Karlsruher Institut für Technologie

Maschinelles Lernen II - Fortgeschrittene Verfahren
V03 Semi – überwachtes Lernen
Semi Supervised Learning (SSL)

Sommersemester 2017

Prof. Dr. J.M. Zöllner, Prof. Dr. R. Dillmann

INSTITUT FÜR ANGEWANDTE INFORMATIK UND FORMALE BESCHREIBUNGSVERFAHREN
INSTITUT FÜR ANTHROPOMATIK UND ROBOTIK

Grundparadigmen

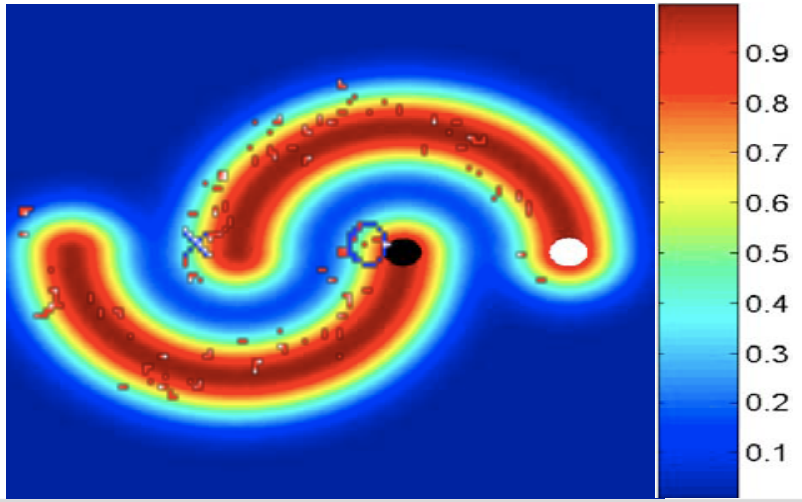
- Überwachtes Lernen
  - Gelabelte Trainingsdaten: Paare (X,Y)
  - Finde eine Funktion h die X (Merkmalsraum) auf Y abbildet (z.B. Klassen)
- Unüberwachtes Lernen
  - Ungelabelte Daten aus dem Merkmalsraum X
  - Strukturen und Labels der Daten (z.B. durch Cluster-Verfahren) finden
  - Oft auch Dichte-(Träger)-Schätzung (siehe ML I – SVM)

- Semi-Überwachtes Lernen
  - Einige, aber meist wenige gelabelte Lerndaten
  - Viele ungelabelte Daten
  - Finde eine Funktion h die X (Merkmalsraum) auf Y abbildet (z.B. Klassen)

Grundannahmen

- Typische Annahmen für das Lernen
  - Gleichmäßigkeit für überwachtes Lernen (Smoothness Assumption):
    - Wenn zwei Datenpunkte x1, x2 „nahe“ beieinander sind dann sollten auch die Ausgaben y1, y2 „ähnlich“ sein
  - Gleichmäßigkeit für Semi-überwachtes Lernen:
    - Wenn zwei Datenpunkte x1, x2 in einer dichten Region „nahe“ beieinander sind, dann sollten auch die Ausgaben y1, y2 „ähnlich“ sein

➔ wenn zwei Datenpunkte
durch einen Pfad hoher Dichte verbunden
sind (i.A. gehören dem gleichen Cluster an)
dann sind ihre Ausgaben ähnlich

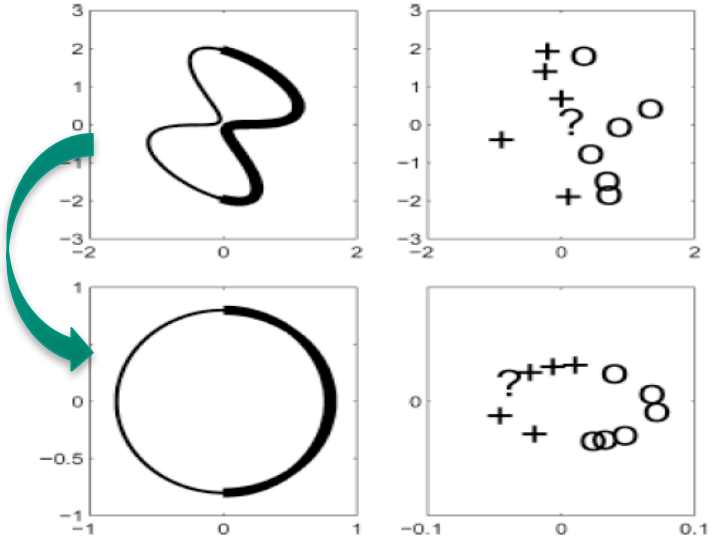


Abgeleitete Grundannahmen

- Cluster oder Dichte Annahme
  - Wenn zwei Datenpunkte im selben („dichten“) Cluster sind, dann sind sie in derselben Klasse
  - ➔ eine Trennung sollte in einer Region niedriger Dichte (zw. den Clustern) liegen

- Manigfaltigkeit-Annahme (Manifold Assumption)
  - Hochdimensionale Daten haben eine Abbildung in einen i.A. anders dimensionalen Raum (Manigfaltigkeitsraum) in dem sich ihre Strukturen abbilden (unterscheiden/ erhalten)

  - ➔ Dieser Raum kann dann für die Berechnung des geodäsischen Abstand benutzt
  - ➔ approximative Implementierung der Gleichmäßigkeitsannahme



Formalisierung

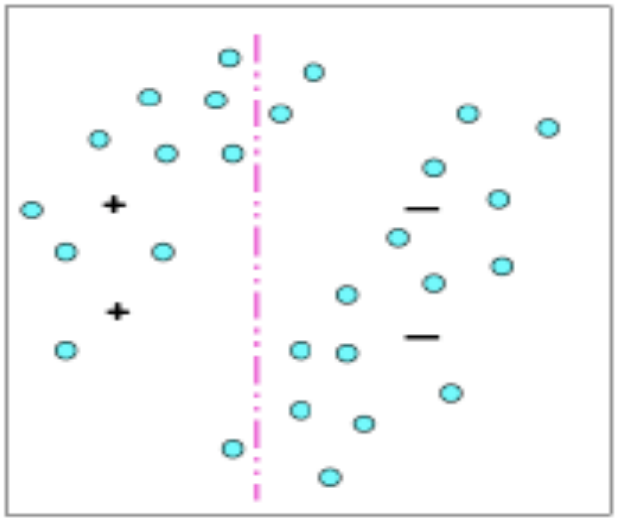
- Instanzen (Feature – Vektor): $x \in X$
  label: $y \in Y$
- Hypothese: $h: \; X \rightarrow Y$
- Gelabelte Daten $(X_l, Y_l) = \{(x_{1:l}, y_{1:l})\}$
- Ungelabelte Daten: $X_u = \{x_{l+1:n}\}$
  vorhanden beim Trainieren
- Üblicherweise gilt: $l << n$
- Neue Daten: $X_{test} = \{x_{n+1:....}\}$
  nicht vorhanden beim Trainieren

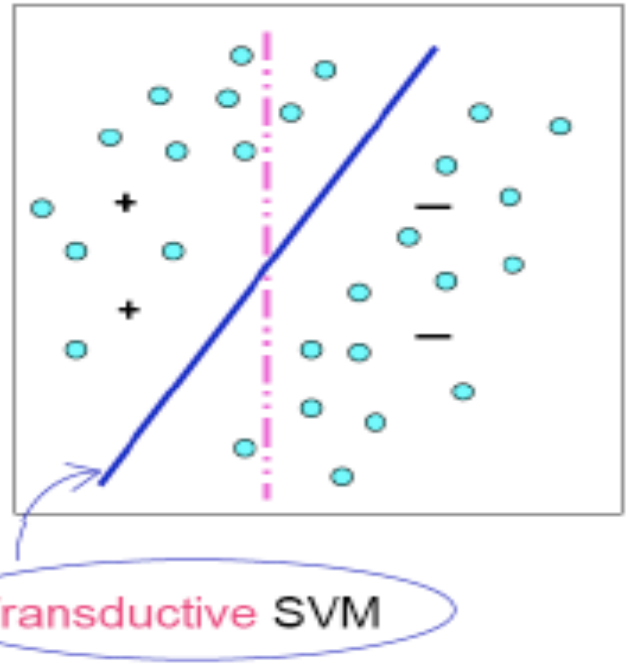
Verschieden Ansätze

- Erste Algorithmen
  - Self-Training & Co-Training
- Generative probabilistische Modelle (Generative Probabilistic Models)
  - EM for Gaussian Mixtures
- Dichte Trennung (Low-Density Separation)
  - „Transduktive“ SVM
- Graph basierte Modelle / Methoden
  - Methoden bei denen die Daten als Knoten eines Graphs repräsentiert sind und die Kanten die jeweiligen Abstände enthalten
- Änderung der Repräsentation
  - unüberwachtes Lernen (z.B.: Clustern) um neue (i.A. niedrig dimensionale) Repräsentationen der Daten zu erhalten
  - Lernen der Zuordnung der Cluster zu Klassen

Dichte Trennung (Low – density separation) mit SVM

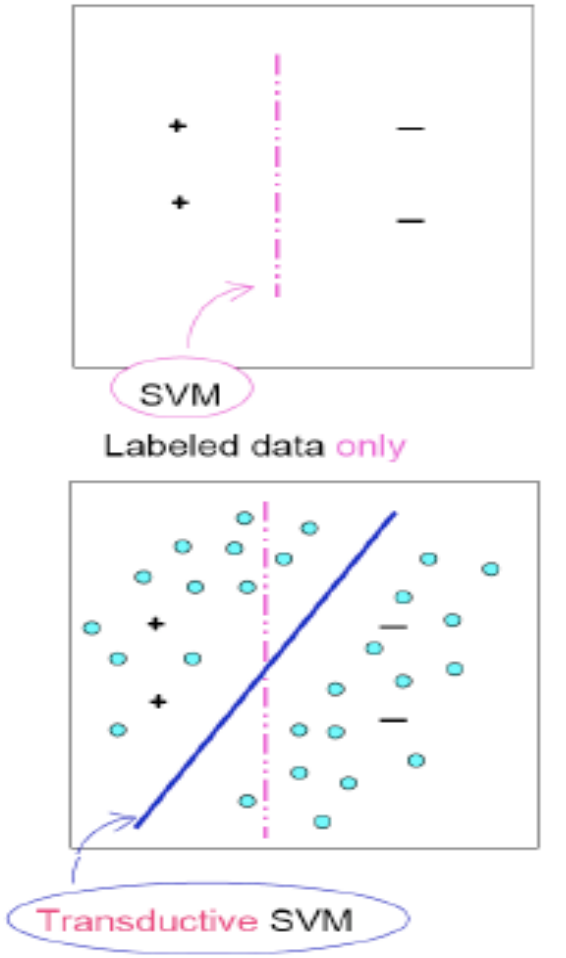
- Ziel





Transduktive SVM

- Annahme
  - Ungelabelte Daten unterschiedlicher Klassen werden mit großem Rand getrennt – aber wie?
- Naiver Ansatz
  - Alle 2^u Möglichkeiten der Labels v. $X_u = \{x_{l+1:l+u}\}$ betrachten
  - Trainiere SVM für alle Möglichkeiten
  - Wähle SVM mit größtem Rand
- Besser
  - Integriere ungelabelte Daten in das Optimierungsproblem



Diskriminative Modelle - Fehler, Risiko, Kosten

- Definiere eine Verlustfunktion oder Kostenfunktion (loss-function, cost-function)

$$L(y, f(\mathbf{x}; \mathbf{w}))$$

- Bestimme empirisch die Kosten (Fehler, Risiko) als

$$R_{\mathrm{emp}}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(\mathbf{x}_i; \mathbf{w}))$$

- Z.B: mit einer quadratischen Kostenfunktion

$$R_{\mathrm{emp}}(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i; \mathbf{w}))^2$$

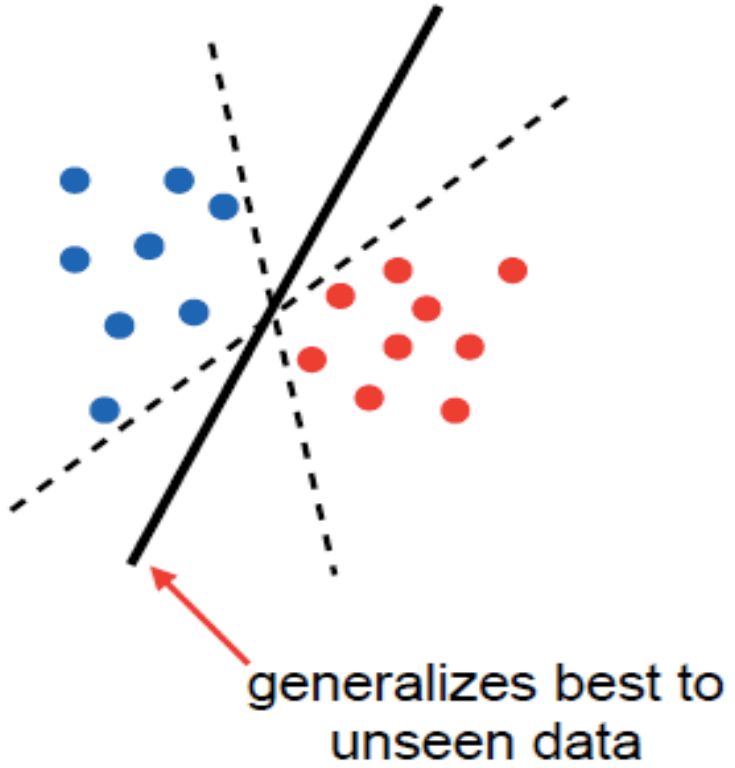
Realer Fehler, Risiko, Kosten

- Von Interesse ist der reale Fehler

$$R(\mathbf{w}) = \int L(y, f(\mathbf{x};\, \mathbf{w}))p(\mathbf{x}, y)\ \mathrm{d}\mathbf{x}\mathrm{d}y$$

