Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Topological Field Chunking with SVM

Martina Liepert

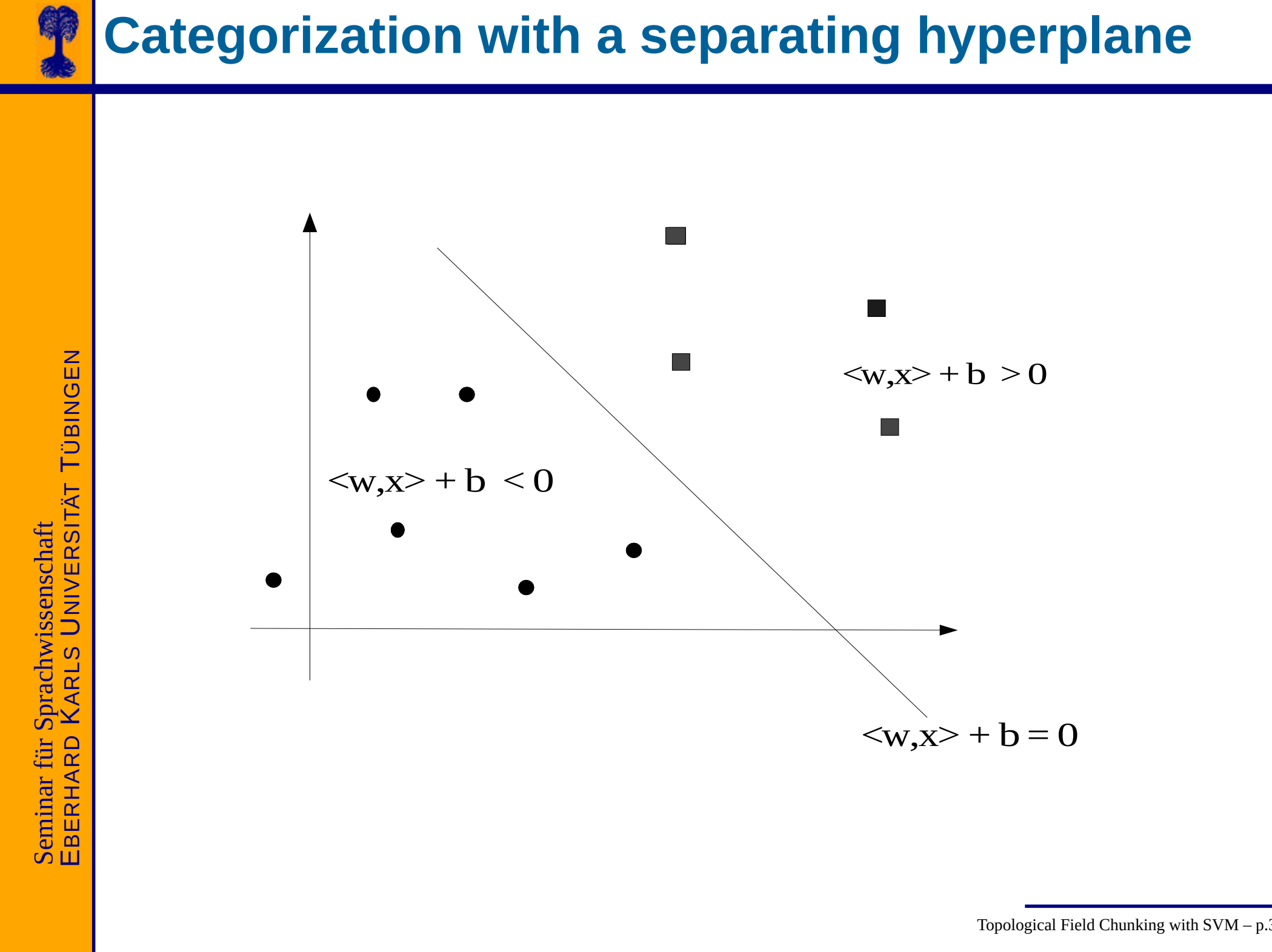
liepert@sfs.uni-tuebingen.de

Seminar für Sprachwissenschaft
