Universität Stuttgart

Pattern Recognition
Chapter 7:
Overview of statistical Methods

Prof. Dr.-Ing. Uwe Sörgel
soergel@ifp.uni-stuttgart.de

ifp

Contents

- Introduction to statistical methods in pattern recognition and image analysis
  - Tasks and solution strategies
  - The feature space
  - Taxonomy of statistical methods
  - Overfitting Problem

Statistical methods of image analysis I

- Objects are **not** primarily described by **models**, but by **statistical properties** of the sensor data in relation to the objects
- Requires a model of statistical properties
- Purpose: Recognition of objects → Classification
- Learning of properties from examples → "Machine Learning"
- Model knowledge may be considered *implicitly* by the selection of suitable features for the classification.

Statistical methods of image analysis II

- Objects correspond to connected regions that are assigned to a certain category.
- Classification can also be seen as a form of segmentation ("semantic segmentation").
- Usually, post-processing of the classification results is required, for example, by morphological operators.
- Output can serve as the basis for high-level processing in knowledge based image analysis.

Contents

- Introduction to statistical methods in pattern recognition and image analysis
  - Tasks and solution strategies
  - The feature space
  - Taxonomy of statistical methods
  - Overfitting Problem

Statistical methods: Task I

- Given:
  - Image primitives P_i, $i \in [0, \dots N\text{-}1]$
    - Pixels or image regions (from segmentation))
  - **Features** $\mathbf{x}_i$ for every primitive P_i with $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots x_{iD}]^T$
    - Derived from sensor data (cf. lecture "Features")
    - Usually real numbers, quantization can lead to discrete values (e.g. grey value: 8 bit → 0, 1, …, 255)
    - D is the dimension of the feature vector
    - Features may be derived from multiple sensors ⇒ Data fusion

Statistical methods: Task II

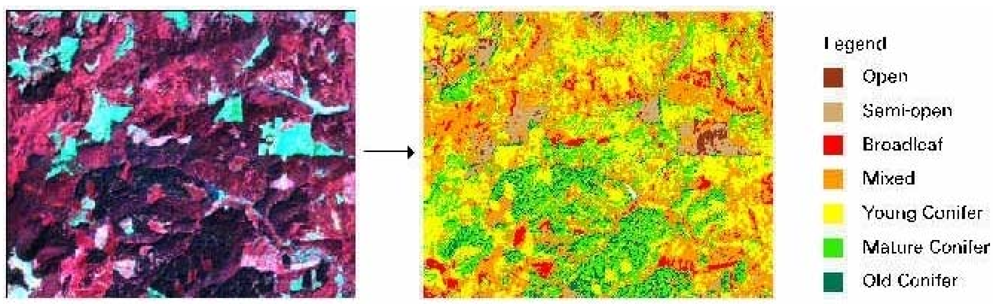
- Wanted:
  - Information about type / class C_i of every primitive P_i
    - Discrete set of M classes: $C_i \in \{C^1, \dots C^M\}$
    - Representation: every class C^j is assigned to a “class label“ j, e.g. $C^1 \leftrightarrow 1$, $C^2 \leftrightarrow 2, \dots$
      - The *superscript* index refers to the label of the *class*, whereas *subscript* index indicates the class a *primitive* is assigned to.
    - “Closed world assumption”: There are no other classes except the given ones.
    - Binary classification: Special case for $M = 2$
      - For example: C^1 = “object”, C^2 = “background”
        class labels: often $\{0, 1\}$ or {-1, 1} for $\{C^2, C^1\}$

Statistical methods: Task III

- Result of classification:
  - Label image **C**, whose “grey value“ C_i at pixel i indicates the class label of the corresponding image primitive

- Example (differentiation of forest types from a satellite image):

spectral information (implicit) → thematic information (explicit)