  - mit $p(\mathbf{x}, y)$ der realen Verteilung der Daten
  - „Berechnung“ auf allen Daten

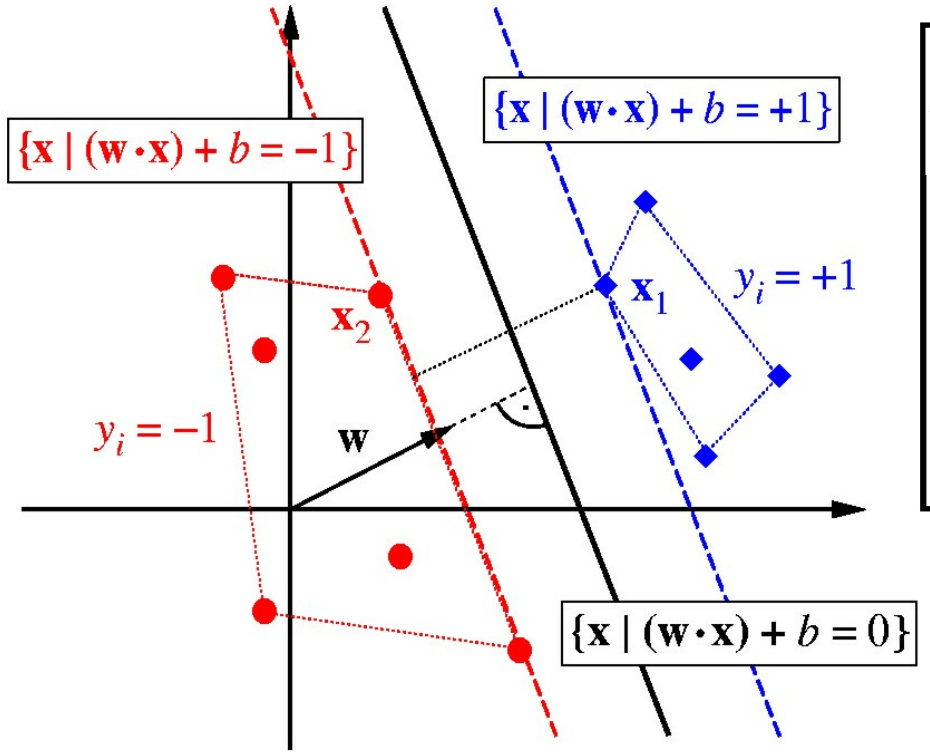
  - $p(\mathbf{x}, y)$ ist fest aber unbekannt
  - der reale Fehler kann nicht berechnet werden
    ➔ eine gute Schätzung ist nötig



Bsp. Empirische Fehler ist in allen Fällen = 0

SV – Methode, Trennebene

Finde die Hyperebene $\{\vec{x} \in S : \vec{w}\vec{x} + b = 0, (\vec{w}, b) \in S \times R\}$



Abstand $\frac{\mathbf{w} \cdot (\mathbf{x}_1 - \mathbf{x}_2)}{\|\mathbf{w}\|}$

$(\mathbf{w} \cdot \mathbf{x}_1) + b = +1$

$(\mathbf{w} \cdot \mathbf{x}_2) + b = -1$

$\Rightarrow \quad (\mathbf{w} \cdot (\mathbf{x}_1 - \mathbf{x}_2)) = 2$

$\Rightarrow \left(\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot (\mathbf{x}_1 - \mathbf{x}_2) \right) = \frac{2}{\|\mathbf{w}\|}$

SV – Methode, Formalisierung

Bedingung für die optimale Hyperebene

$$\min_{i=1\ldots n}\left|\vec{w}\vec{x}_i + b\right| = 1, \quad x_i - Lerndaten$$

➔ Der nächste Punkt hat den Abstand $\frac{1}{\|\vec{w}\|}$

➔ Der Abstand zw. den 2 Klassen $\frac{2}{\|\vec{w}\|}$

➔Die Entscheidungsfunktion eines Hyperebenen - Klassifikators

$$f_{\vec{w},b}(\vec{z}) = sign(\vec{w}\vec{z} + b)$$

SV – Optimierung, Problem

Hyperebene mit maximalem Abstand (margin)

Maximiere $\frac{2}{\|\vec{w}\|}$ ➔ Minimiere $\|\vec{w}\|^2$ $oder$ $\frac{1}{2}\|\vec{w}\|^2$

Unter den Bedingungen

$$y_i\left(\vec{w}\vec{x}_i + b\right) \geq 1 \quad i = 1 \ldots n$$

d.h. die Daten werden korrekt klassifiziert

➔ Laut Vapnik die Lernmaschine mit der kleinsten möglichen VC-Dimension (falls die Klassen linear trennbar sind)

➔ Kleinste realer Fehler

Minimierung: Lagrange - Methode

Äquivalentes Problem (*Primäres Optimierungsproblem*):

Finde den (eindeutigen) Sattelpunkt der Funktion

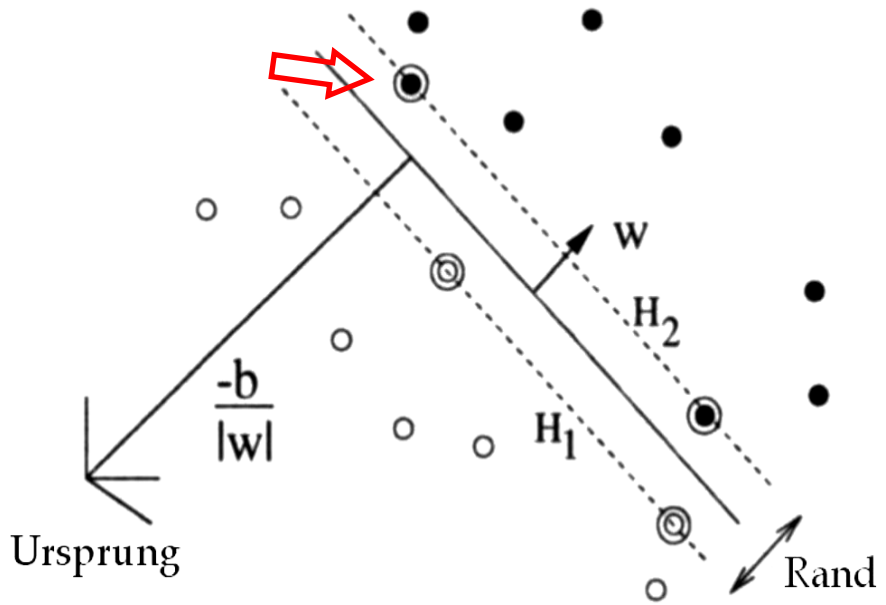
$$L_P = L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{N} \alpha_i (y_i(\vec{w}\vec{x}_i + b) - 1)$$

und $\alpha_1, \alpha_2, ..., \alpha_N \geq 0$ (Lagrange-Multiplikatoren)

Support Vektoren

- Die meisten Bed. sind erfüllt

  ➔ die meisten $\alpha_i = 0$ (Sattelpunkt-Bedingung)



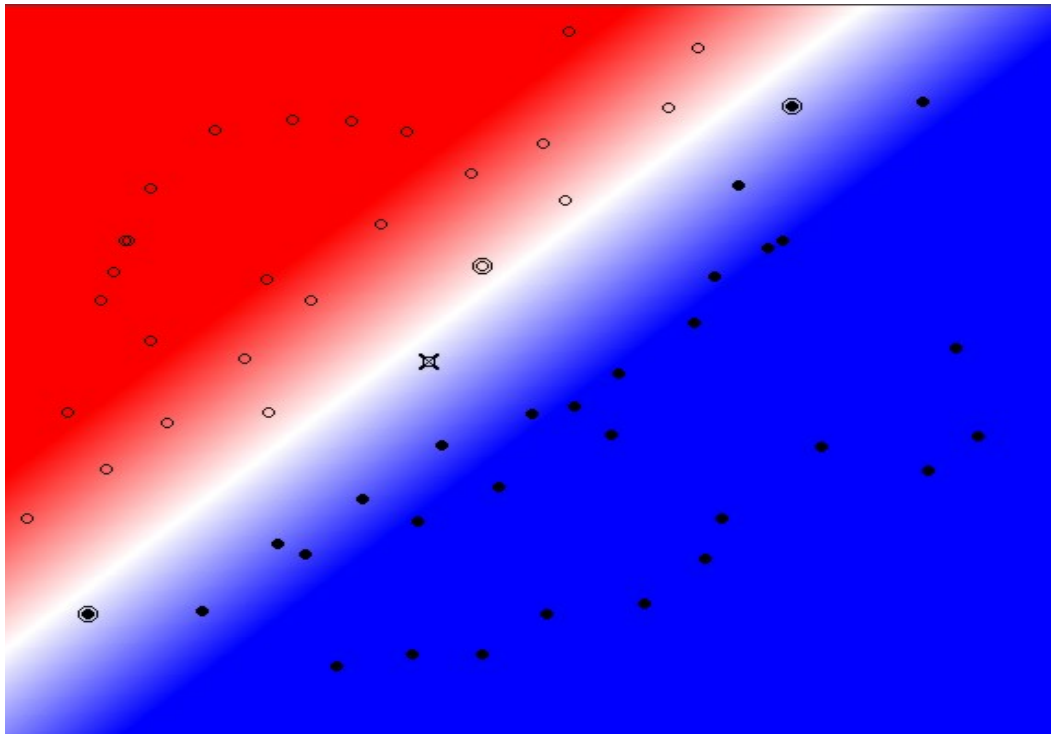
- Support-Vektoren: $\vec{x}_i$ mit $\alpha_i > 0$

- $\vec{w}$ ist eine Linearkombination weniger Vektoren $\vec{x}_i$ (Support Vektoren)

$$\vec{w} = \sum_{i=1}^{N} \alpha_i y_i \vec{x}_i$$

Support Vektoren liegen am nächsten zur Trennebene

Soft Margin - Hyperebene



Idee:
Erlaube eine geringe Zahl von Miss-klassifikationen
➔ Höhere Generalisierung

Änderung der Randbedingungen durch Schlupf-Variablen (slack variables):

$$y_i\left(\vec{w}\vec{x}_i + b\right) \geq 1 - \xi_i \quad i = 1 \ldots n, \xi_i \geq 0$$

Generalisierte optimale Hyperebene

Minimiere

$$\min_{\vec{w},b,\xi_i} \frac{1}{2}\|\vec{w}\|^2 + C\left(\sum_{i=1}^{l}\xi_i\right)^p$$

Bedingungen

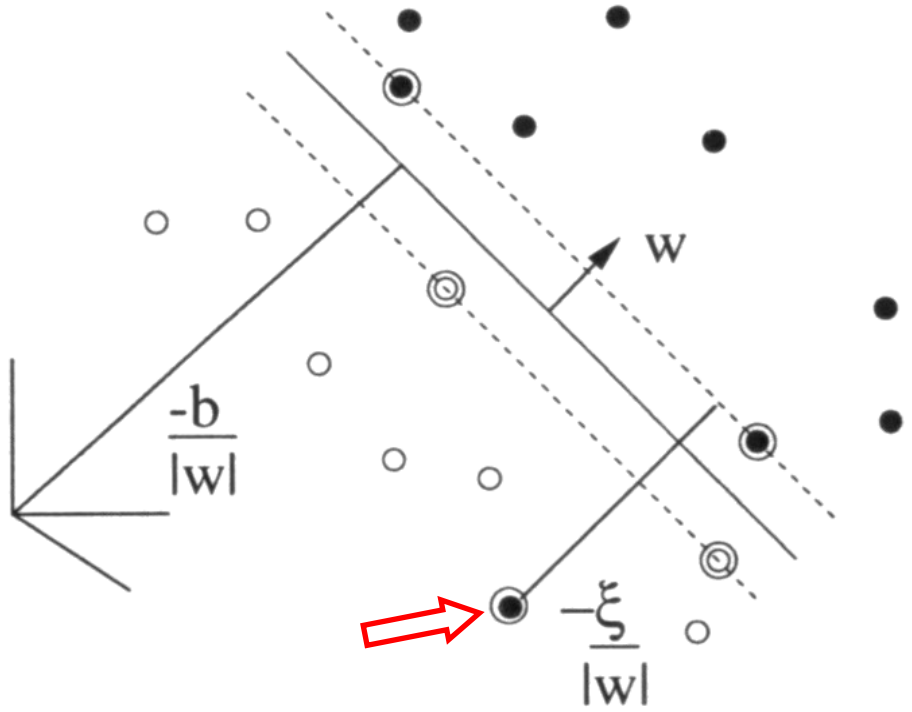
$$y_i\left(\vec{w}\vec{x}_i + b\right) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