Abteilung Computerlinguistik
Universität Tübingen

overview

- introduction to SVMs
- Topological field chunking with LibSVM

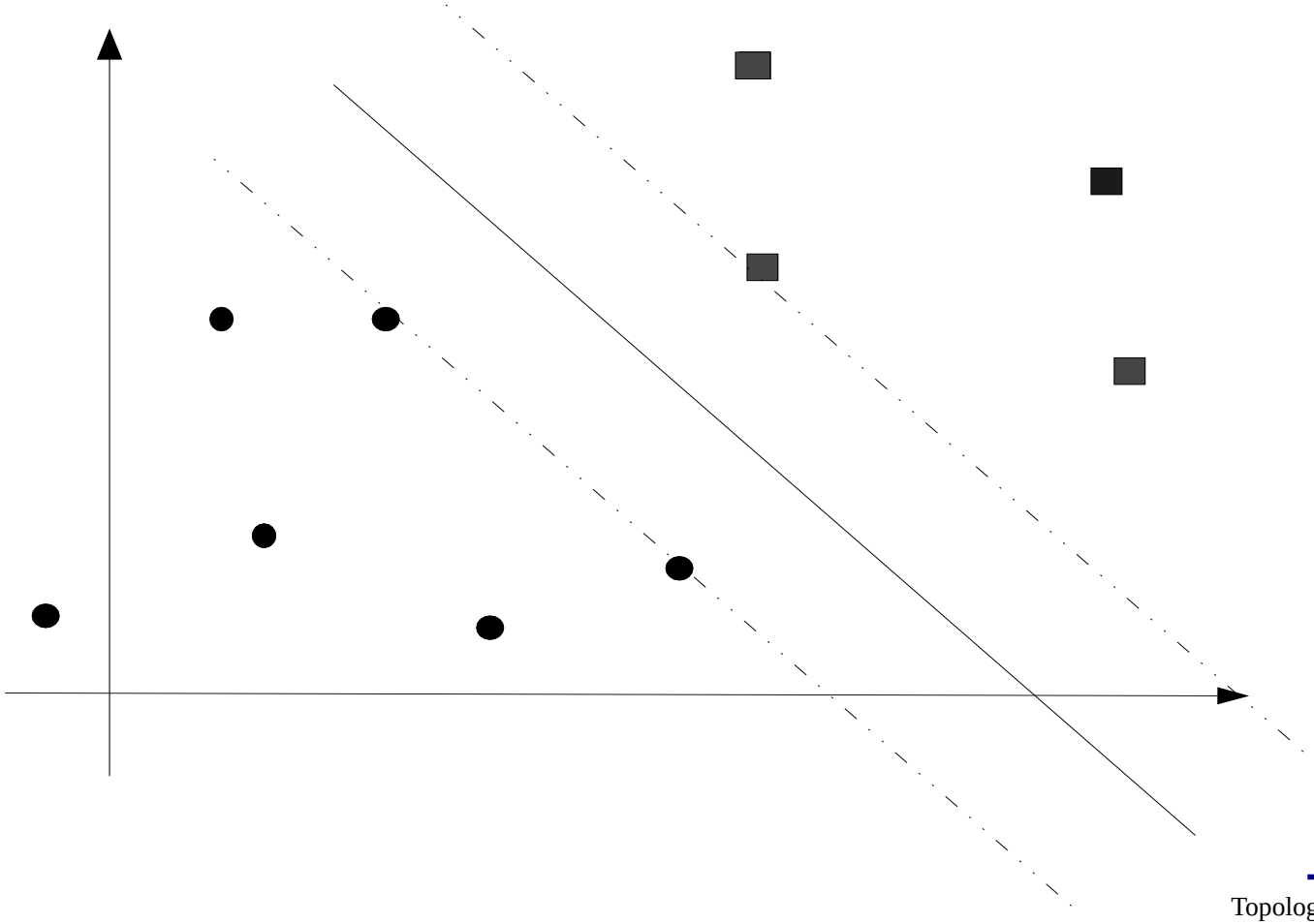
Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN



Categorization with a separating hyperplane

maximal margin

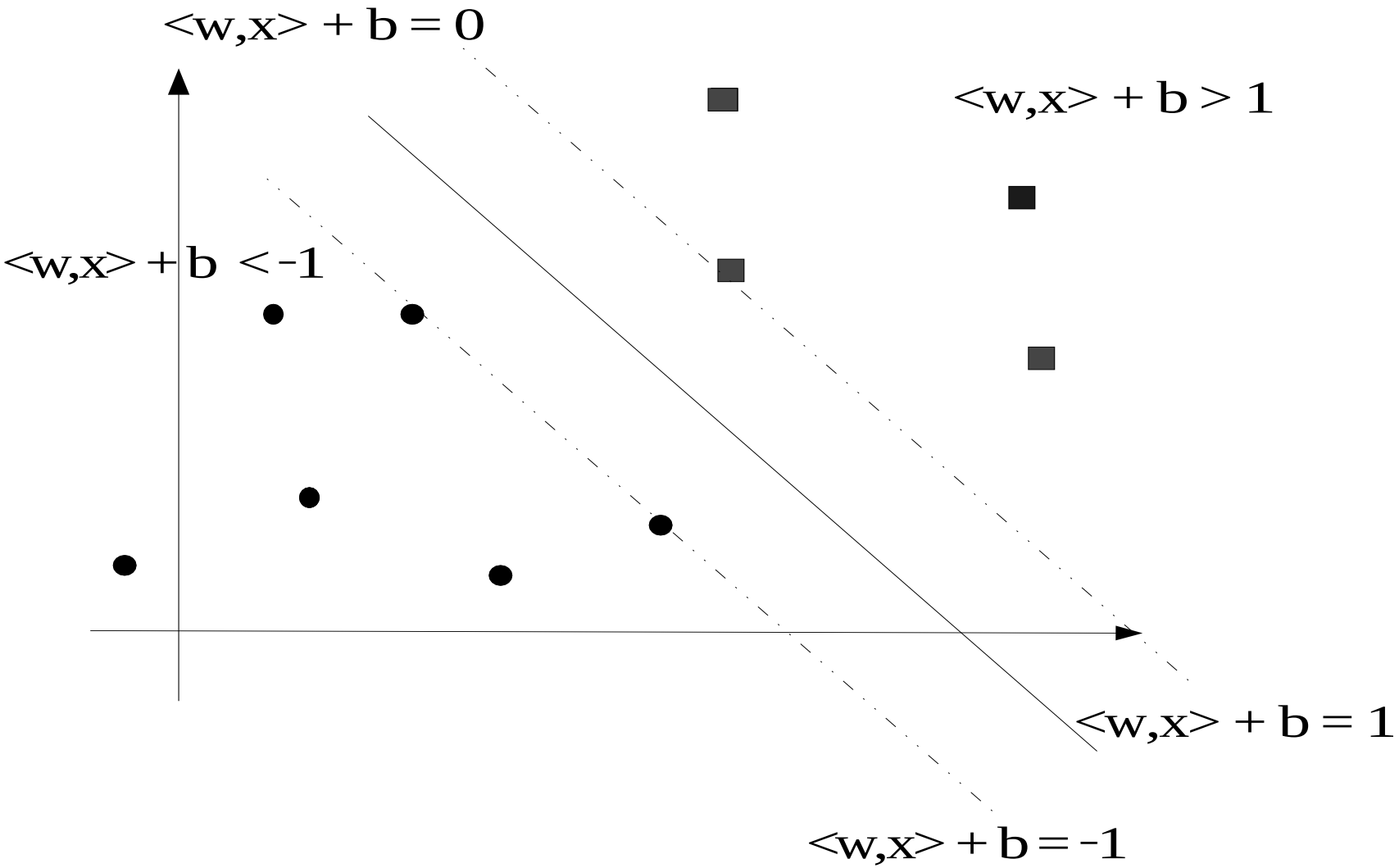
- margin: ’corridor’ around the hyperplane
- maximize the margin



Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

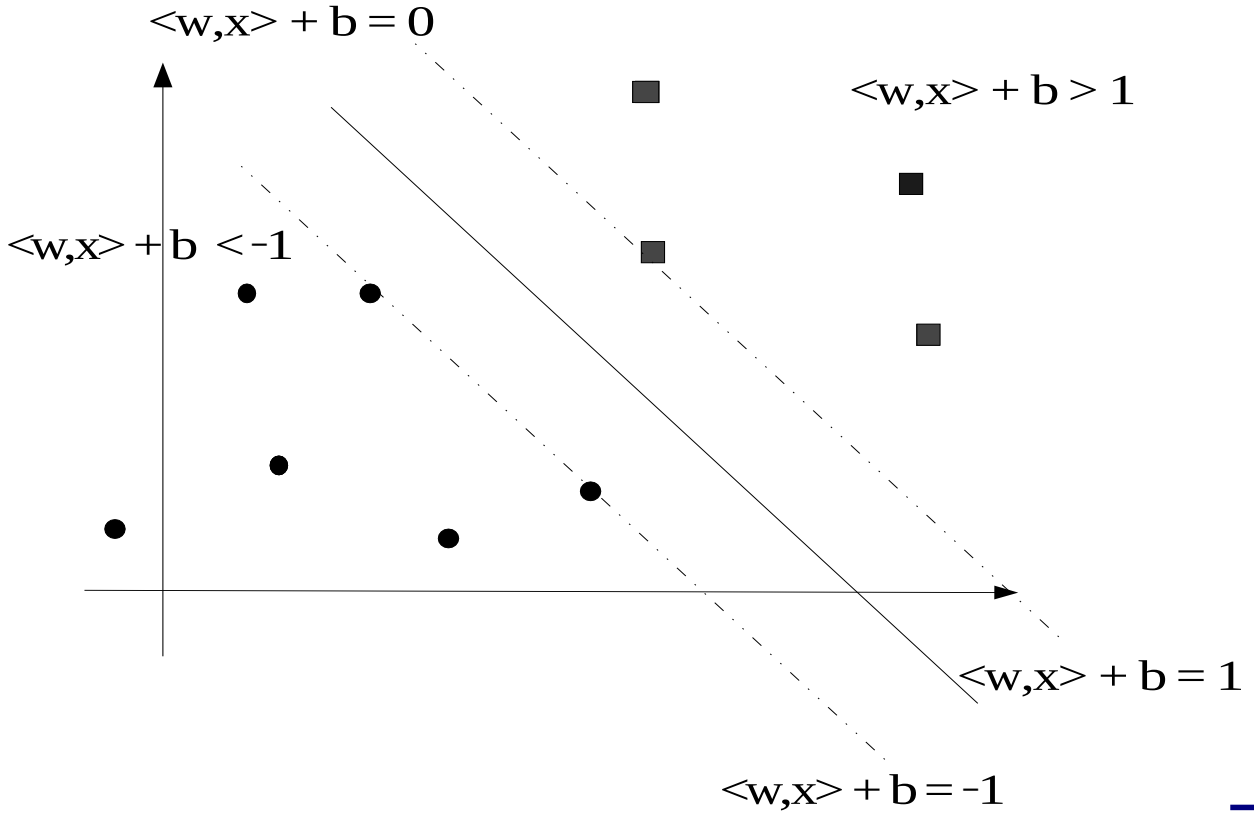
maximal margin

- scale w and b



maximal margin

- $y_i \langle w, x_i \rangle + b \geq 1$
- support vector: point on the margin
- margin = $\frac{2}{\|w\|}$