Statistical methods: Probabilistic approach I

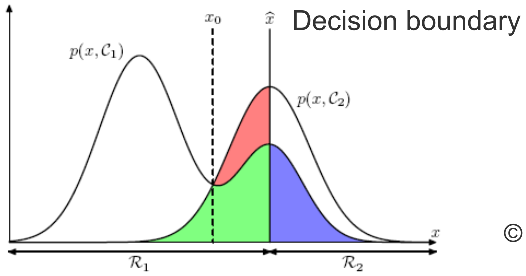
- Both the features $\mathbf{x}$ and the class labels C are considered to be **random variables**.
- The joint distribution of $\mathbf{x}$ and C is described by the **probability density** $p(\mathbf{x},C)$, whose parameters can be determined from training data.
- C is determined so that the conditional probability $p(C|\mathbf{x})$ for the class label C given the observed data $\mathbf{x}$ is maximized:

maximum a posteriori (MAP)

$$C = \underset{C}{\operatorname{argmax}}\left(p\left(C|\mathbf{x}\right)\right)$$

Statistical methods: Probabilistic approach II

- MAP corresponds to the minimization of classification errors
- Example (two-class-problem, single feature x):
  - The probability for classification errors corresponds the sum of the colored areas.
    - Blue: Probability for assigning a feature x to C_2 although it belongs to C_1.
    - Sum of green and red areas: Probability for assigning a feature x to C_1 although it belongs to C_2.
  - Variation of threshold leads to change of red area, while the *sum* of green and blue areas is constant.
  - At position x_0 holds $p(x,C_1) = p(x,C_2) \rightarrow p(C_1|x) = p(C_2|x)$, there is the red area 0 and therefore **the probability for a classification error is minimal.**



Statistical methods: Non-probabilistic approach

- Probabilities are not modeled directly
- The goal is to find the optimal separating surface between the classes in feature space on the basis of training data.

- Different criteria for optimality, e.g.
  - "Maximum margin": Maximize distance of the separating surface from the nearest training points.
  - Minimize the training error

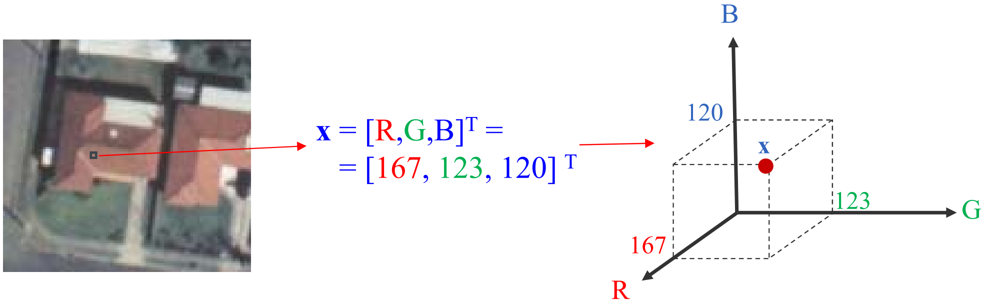
- The class C an image primitive is determined according to the position of ist feature vector $\mathbf{x}$ relative to the separating surface.

Contents

- Introduction to statistical methods in pattern recognition and image analysis
  - Tasks and solution strategies
  - The feature space
  - Taxonomy of statistical methods
  - Overfitting Problem

The feature space I

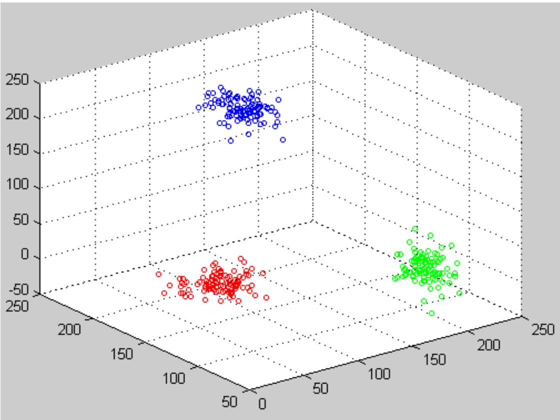
- For every image primitive (pixel, segment), a feature vector $\mathbf{x} = [x_1, x_2, \dots x_D]^T$ is determined from sensor data.
- $\mathbf{x}$ can be interpreted as a point in a D-dimensional feature space.
- Example: Color image with three channels (R,G,B):



- The components of $\mathbf{x}$ may be derived from different sensors
  $\Rightarrow D > 100$ can occur!