Lösung: Lagrange - Methode

Rolle von *C*

- Regulierungsparameter
- *C* – groß ➔ wenig Missklassifikationen
- *C* – klein ➔ maximale Margins



Hinge Funktion

- Gegeben die Maximumsfunktion (hinge function):

$$(\xi_i)_+ = \max(\xi_i, 0)$$

- Gilt:

$$y_i\left(\vec{w}\vec{x}_i + b\right) \geq 1 - \xi_i \Leftrightarrow$$

$$\left.\begin{aligned} \xi_i &\geq \left(1 - y_i\left(\vec{w}\vec{x}_i + b\right)\right) \\ \xi_i &\geq 0 \end{aligned}\right\} \Rightarrow \xi_i \geq \left(1 - y_i\left(\vec{w}\vec{x}_i + b\right)\right)_+$$

- Äquivalentes Optimierungsproblem

$$\min_{\vec{w},b} \frac{1}{2}\left\|\vec{w}\right\|^2 + C\left(\sum_{i=1}^{l}\left(1 - y_i\left(\vec{w}\vec{x}_i + b\right)\right)_+\right)^p$$

Hinge Funktion

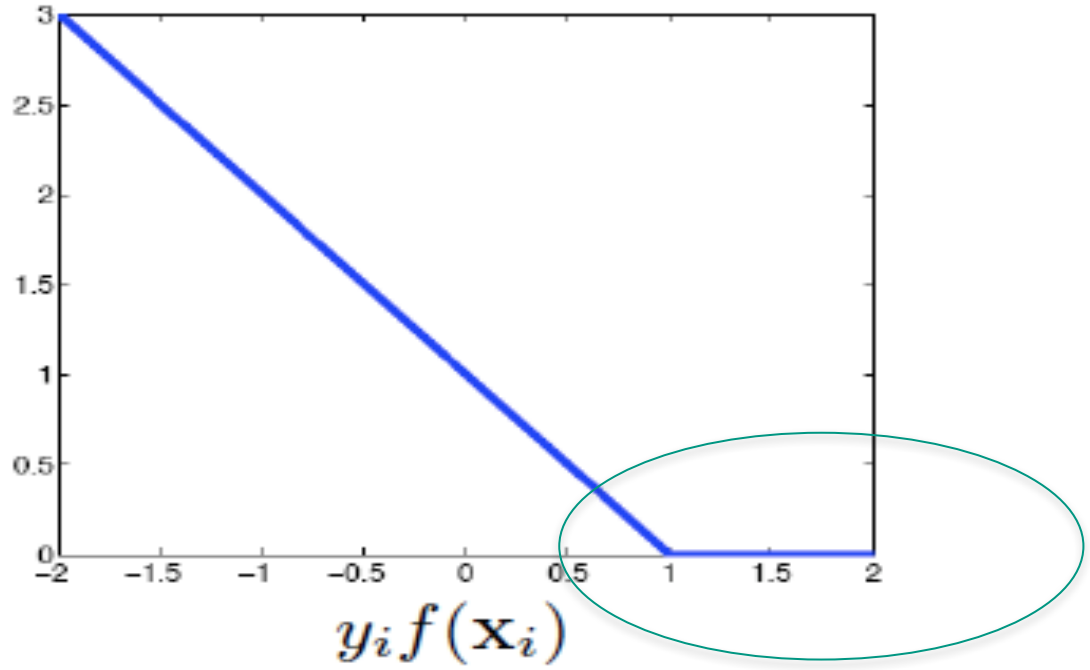
- Minimierungsproblem:

$$\min_{\vec{w},b} \frac{1}{2}\|\vec{w}\|^2 + C\left(\sum_{i=1}^{l}\left(1 - y_i\left(\vec{w}\vec{x}_i + b\right)\right)_+\right)$$

- Bevorzugt gelabelte Daten auf der „richtigen“ Seite der Trennhyperebene:

$$(1 - y_i f(\mathbf{x}_i))_+$$

$$f(\vec{x}_i) = \vec{w}\vec{x}_i + b$$



„Transduktive“ SVM - Anpassung

- Einbinden der ungelabelten Daten:
  - Problem: y_i unbekannt für ungelabelte Daten
  - Für korrekte Labels würde gelten: $y_i = \text{sign} f(\mathbf{x}_i) \text{ for } \mathbf{x}_i \in X_u$
  - Hinge – Funktion für ungelabelte Daten:

$$(1 - y_i f(\mathbf{x}_i))_+ = (1 - |f(\mathbf{x}_i)|)_+$$

    - Weil:

$$\text{sign}(f(\mathbf{x}_i)) f(\mathbf{x}_i) = |f(\mathbf{x}_i)|$$

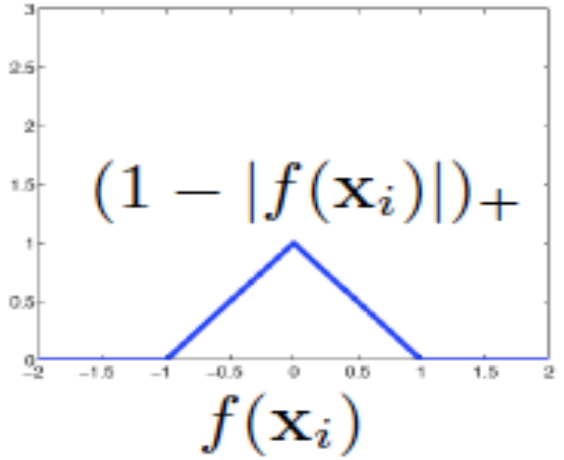
  - Minimierungsproblem kann erweitert werden:

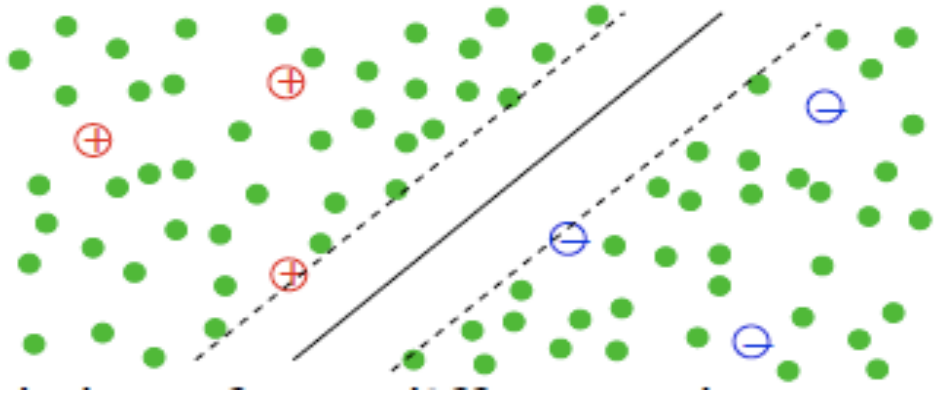
$$\min_{\mathbf{w},b} \left\{ \frac{1}{2}||\mathbf{w}||^2 + C_1 \sum_{i=1}^{l} (1 - y_i f(\mathbf{x}_i))_+ + C_2 \sum_{i=l+1}^{n} (1 - |f(\mathbf{x}_i)|)_+ \right\}$$

Bedeutung – niedrige Dichte

$$\min_{\mathbf{w},b}\left\{\frac{1}{2}||\mathbf{w}||^2 + C_1\sum_{i=1}^{l}(1-y_i f(\mathbf{x}_i))_+ + C_2\sum_{i=l+1}^{n}(1-|f(\mathbf{x}_i)|)_+\right\}$$

- Letzter Term: $(1-|f(\mathbf{x}_i)|)_+$
  - Bevorzugt $f(\mathbf{x}_i) \geq 1 \quad f(\mathbf{x}_i) \leq -1$
  - Insbesondere Daten die nicht nahe an der Entscheidungsfunktion mit $f(x) = 0$ liegen
  - bevorzugt Daten außerhalb des Randes
  - Äquivalent damit, dass die Trennung $f(x) = 0$ in einer Region ohne gelabelte und ungelabelte Daten liegt (➔ **niedrige Dichte**)

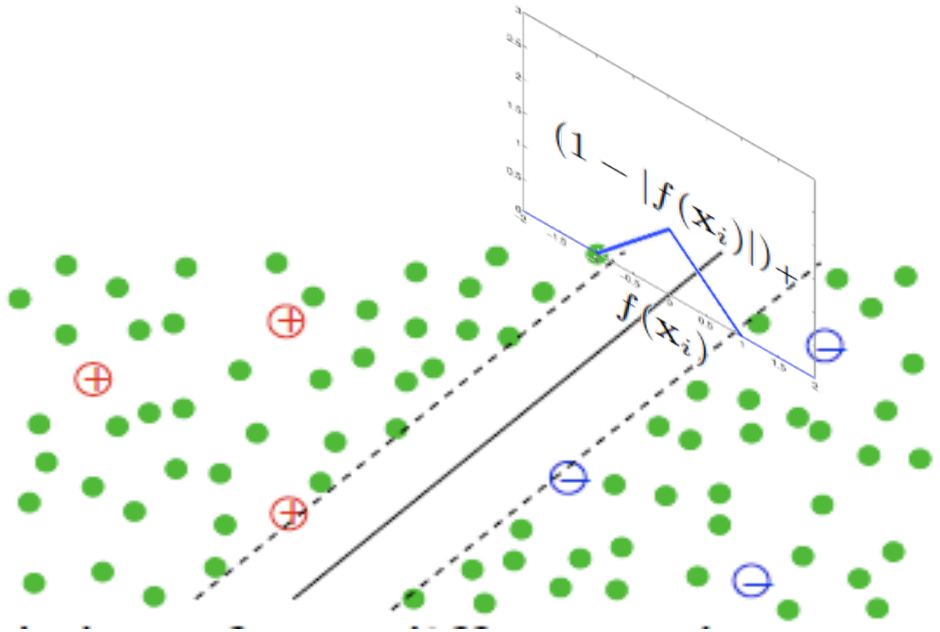




Bedeutung – niedrige Dichte

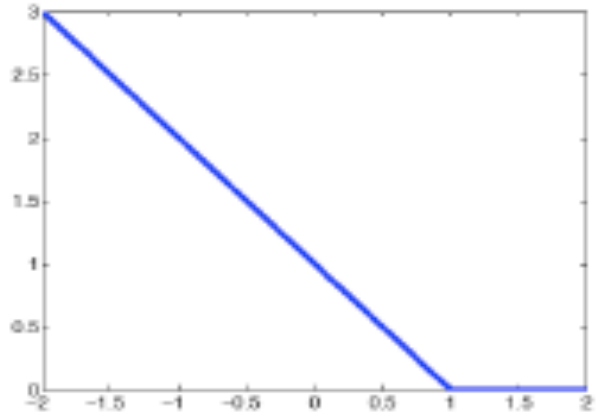
$$\min_{\mathbf{w},b}\left\{\frac{1}{2}||\mathbf{w}||^2 + C_1\sum_{i=1}^{l}(1-y_i f(\mathbf{x}_i))_+ + C_2\sum_{i=l+1}^{n}(1-|f(\mathbf{x}_i)|)_+\right\}$$

- Letzter Term: $(1-|f(\mathbf{x}_i)|)_+$
  - Bevorzugt $f(\mathbf{x}_i) \geq 1 \quad f(\mathbf{x}_i) \leq -1$
  - Insbesondere Daten die nicht nahe an der Entscheidungsfunktion mit $f(x) = 0$ liegen
  - bevorzugt Daten außerhalb des Randes
  - Äquivalent damit, dass die Trennung $f(x) = 0$ in einer Region ohne gelabelte und ungelabelte Daten liegt (➔ **niedrige Dichte**)

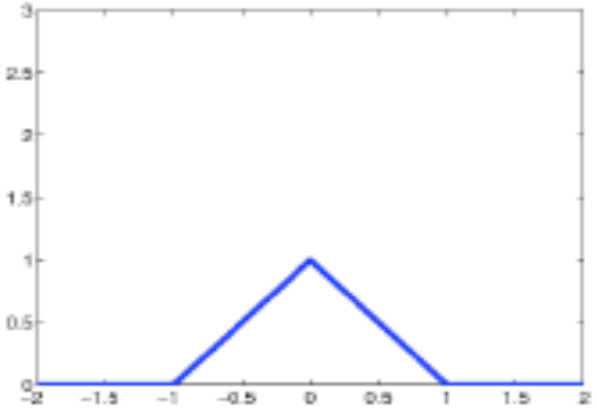


„Transduktive“ SVM – Optimierungsherausforderung

- Die Standard – SVM hat ein konvexes Optimierungsproblem



- Die Transduktive SVM hat ein **nicht-konvexes Optimierungsproblem**
  - Finden einer Lösung ist komplizierter (NP-hart) und insbesondere nicht mehr eindeutig
  - Lokale Minima und falsche Lösungen
  - ➔ einige Lösungsansätze wurden vorgestellt: SVM^{light}, ΔS^3VM , continuation S^3VM, Branch and Bound