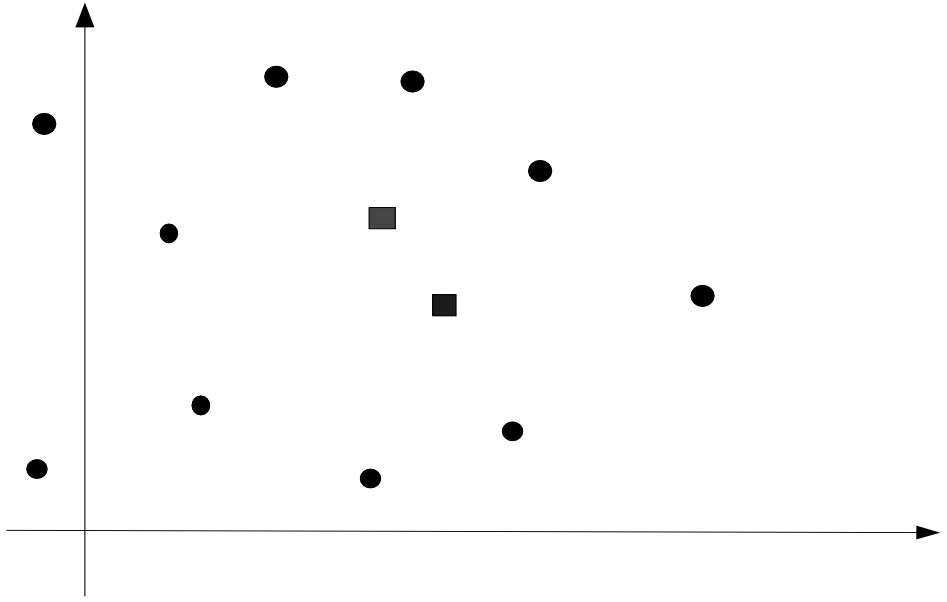


Quadratic Optimization Problem

- maximize the margin $\Longrightarrow$ maximize $\frac{2}{\|w\|}$
- minimize $\frac{1}{2}\|w\|^2$ $\qquad (\|w\| = \sqrt{\langle w, w \rangle})$
  subject to $y_i(\langle x_i \cdot w \rangle + b) \geq 1$
- **Langrangian**
  $L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_i \alpha_i (y_i(\langle x_i \cdot w \rangle + b) - 1)$
  $\Longrightarrow \sum_i \alpha_i y_i = 0 \qquad w = \sum_i \alpha_i y_i x_i$
  maximize the Lagrangian with respect to α_i and minimize it with respect to w and b.
- **dual form of the optimization problem**
  maximize $W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$
  subject to $\alpha_i \geq 0$ and $\sum_i \alpha_i y_i = 0$

Nonlinear Support Vector Classifier

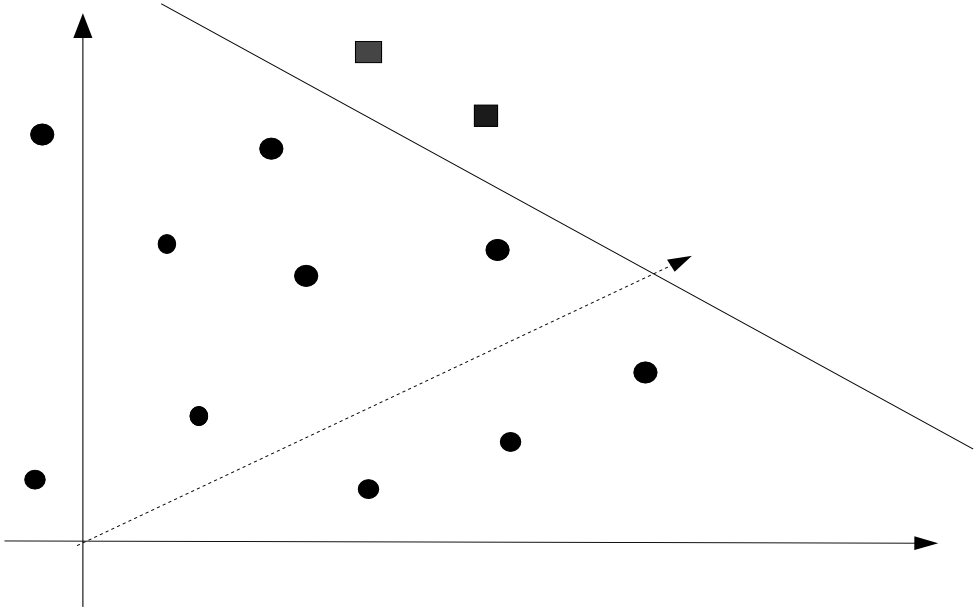
- **problem:** data not separable by a linear hyperplane



Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Nonlinear Support Vector Classifier