The feature space II

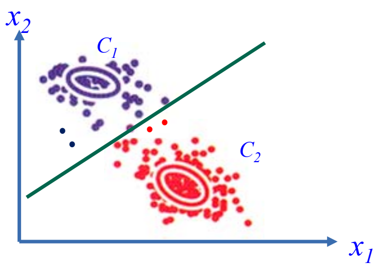
- Image primitives (pixel, segments) of the same class have similar properties, therefore their feature vectors are “close” in feature space.
- Consequently, the classes correspond to clusters in feature space.
- Training: Search for the clusters and determine their parameters.
- Classification: Every image primitive is assigned to the most similar cluster in feature space.



The feature space III

- Example:

$$\mathbf{x} = [x_1, x_2]^T$$
$$C \in \{C^1, C^2\}$$



- On the basis of x_1 alone, C^1 and C^2 cannot be separated properly.
- Increase the dimension of the feature space
  → C^1 and C^2 can be seperated
- The selection of the features is crucial for the success of the classification.
- In remote sensing, for example, each multi-spectral image corresponds to one dimension of the feature space.
- Selection of the features: **often based on** model knowledge
- **Learning of features →** deep learning

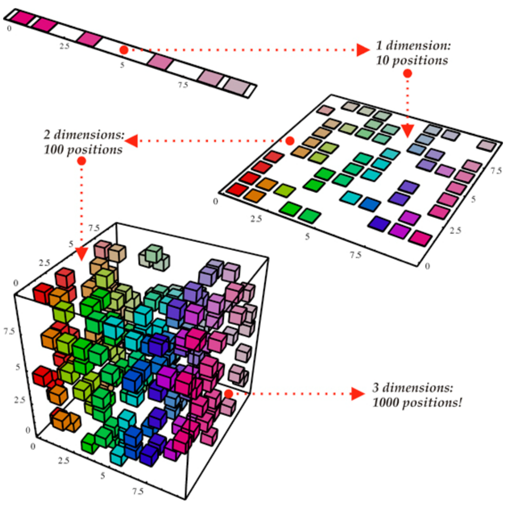
Problem: Too many features/dimensions

- Curse of Dimensionality
  - Huge data amount required for training
    - If we have D features with Q possible values per feature

    → Q^D probabilities need to be determined!

In order to maintain the same density of training data in the feature space, the data volume increases exponentially with dimension D, here ($Q = 10$):

- 1-dim: 10^1
- 2-dim: 10^2
- 3-dim: 10^3



The feature space: Summary

- Methods of statistical image analysis are differentiated according to
  - The way in which the clusters are determined
  - The parameters used to describe the clusters in feature space
  - The methods used to determining the parameters
  - The methods used for assigning a primitive to a particular class

- In principle, a larger amount of features could be considered to increase the prospects for a good classification result.

- However, the more features are used, the more training data are required.

  → One should avoid the use of heavily correlated features.

Contents

- Introduction to statistical methods in pattern recognition and image analysis
  - Tasks and solution strategies
  - The feature space
  - Taxonomy of statistical methods
  - Overfitting Problem

Taxonomy of statistical methods I

1. According to the image primitives that are classified:

- Pixel-based classification
- Segment-based classification (also referred to as object-based classification, which is a bit misleading)

2. According to the requirements w.r.t. training data:

- Supervised classification or supervised learning
- Unsupervised classification or unsupervised learning

3. According to the classification procedure:

- Individual classification of the image primitives: image primitives are considered to be independent
- Simultaneous classification of all image primitives: modelling of dependencies → Consideration of context

Taxonomy of statistical methods II

4. According to the type of the statistical model:

- Probabilistic methods: Classification on the basis of probabilities (or related concepts)
  - Generative methods: Based on a model of the joint distribution $p(C,\mathbf{x})$ of features and classes; synthetic data sets can be *generated* by appropriate sampling techniques.
  - Discriminative methods: Such methods directly model the posterior probability $p(C|\mathbf{x})$; it is not possible to generate synthetic data sets by sampling from $p(C|\mathbf{x})$.

- Non-probabilistic methods: Prediction of the class labels without modelling any probabilities; these methods are often referred to as discriminative classifiers as well.

Taxonomy of statistical methods III

5. According to the models used in probabilistic methods:

- Parametric techniques: Require assumptions about the distributions of the data and/or the classes; the parameters of the corresponding analytical functions for the probability densities are estimated from training data.
- Non-parametric techniques: No assumptions about distributions are made, but the probabilities are derived directly from training data.