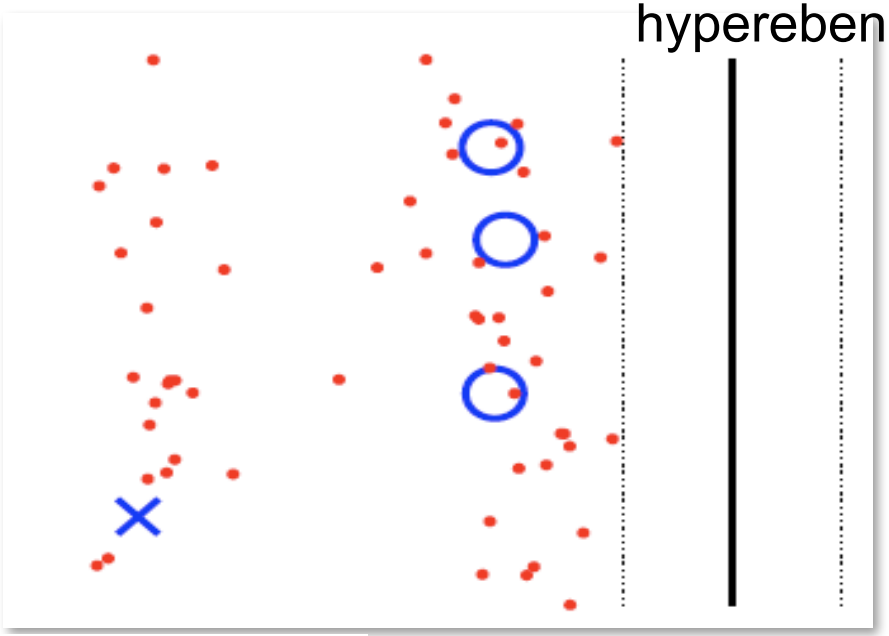


Transduktive SVM - Probleme

- Daten sind häufig unausgewogen
  ➔ meisten Daten werden einer Klasse zugeordnet

- Einführen einer Heuristik für die Ausbalancierung der Klassen (Erhalt der relativen Verteilung)

➔ Randbedingung für Optimierung:



$$\min_{\mathbf{w},b}\left\{\frac{1}{2}||\mathbf{w}||^2 + C_1\sum_{i=1}^{l}(1-y_i f(\mathbf{x}_i))_+ + C_2\sum_{i=l+1}^{n}(1-|f(\mathbf{x}_i)|)_+\right\}$$

- So dass: $\frac{1}{n-l}\sum_{i=l+1}^{n} f(x_i) = \frac{1}{l}\sum_{i=1}^{l} y_i$

SVM^{light} - Ansatz

- Heuristisches, iteratives „Labeln“ mit Ausbalancierung
- Trainiere SVM auf (X_l, Y_l)
- Labeln X_u durch $f(\mathbf{x}_i)$, Label $\hat{y}_i \leftarrow$ -1 / 1
- Von $\tilde{C} \leftarrow 10^{-5} C_2$ bis C_2 (schrittweise Iterieren)
  - Wiederhole
    - Trainiere SVM mit

$$\min_{\mathbf{w},b} \left\{ \frac{1}{2}||\mathbf{w}||^2 + C_1 \sum_{i=1}^{l} (1 - y_i f(\mathbf{x}_i))_+ + \tilde{C} \sum_{i=l+1}^{n} (1 - \hat{y}_i f(\mathbf{x}_i))_+ \right\}$$

    - Tausche Labels $\hat{y}_i, \hat{y}_j$ wenn $\exists (i,j)$ switchable (s. nächste Folie)
  - Bis keine tauschbaren Labels

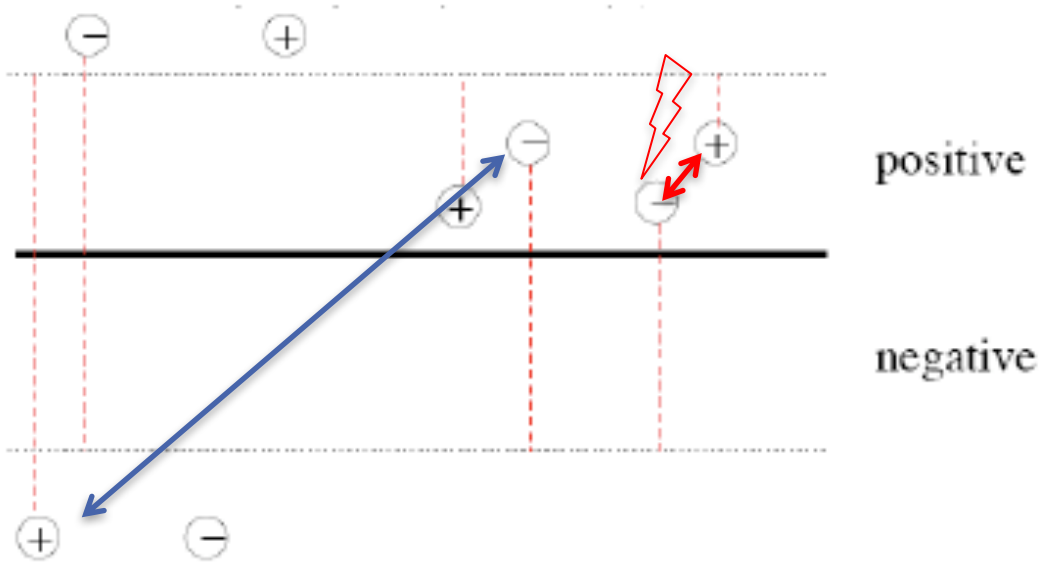
SVM^{light} - Ansatz

- (i, j) switchable, if $\hat{y}_i = +1, \hat{y}_j = -1$ and

$$loss(\hat{y}_i = +1, f(\mathbf{x}_i)) + loss(\hat{y}_j = -1, f(\mathbf{x}_j)) >$$
$$loss(\hat{y}_i = -1, f(\mathbf{x}_i)) + loss(\hat{y}_j = +1, f(\mathbf{x}_j))$$

$$loss(y, f(\mathbf{x}) = (1 - yf(\mathbf{x}))_+$$

- Beispiel
  - Rot: „Fehler“ (loss)
  - Tauschbar: 1_+ und 3_-
  - Nicht tauschbar: 4_+ und 4_-



S³VM – probabilistische Sicht für überwachtes Lernen

- Wahrscheinlichkeit für die Ausgabe y gegeben x

$$p(y|\mathbf{x}) = 1/\left(1 + \exp(-yf(\mathbf{x}))\right) \qquad f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$$

- Log- Gesamtwahrscheinlichkeit (über alle Daten)

$$\sum_{i=1}^{l} \log p(y_i|\mathbf{x}_i, \mathbf{w}, b)$$

- MAP – Training (inklusive Rand)

$$\max_{\mathbf{w},b} \sum_{i=1}^{l} \log\left(1/\left(1 + \exp(-y_i f(\mathbf{x}_i))\right)\right) - \lambda \|\mathbf{w}\|^2$$

  - Bzw.: $\min_{\mathbf{w},b} \sum_{i=1}^{l} \log\left(1 + \exp(-y_i f(\mathbf{x}_i))\right) + \lambda_1 \|\mathbf{w}\|^2$

S³VM – probabilistische Sicht Erweiterung für SSL

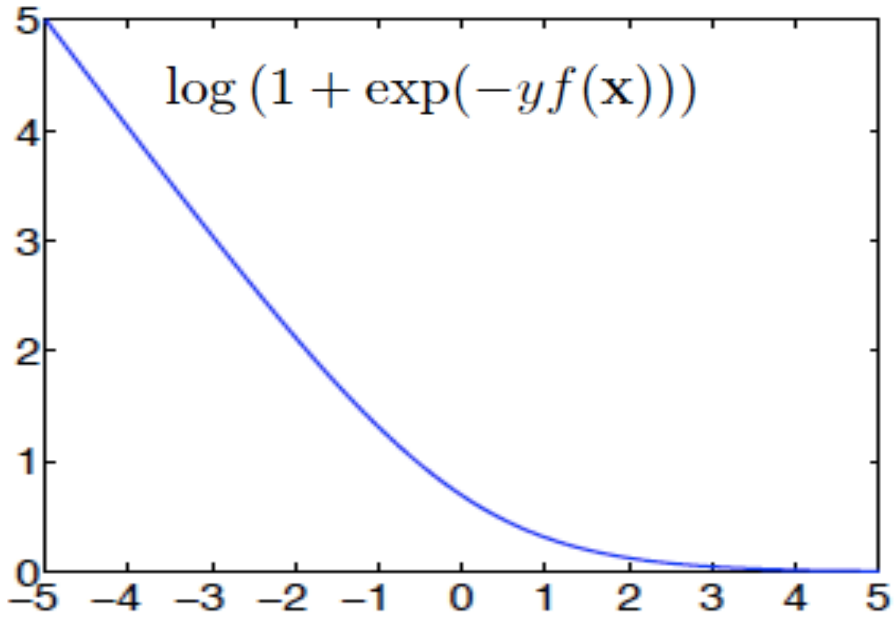
- Wenn zwei Klassen gut trennbar sind dann sollte
  - $p(y|x)$ nahe 0 oder 1 sein für alle Instanzen ungelabelter Daten
  - ➔ Entropie $H(p) = -p \log p - (1-p) \log(1-p)$ ist klein !!

- Entropie – Regularisierung
  - Definiere Entropie für ungelabelte Daten $\Omega(f) = \sum_{j=l+1}^{l+u} H(p(y=1|\mathbf{x}_j, \mathbf{w}, b))$
  - ➔ Neues Minimierungsproblem:

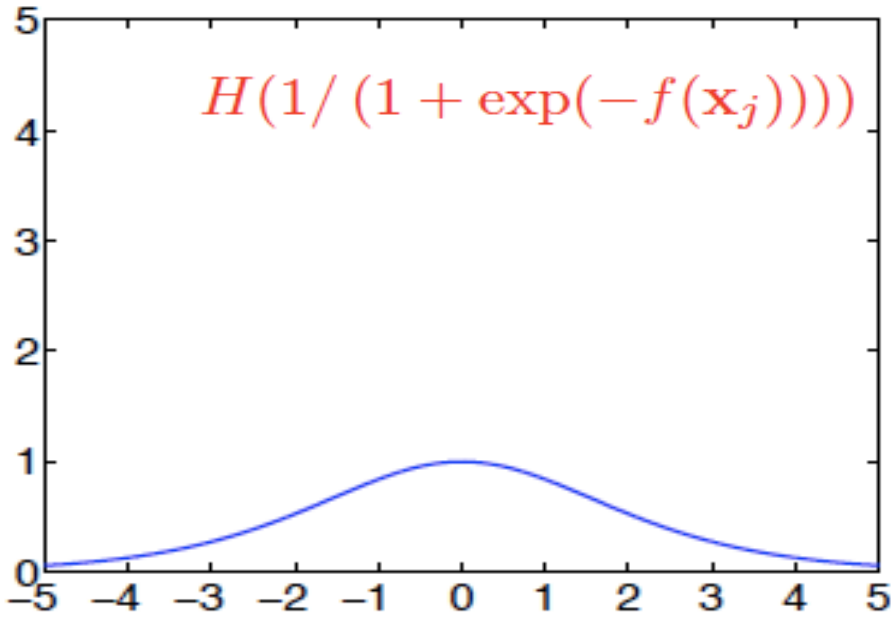
$$\min_{\mathbf{w},b} \quad \sum_{i=1}^{l} \log\left(1+\exp(-y_i f(\mathbf{x}_i))\right) + \lambda_1 \|\mathbf{w}\|^2 + \lambda_2 \sum_{j=l+1}^{l+u} H\left(1/\left(1+\exp(-f(\mathbf{x}_j))\right)\right)$$

S³VM – probabilistische Sicht

- Probablisitische Sicht führt zur Kostenfunktion (a)
- Entropie Betrachtung / Regularisierung führt zur Kostenfunktionen (b)