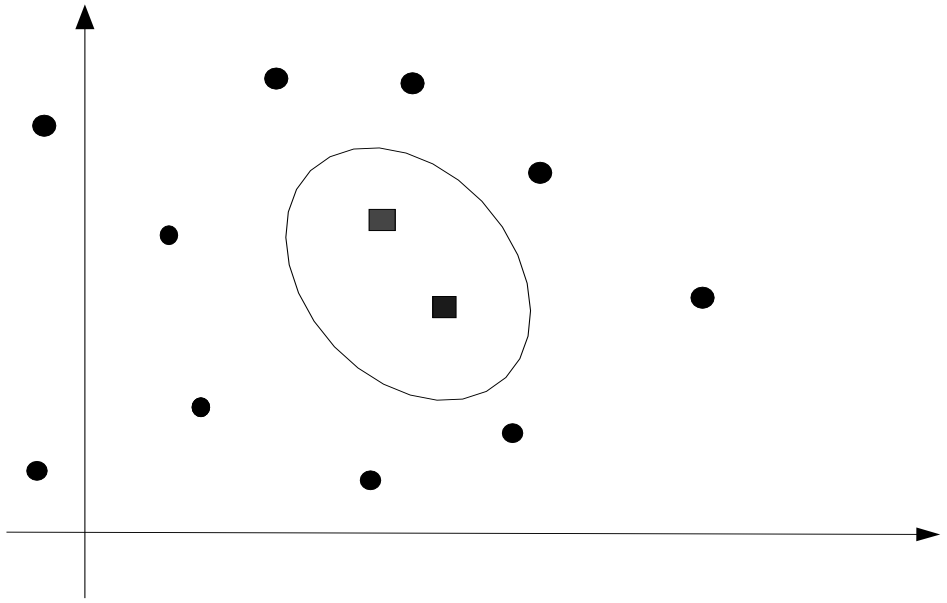
- **idea:** map input data in a higher dimensional feature space. $x \rightarrow \Phi(x)$



Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Nonlinear Support Vector Classifier

- this corresponds to a nonlinear classification in the input space



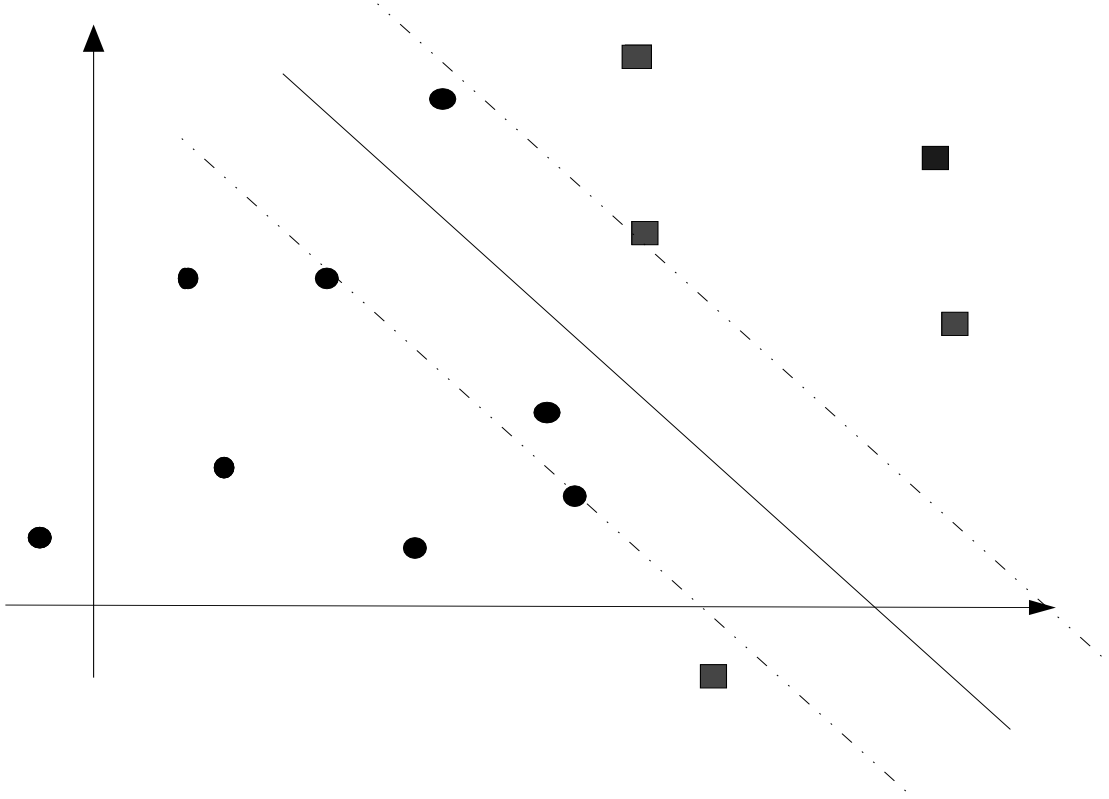
Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Nonlinear SVMs - Kernel based Classification