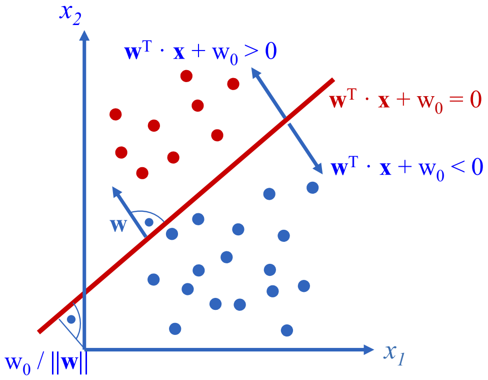
Contents

- Introduction to statistical methods in pattern recognition and image analysis
  - Tasks and solution strategies
  - The feature space
  - Taxonomy of statistical methods
  - Overfitting Problem

Discriminant function: Usually a linear function

- Often we strive to find a so-called discriminant function, which separates optimally the classes in feature space (geometric interpretation: hyper plane)
- The simplest and most common model is a linear combination of the input feature vectors $\mathbf{x}$ of dimension D:

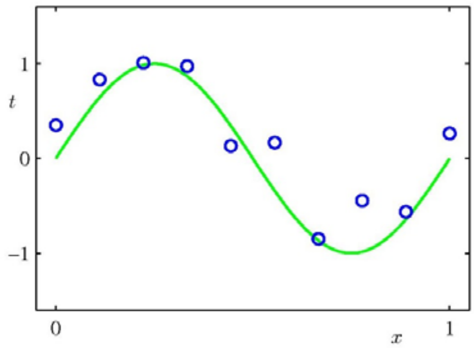
$$C(\mathbf{w},\mathbf{x}) = w_0 + w_1 x_1 + \ldots + w_D x_D = w_0 + \sum_{i=1}^{D} w_i x_i = w_0 + \mathbf{w}^T \cdot \mathbf{x}$$



Problem: Overfitting

- Consider the task to approximate a set of given data points $\mathbf{x}$ by some polynomial of degree M :

$$y(\mathbf{w},\mathbf{x}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{i=0}^{M} w_i x^i$$



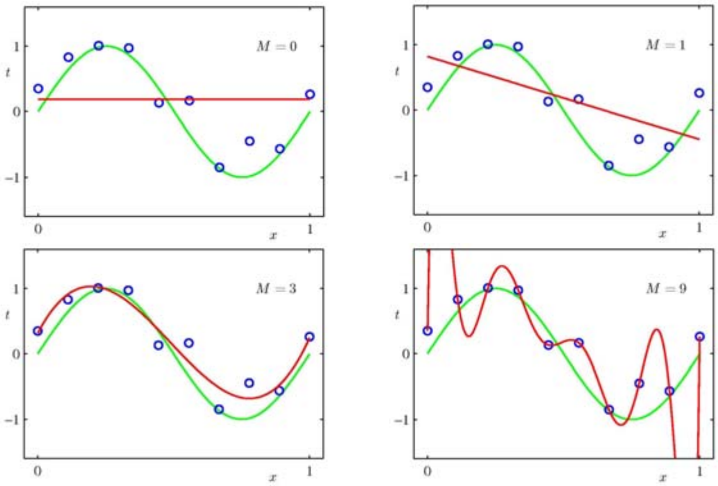
Example of $N = 10$ observations of input variable x along with the corresponding target variable t. The green curve shows the (unknown) function $sin(2\pi x)$ used to generate the data. Our goal is to predict the value of t for some new value of x.

Problem: Overfitting – Least squares as objective

- In case of least squares constraint the objective function is:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N}\left(y(x_n,\mathbf{w}) - t_n\right)^2 \qquad y(\mathbf{w},\mathbf{x}) = \sum_{i=0}^{M} w_i x^i$$

- We yield for different choices of M:



The solution $M=9$ yields minimal error according to least squares; unfortunately this is due undesired overfitting