(a) the logistic loss



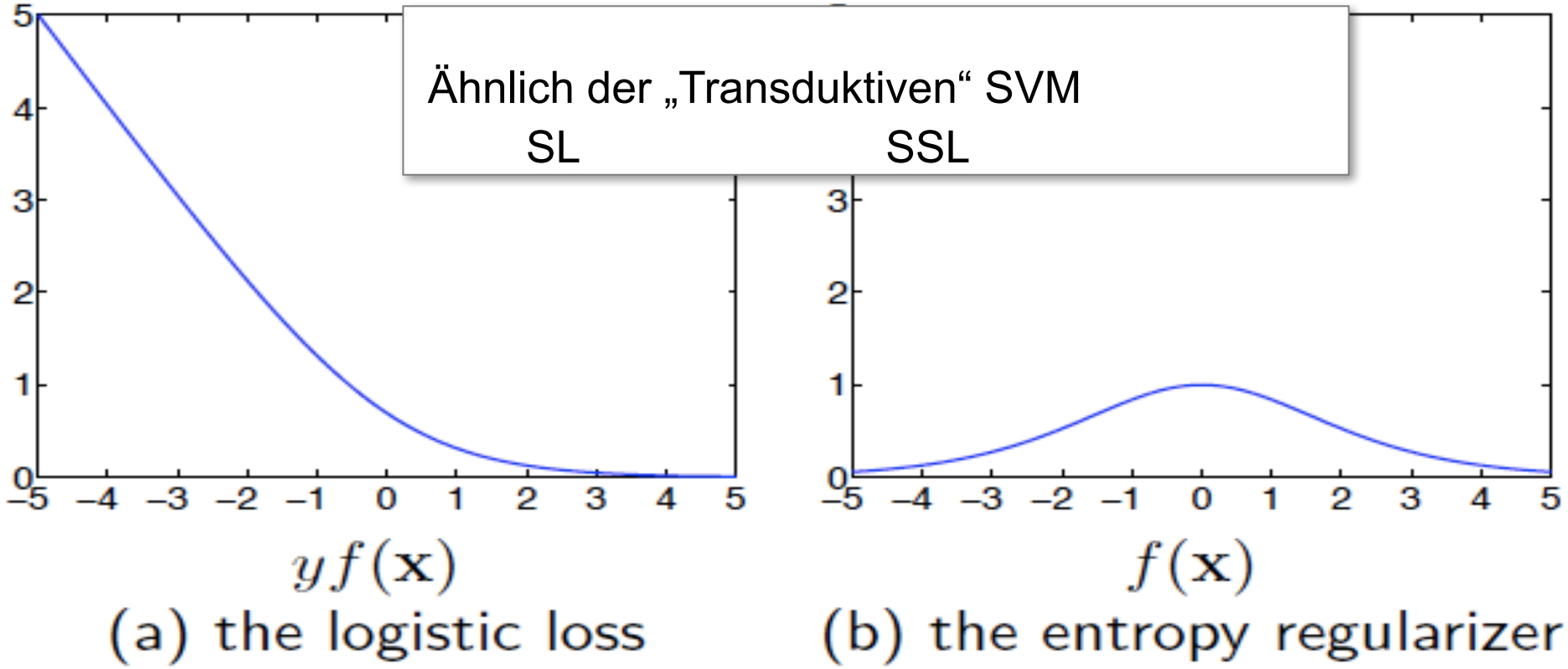
(b) the entropy regularizer

S³VM – probabilistische Sicht

- Probablisitische Sicht führt zur Kostenfunktion (a)
- Entropie Betrachtung / Regularisierung führt zur Kostenfunktionen (b)

Ähnlich der „Transduktiven“ SVM

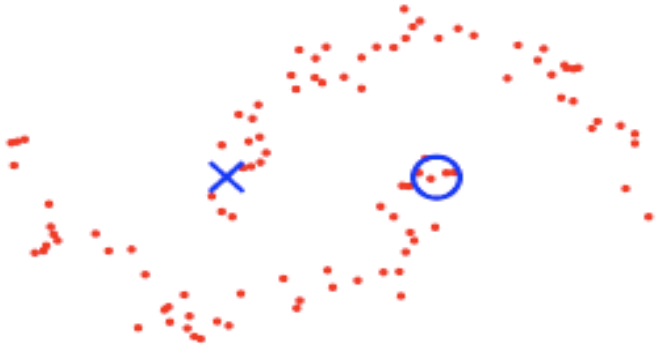
SL SSL



Vergleich verschiedener Ansätze I

"Two Moons" toy data

- easy for human (0% error)
- hard for S^3VMs!



S^3VM optimization method		test error	objective value
global min.	{Branch & Bound	0.0%	7.81
find local minima	CCCP	64.0%	39.55
	S^3VM^{light}	66.2%	20.94
	∇S^3VM	59.3%	13.64
	cS^3VM	45.7%	13.25

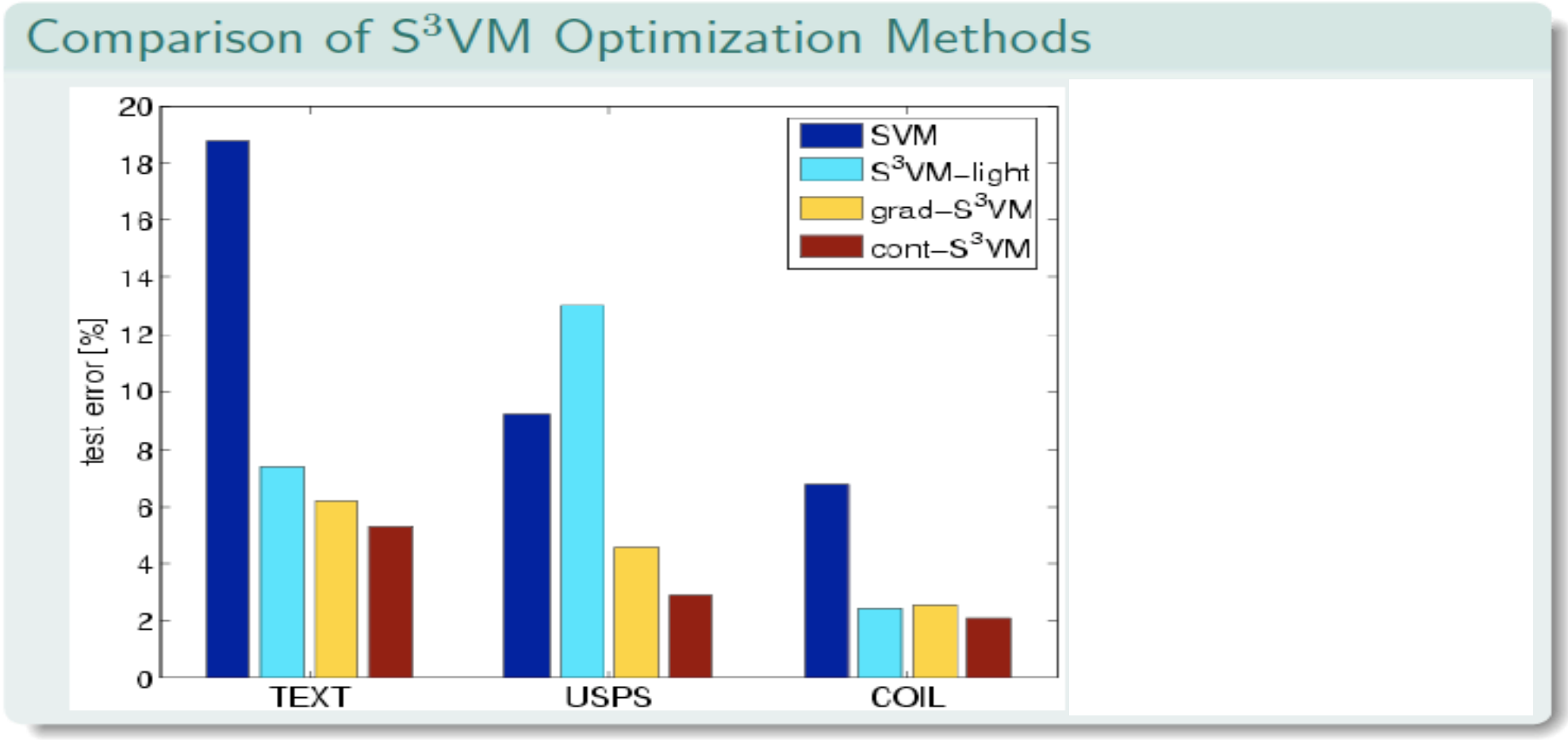
Vergleich verschiedener Ansätze II

Comparison of S^3VM Optimization Methods

On three tasks (with ∼2000 points each, 100 of which labeled)

- TEXT:
  - do newsgroup texts refert to mac or to windows?
    ⇒ binary classification
  - bag of words representation: ∼7500 dimensions, sparse
- USPS
  - recognize handwritten digits
  - 10 classes ⇒ 45 one-vs-one binary tasks
  - 16×16 pixel image as input (256 dimensions)
- COIL
  - recognize 20 objects in images: 20 classes
  - 32×32 pixel image as input (1024 dimensions)

Vergleich verschiedener Ansätze II



[Chapelle, Chi, Zien; ICML 2006]

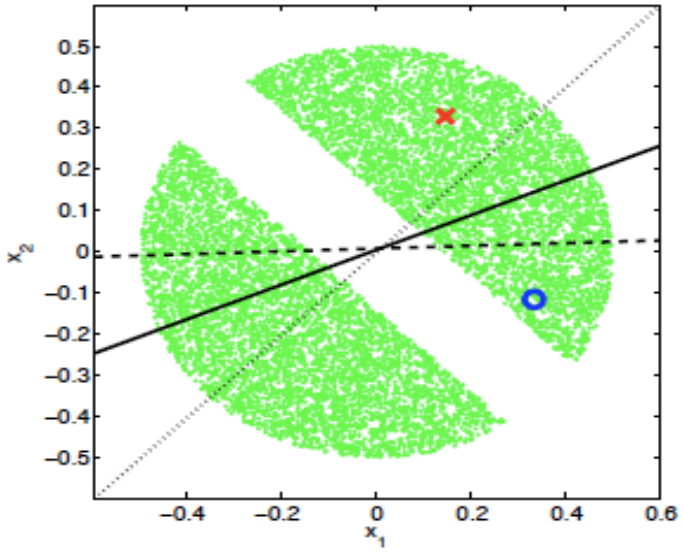
„Transduktive“ SVM / S^3VM - Diskussion

- Vorteile
  - Anwendbar wenn SVM anwendbar
  - Klar formuliertes mathematisches Rahmenwerk

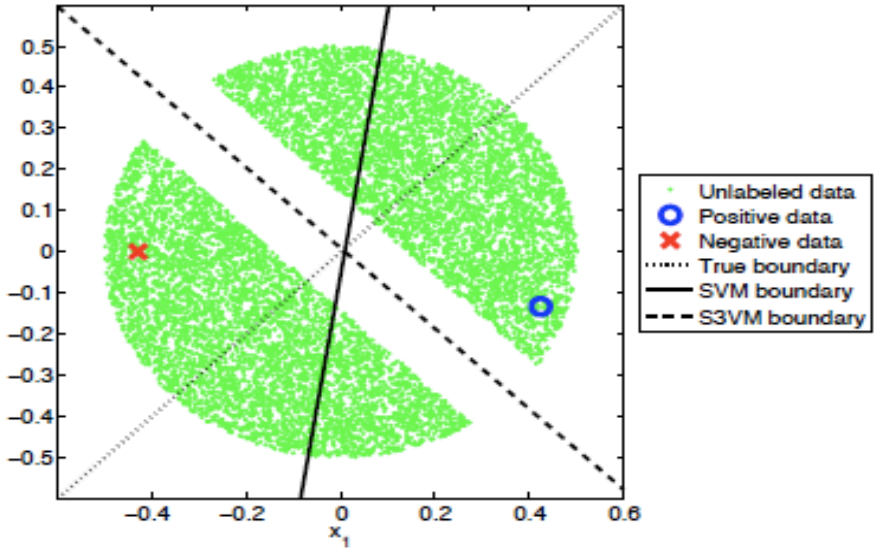
- Nachteile
  - Optimierungsproblem nicht mehr konvex
  - Optimierung – kompliziert
  - Lokale Minima
  - Schwächere Annahme (Dichte) als generative Modelle oder graphbasierte Methoden ➔ möglicherweise schlechtere Ergebnisse

S³VM – funktioniert nicht immer, ☹

- Insbesondere wenn die Grundannahme (Maximierung → Trennung niedrige Dichte) falsch ist:



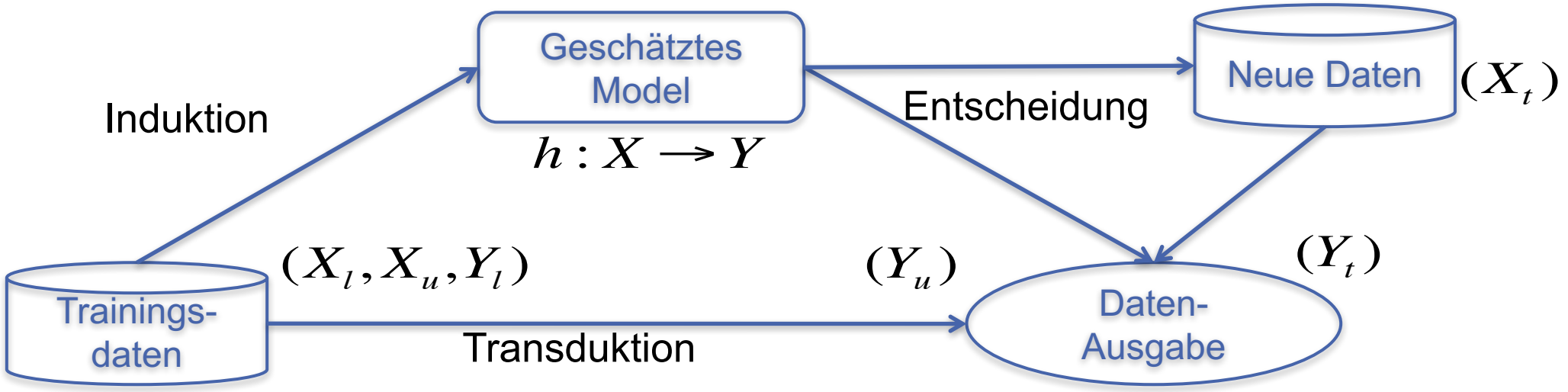
S3VM in local minimum



S3VM in wrong gap

SVM error: 0.26 ± 0.13
S3VM error: 0.34 ± 0.19

Induktion (Deduktion) Transduktion



[Learning from Data: Concepts, Theory and Methods.
V. Cherkassky, F. Mulier. Wiley, 1998.]