- **problem:** data not separable by a linear hyperplane
- **idea:** mapping into higher dimensional feature space: $x \rightarrow \Phi(x)$
  maximize
  $W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \Phi(x_i), \Phi(x_j) \rangle$
- **kernel function**
  calculate directly $k(x, x_i) = \langle \Phi(x_i), \Phi(x_j) \rangle$
- **common kernels**
  polynomial kernel: $k(x, x_i) = (x \cdot x_i + b)^d$
  Gaussian kernel (radial basis function):
  $$exp(\frac{-\|x - x_i\|^2}{c})$$
  tanh activation function: $tanh(\kappa \langle x, x_i \rangle + \Theta)$

Soft Margin Classifier

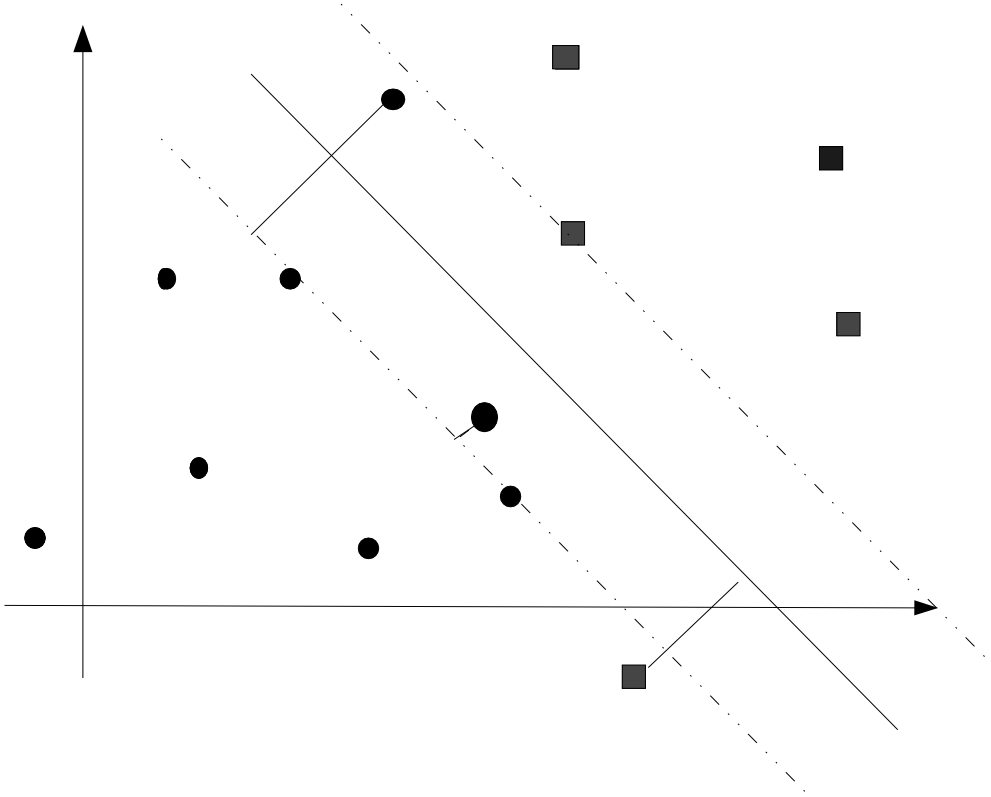
- **problem:** Classes still overlap in the high dimensional feature space
- single outliers (e.g. noise) should not have too much influence



Soft Margin Classifier

- introduce **slack variable** ε
- allow for a relaxed separation constraint

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \varepsilon_i$$



Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Soft Margin Classifier - C-SVC and nu-SVC

- introduce slack variable ε
- penalize ε_i (i.e. the training error) with constant C
- minimize $\tau(w, \varepsilon) = \frac{1}{2}\|w\|^2 + C\sum_i \varepsilon_i$
- **C-SVC**: choose C between e.g. 10^{-4} and 10^4
- **nu-SVM**: equivalent to C-SVC; choose $\nu \in [0, 1]$

Unbalanced Data

- there's no a priori rule to choose an optimal C (or ν) for a given problem
- in practice the optimal kernel parameters and C are often chosen by using cross validation
- unbalanced data = asymmetric number of examples for each of the two classes
- **problem:** unbalanced data leads to a bias of the learner towards the more frequent class
- **solution:** choose different Cs for the two classes; rule of thumb: antiproportional to the number of examples
- minimize $\frac{1}{2}||w||^2 + C^+ \sum_i \varepsilon_i + C^- \sum_i \varepsilon_i$