Regularization: Idea

- Too tight approximation to **data** involves the danger of **overfitting**.
- Hence, we add a **model term** to the objective function, which prevents overfitting by enforcing some desired property of the optimal solution.
- This desired property depends on our purpose at hand, e.g. yield
  - As few as possible significant polygon coefficients (i.e., weights) of small magnitude
  - Preferably straight contours of roads
  - Preferably right-angled building footprints
  - …
- The hyper parameter λ weights these terms, it needs to be chosen carefully (often additional pre-training step like cross validation)

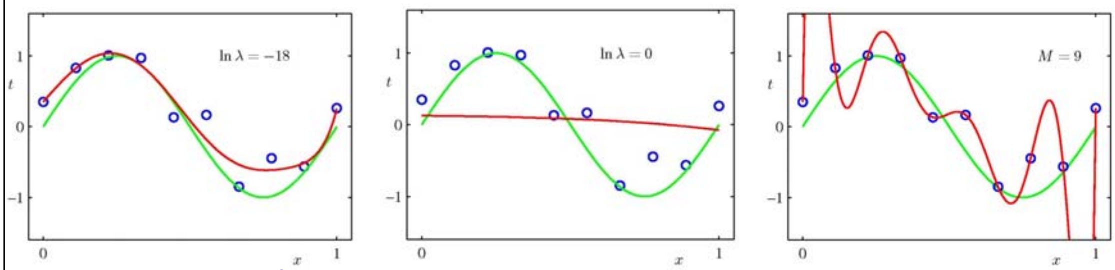
$$\tilde{E}(\mathbf{w}) = \text{data term} + \lambda \cdot \text{model term}$$

Problem: Overfitting – Regularization I

- We add a term that is typically chosen to impose a penalty on the complexity of $\mathbf{w}$:

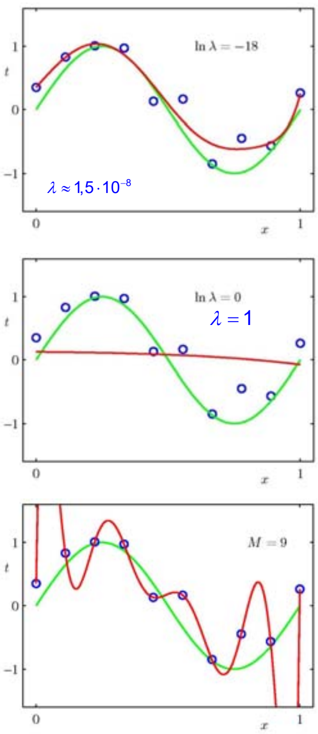
$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N}\left(y(x_n,\mathbf{w}) - t_n\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|$$

- Occam's Razor: "Select hypothesis with the fewest assumptions!"
- We yield for different choice of λ:



$\lambda \approx 1{,}5 \cdot 10^{-8}$ $\lambda = 1$ $\lambda = 0$

Problem: Overfitting – Regularization II



	$\lambda = 0$ $\ln\lambda = -\infty$	$\ln\lambda = -18$	$\lambda = 1$ $\ln\lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

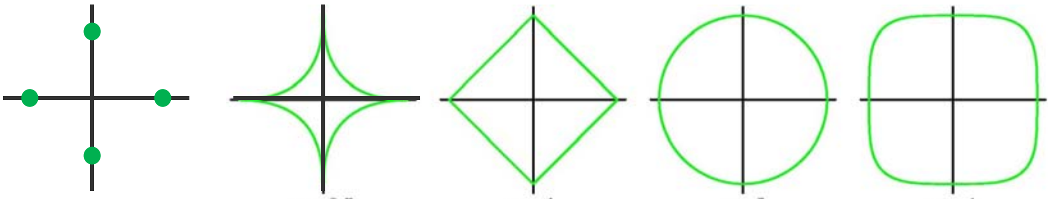
For $\lambda = 0$ (left in table) one yields very large values of coefficients, whereas for large λ small coefficients.

Problem: Overfitting – Regularization III

- In general: $$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N}\left(y(x_n,\mathbf{w}) - t_n\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{M}\left|w_j\right|^q$$

- The right term is often called *p-norm or L^p-norm* ($q=p$) $$\sum_{j=1}^{M}\left|w_j\right|^p \doteq \left\|\mathbf{w}\right\|_p$$

- We yield for different choice of p or q:



$q=0$ $q=0.5$ $q=1$ $q=2$ $q=4$

© Bishop, 2006