- Vorsicht: einige Verfahren heißen zwar „transductive ...“ sind aber eher induktiv (z.B. die ursprüngliche transductive SVM)

Verschieden Ansätze

- Erste Algorithmen
  - Self-Training & Co-Training
- Generative probabilistische Modelle (Generative Probabilistic Models)
  - EM for Gaussian Mixtures
- Dichte Trennung (Low-Density Separation)
  - Transduktive SVM
- Graph basierte Modelle / Methoden
  - Methoden bei denen die Daten als Knoten eines Graphs repräsentiert sind und die Kanten die jeweiligen Abstände enthalten
- Änderung der Repräsentation
  - unüberwachtes Lernen (z.B.: Clustern) um neue (i.A. niedrig dimensionale) Repräsentationen der Daten zu erhalten
  - Lernen der Zuordnung der Cluster zu Klassen

➔ Literatur

Karlsruher Institut für Technologie

Maschinelles Lernen II - Fortgeschrittene Verfahren
V03 Aktives Lernen

Sommersemester 2017
Prof. Dr. J.M. Zöllner, Prof. Dr. R. Dillmann

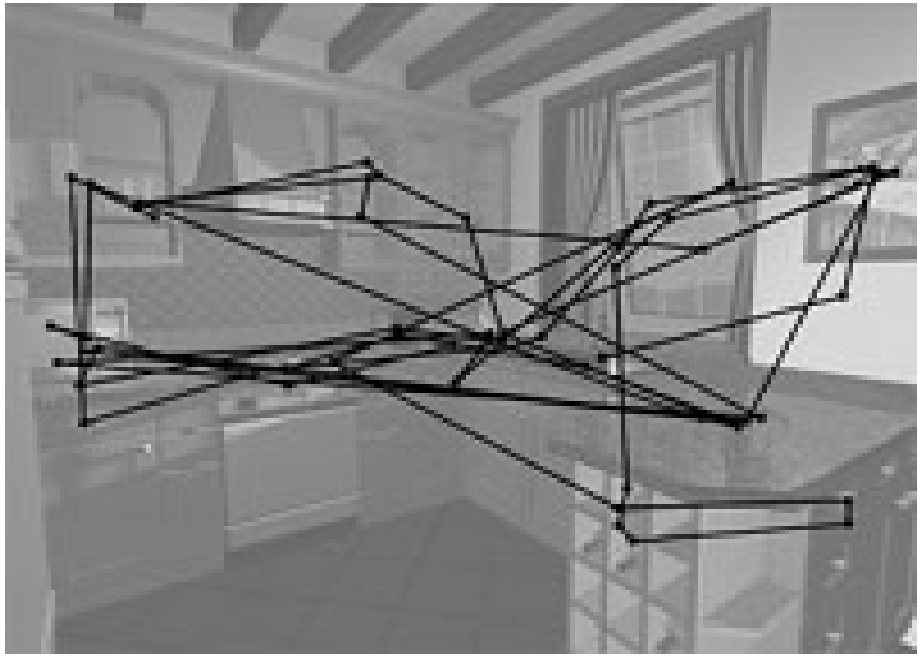
INSTITUT FÜR ANGEWANDTE INFORMATIK UND FORMALE BESCHREIBUNGSVERFAHREN
INSTITUT FÜR ANTHROPOMATIK UND ROBOTIK

Übersicht Aktives Lernen

- Motivation
- Formalisierung
  - (Lernszenarien)
- Gedankenexperiment
- Methoden basierend auf Unsicherheit (u. weiteren Maßen)
- Version-Space Ansätze
  - QBC
  - SVM

Feedback driven learning

The eyes focus on the interesting and relevant features, and do not sample all the regions in the scene in the same way.



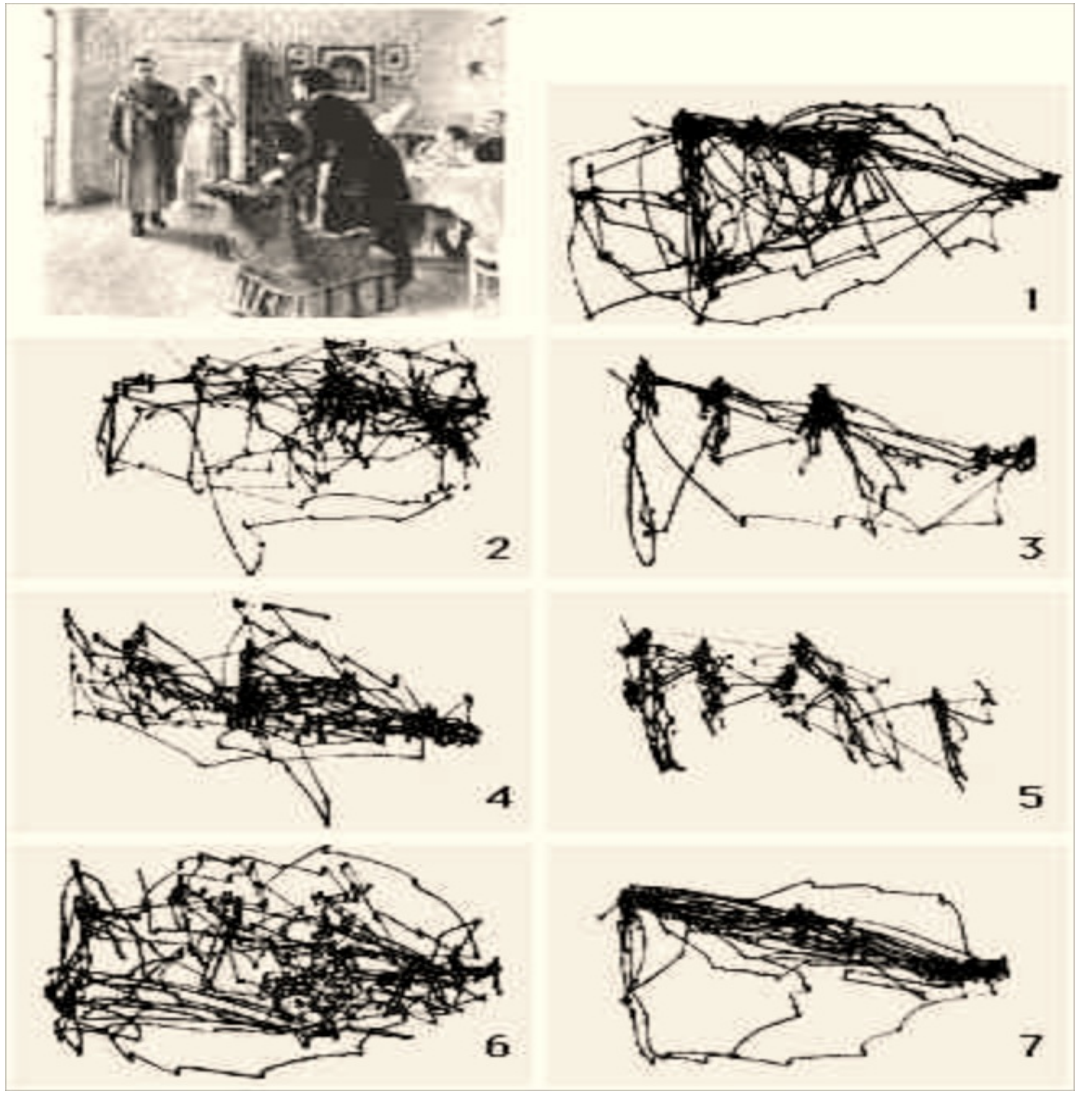
Experiment – Wohin sieht der Mensch?

Analysiere:

- Wie alt sind die Personen?
- Welche Kleider tragen die Personen?
- Wo stehen die Personen?
- Wie lang war der Besucher weg?
- .....

(Yarbus 1967)

Experiment – Augenbewegung



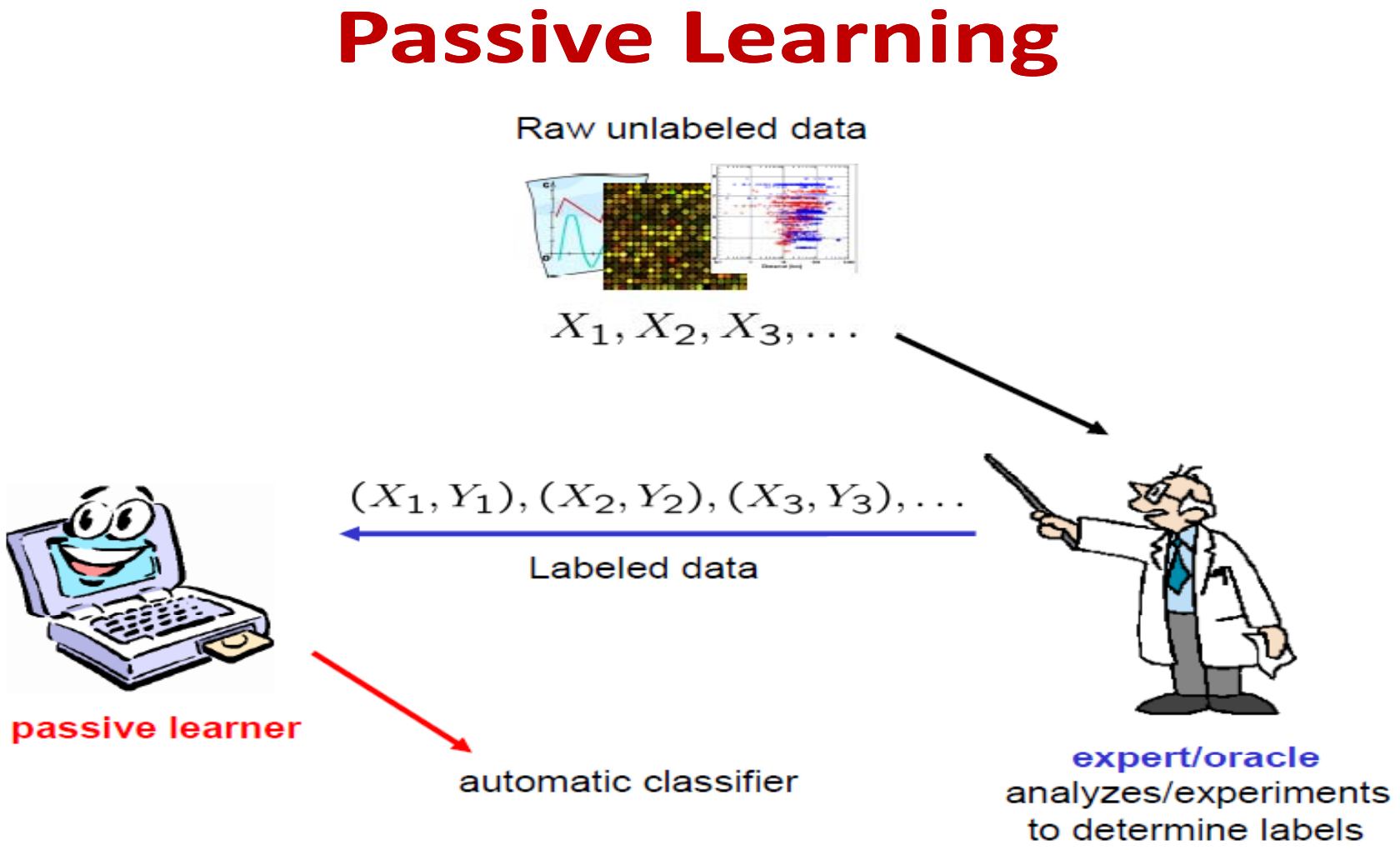
Aufnahmen der Augenbewegungen von Probanden (jeweils 3 Minuten pro Frage):

1. Freies Explorieren
2. Reich oder Arm?
3. Alter der Personen?
4. Was hat die Familie gemacht?
5. Welche Kleider tragen die Personen?
6. Wo stehen die Personen?
7. Wie lang war der Besucher weg?

