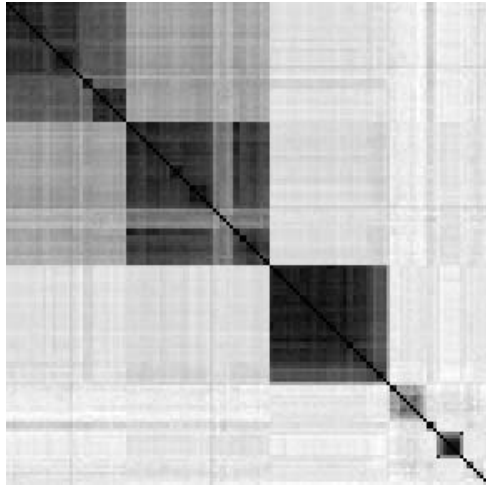
Example for Vectorial Data



Data of 20 Gaussian sources in \mathbb{R}^{20} , projected onto two dimensions with Principal Component Analysis.

Example of Relational Data

Pairwise dissimilarity of 145 globins which have been selected from 4 classes of α -globine, β -globine, myoglobins and globins of insects and plants.



Scales for Data

Nominal or **categorical scale**: qualitative, but without quantitative measurements,
e.g. *binary scale* $\mathcal{F} = \{0, 1\}$ (*presence or absence of properties*) or
taste categories “*sweet, sour, salty and bitter*.”

Ordinal scale : measurement values are meaningful only with respect to other measurements, i.e., the rank order of measurements carries the information, not the numerical differences (e.g. *information on the ranking of different marathon races!?*)

Quantitative scale:

- **interval scale**: the relation of numerical differences carries the information. Invariance w.r.t. translation and scaling (*Fahrenheit scale of temperature*).
- **ratio scale**: zero value of the scale carries information but not the measurement unit. (*Kelvin scale*).
- **Absolute scale**: Absolute values are meaningful. (*grades of final exams*)