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Multi-class Classification

- SVMs are binary classifiers
- to get M-class classifiers: construct a set of binary classifiers and combine them

1. **one-versus-rest:**
   - Given m classes. Construct m classifiers which separate class m_i from the rest.
   - category 1 versus not-1; category 2 versus not-2, ..., category m versus not-m
   - 'winner takes all': choose the solution of classifier i with maximal output $(max_{j=1,..m} \sum_i y_i \alpha_i^j k(x_i, x) + b)$

Multi-class Classification

1. **one-versus-rest**
2. **one-versus-one**
   - train a classifier for each possible pair of m classes
   - 1 vs 2, 1 vs 3, ... 1 vs m, 2 vs 3, ... 2 vs m, ... m-1 vs m
   - we have to train $\frac{m(m-1)}{2}$ classifiers
   - combine classifiers by majority vote
   - alternative combination: use a directed acyclic graph (DAG)
   - one-versus-one is (only) slightly better then one-versus-rest
   - disadvantage: a lot of training time necessary

Applicaton: Topological field chunking

- find left and right sentence bracket (verb complex), C-field
- task performed by Jorn (Timble), Tylman (PCFG) and Frank (rule based) and presented at the CONLL 2002

Example:
[$_C$ Bevor] die Kommission grünes Licht [$_{VC}$ geben wollte], [$_{LK}$ sollte] ein gesellschaftlicher Konsens über derartige Experimente [$_{VC}$ herbeigeführt werden].

Topological field chunking with SVM

Characteristics of the learning task

- 5000 sentences, ca. 90 000 learning points. Very many for SVM; long training time.
- very unbalanced data (e.g. O: 80 000, I-VC: 7000, I-C: 2000)
- 4 target categories, multi-class problem
- LibSVM

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Features and Kernel

- LibSVM demands numerical features
- look at a window of the two preceeding words and the following word
- first try with POS and words:
  lexicon with 20 000 entries * 8 attributes = 160 000 features
  LibSVM couldn’t separate the classes
- second try with POS:
  lexicon with 55 entries * 4 attributes = 220 features
- rbf-Kernel

Multiclass Classification with unbalanced data

- libSVM provides multi-class (one-vs-one) classification, but uses the same feature values for all binary classifiers
- my approach: one-vs-one-multiclass classification with Majority vote where the C-values can be adjusted for each class and classifier
- start with C-values antiproportional to the number of training points in the classifier
- adjust C-values heuristically depending on number of false negatives and false positives in the overall classifier or by brute force (trying several values, choose the best combination)

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Results

multiclass, rbf-kernel, TnT-POS

	ALL	LK	VC	C
LibSVM MC	91.61	95.18	90.65	83.33
my MC C-antiprop	86.76	95.41	90.77	61.39
my MC C-best	93.13	95.59	91.34	90.61
my MC C-antiprop gold	95.51	97.51	96.12	91.31

- polynomial kernel with the best C-values for rbf-kernel: SVM couldn't separate the classes
- features have to be re-adjusted for every kernel, if gold instead of TnT-tags are used ...

comparison of the results

	ALL	LK	VC	C
FSA TnT	94.1	96.2	92.0	93.8
FSA gold	98.4	98.8	98.3	97.5
PCFG TnT	94.4	97.0	92.2	92.3
PCFG gold	98.1	98.9	98.1	96.1
MBL TnT	93.3	96.0	90.0	91.6
MBL gold	97.2	98.0	96.7	96.6
SVM TnT	93.13	95.59	91.34	90.61
SVM gold	95.51	97.51	96.12	91.31

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

feature combinations

- features to be chosen:
  - kernel
  - kernel features:
    polynomial kernel: b, gamma, degree
    rbf kernel: gamma
  - C-value(s) or nu
- polynomial kernel for TopF:
  3 values for degree, 5 values for gamma and b and 5 values for C^+ and C^- in each binary classifier: 3* $5^9 = 5.8$ Mio possible feature combinations
- finding the optimal feature combinations becomes the major problem when classifying with SVMs