➔ Zielgerichteter Einsatz der Augenbewegung

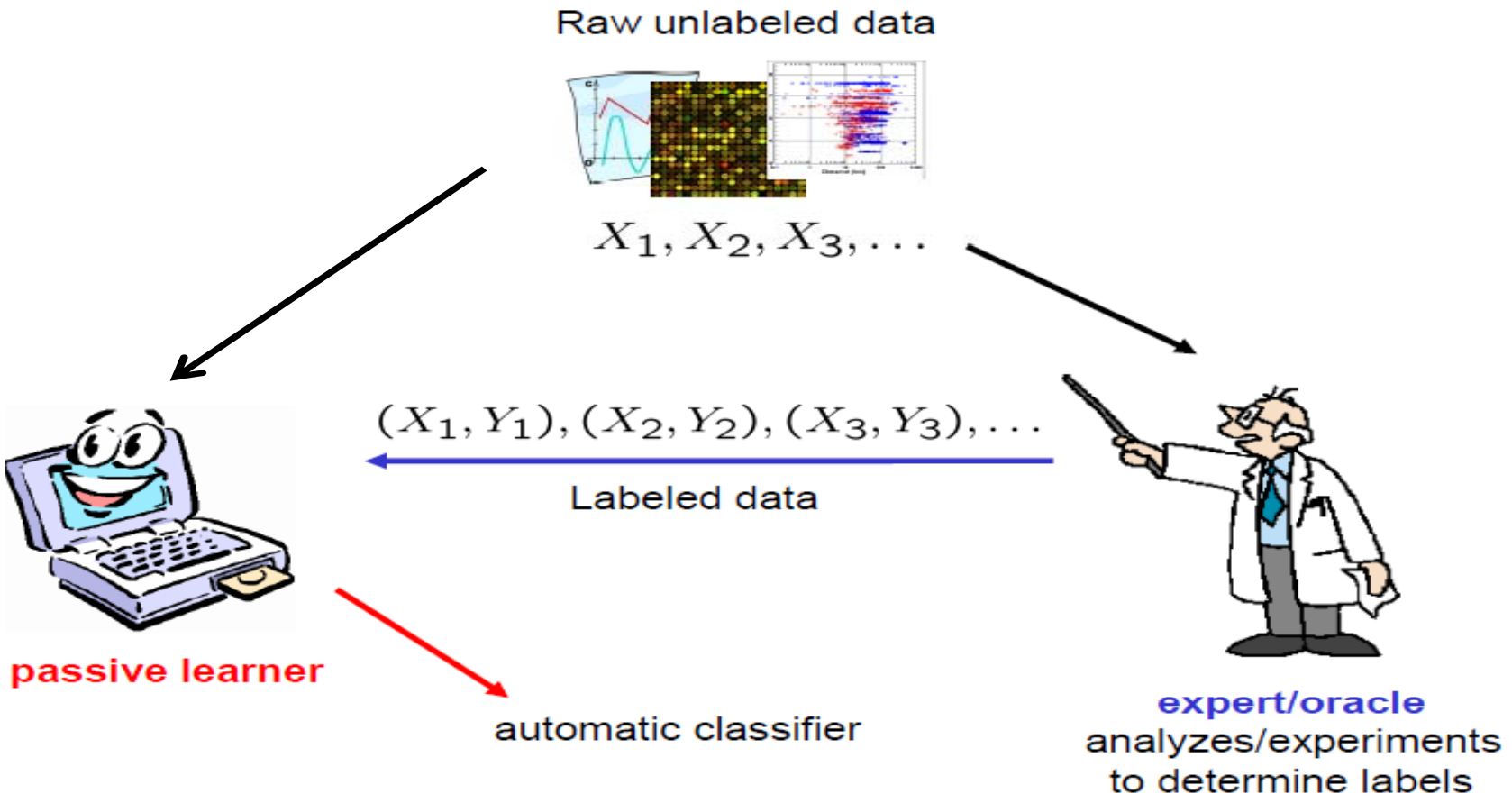
➔ Auswahl von Lerndaten!!

Vergleich Passiv-SSL-Aktiv



Vergleich Passiv-SSL-Aktiv

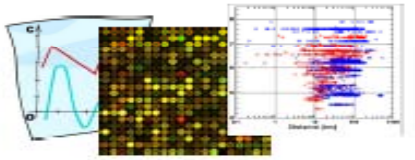
Semi-supervised Learning



Vergleich Passiv-SSL-Aktiv

Active Learning

Raw unlabeled data



$X_1, X_2, X_3, \ldots$

Learner requests labels for **selected** data

$(X_1, ?)$

(X_1, Y_1)

$(X_3, ?)$

(X_3, Y_3)

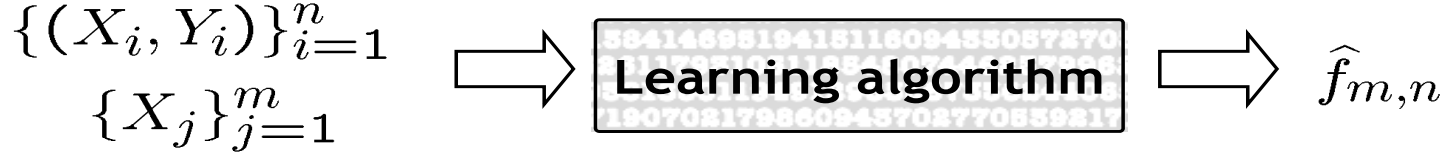
active learner

expert/oracle
analyzes/experiments
to determine labels

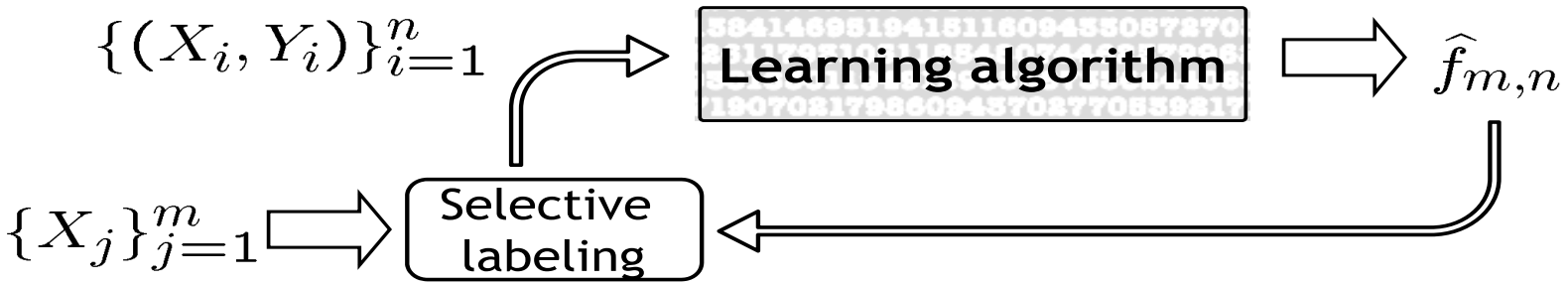
automatic classifier

Lernen aus nicht-gelabelten Daten

- SSL
  - Lernmaschine die mit wenigen überwachten und vielen unüberwachten Daten lernt
  - Annahme: zusätzliche Information über Datenverteilung ermöglicht Hypothesenfindung

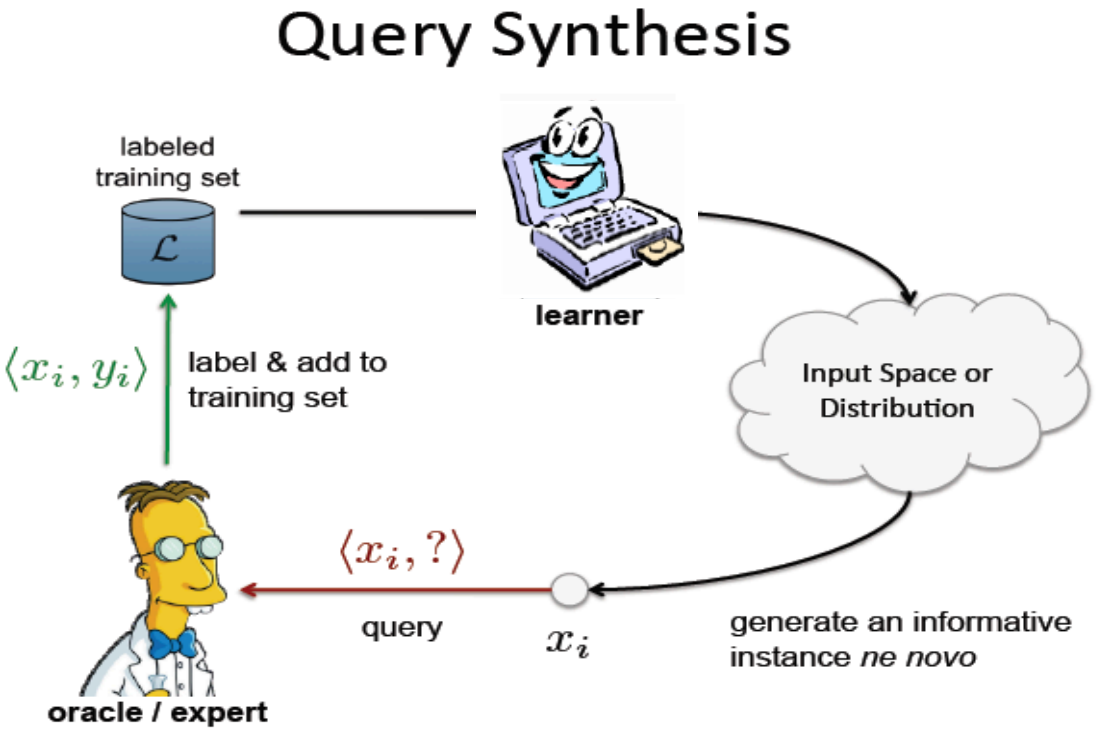


- Aktives Lernen (allgemeiner)
  - Lernmaschine die ggf. mit wenigen überwachten aber wesentlich mit selektiv gewählten unüberwachten Daten lernt
  - Annahme: Einige Daten enthalten wesentlich mehr Information als andere



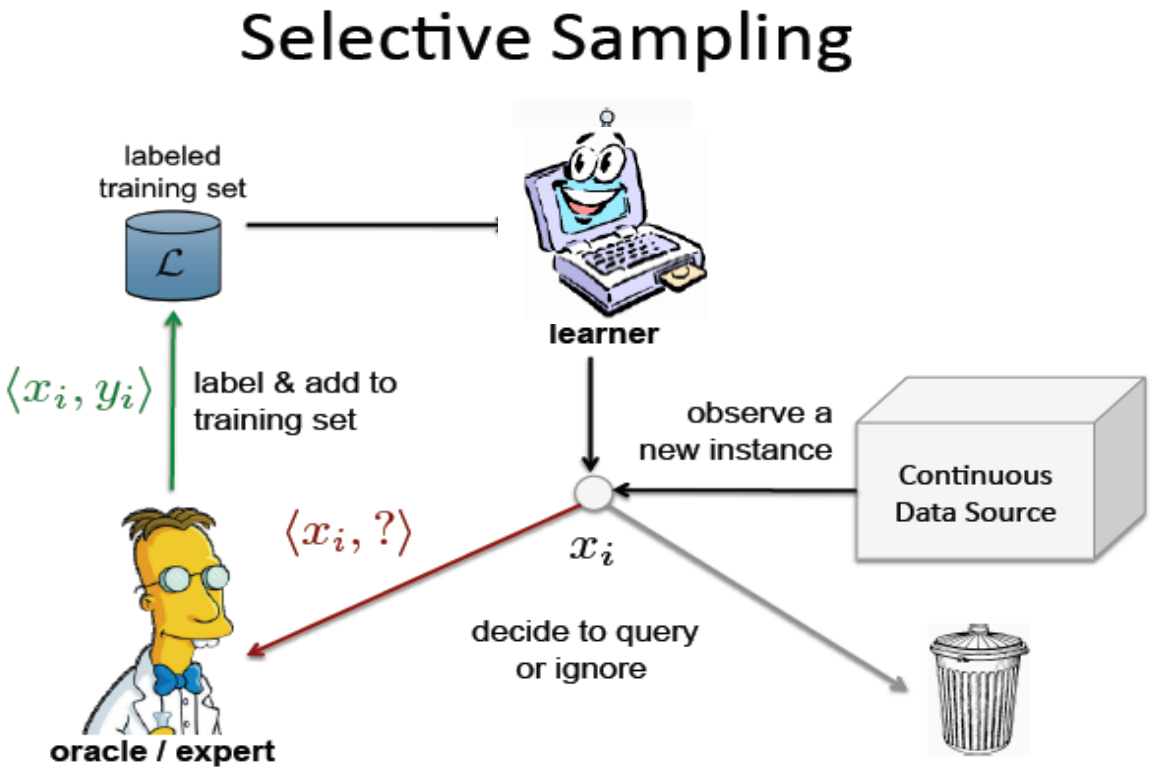
Lernszenarien

- Grundsätzlich (anwendungsbedingt) gibt es 3 Möglichkeiten des aktiven Lernens:
  - Query Synthesis
    (Erzeugung synthetischer Daten)
  - Selective Sampling
    (Selektive Entnahme aus Daten-Strom)
  - Pool Based
    (Auswahl aus Daten-Pool)



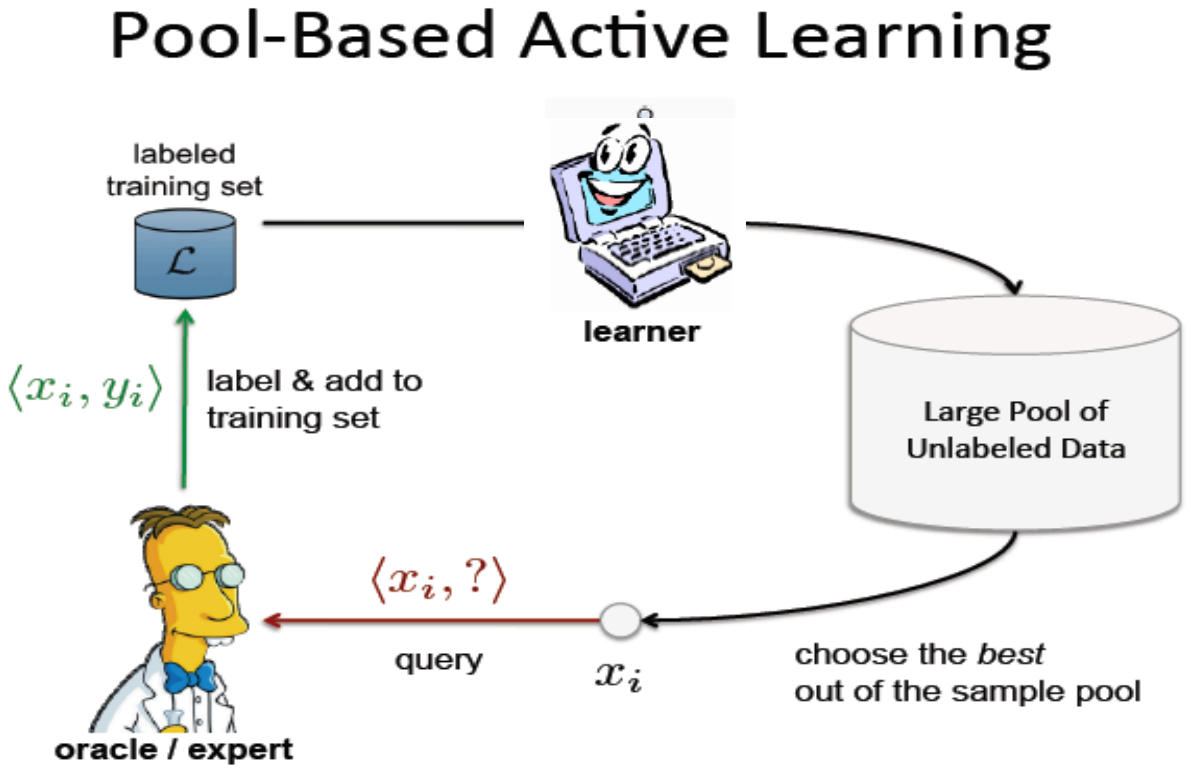
Lernszenarien

- Grundsätzlich (anwendungsbedingt) gibt es 3 Möglichkeiten des aktiven Lernens:
  - Query Synthesis
    (Erzeugung synthetischer Daten)
  - Selective Sampling
    (Selektive Entnahme aus Daten-Strom)
  - Pool Based
    (Auswahl aus Daten-Pool)



Lernszenarien

- Grundsätzlich (anwendungsbedingt) gibt es 3 Möglichkeiten des aktiven Lernens:
  - Query Synthesis
    (Erzeugung synthetischer Daten)
  - Selective Sampling
    (Selektive Entnahme aus Daten-Strom)
  - Pool Based
    (Auswahl aus Daten-Pool)



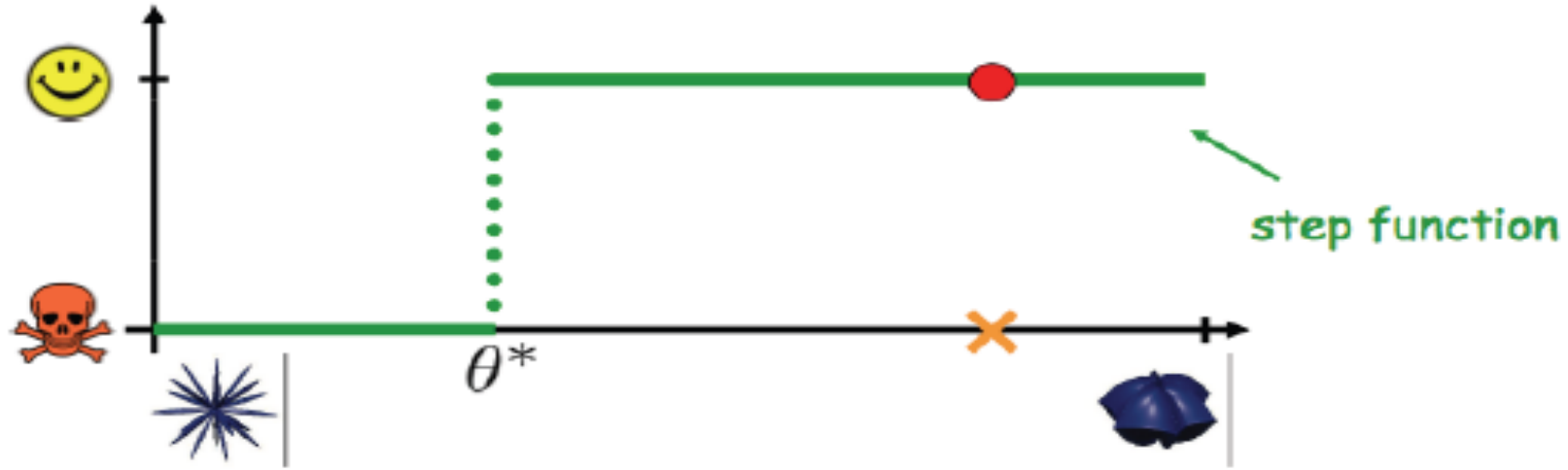
Gedankenexperiment

- Annahme – Ein Team ist in einer fremden Gegend und kennt die Früchte nicht einige sind giftig andere gut

- Wissen über Güte erst nach dem Ausprobieren ➔ Risiko hoch!
- Es gibt verschiedene Ausprägungen. Welche sind giftig?

Gedankenexperiment - Problem

- Lernen eines Schwellwertes, bzw. einer Klassifikation



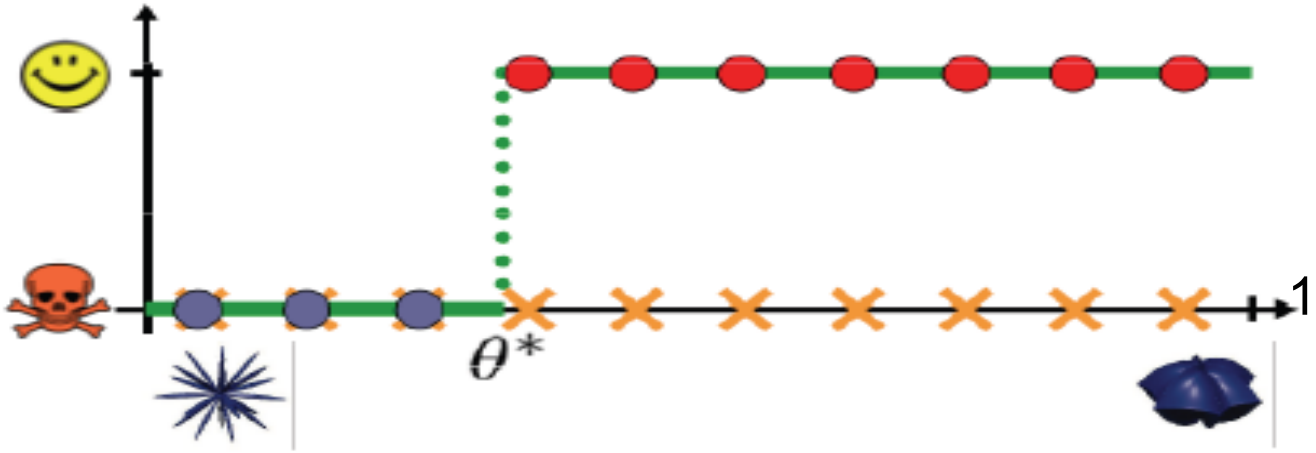
- Möglichst schnell und genau und mit wenig Daten d.h. wenig Tests um die Lernkosten (Risiko) niedrig zu halten

Gedankenexperiment – Idee?

- Im Vergleich dazu bei passivem überwachten Lernen
  → alle vorab (ggf. auch inkrementell) testen

Gedankenexperiment – Passives Lernen

- Lernen eines Schwellwertes einer Klassifikation

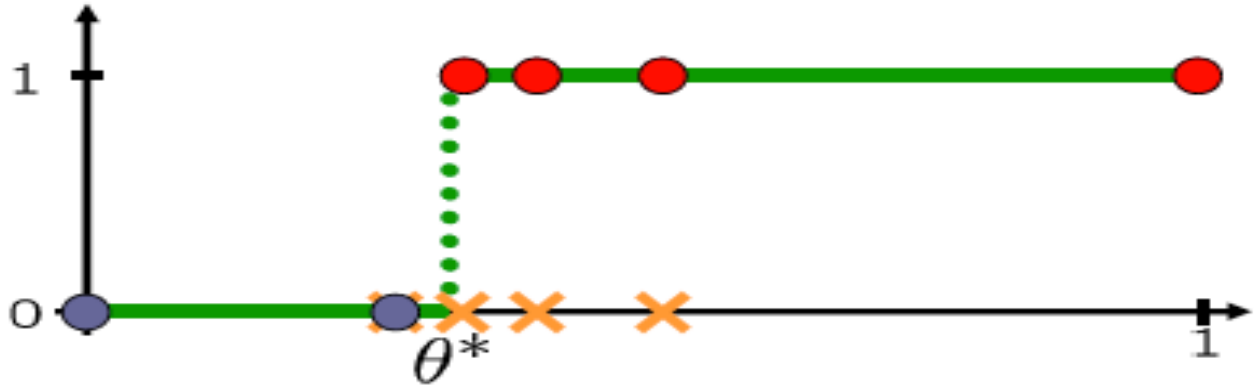


- Bei passivem Lernen müssen die Instanzen und ihr Label vor dem Lernen bekannt sein
- Fehlerreduktion bei n äquidistanten Lerndaten, pro Lernschritt i

$$\left|\widehat{\theta}_{i+1} - \widehat{\theta}_i\right| \approx \frac{1}{n}$$

Gedankenexperiment – Aktives Lernen

- Lernen eines Schwellwertes einer Klassifikation

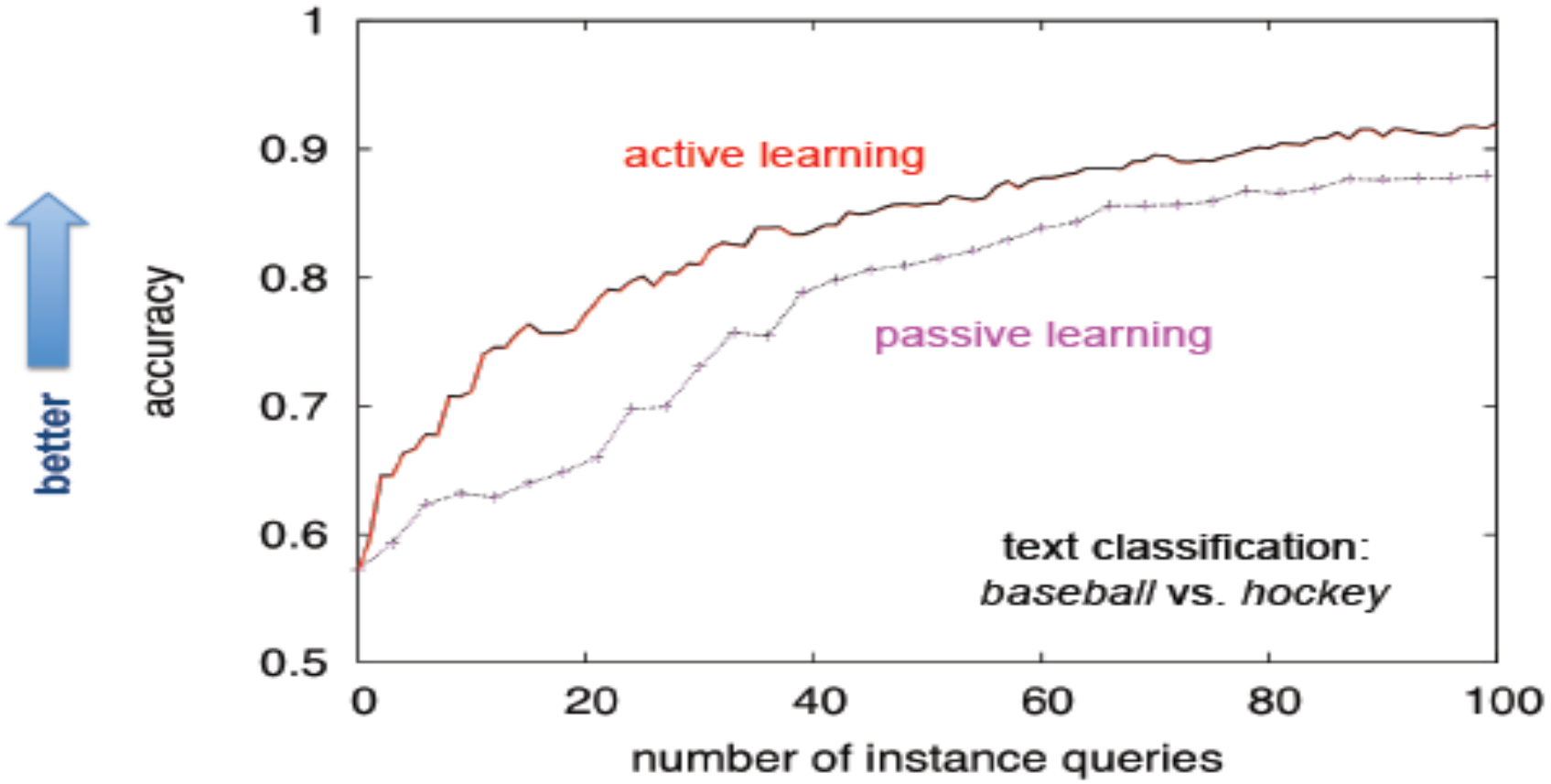


- Bei aktivem Lernen werden die Instanzen auf Basis des vorhergehenden Tests gezogen
- Fehlerreduktion bei n Daten mit Halbierung des Intervalls:

$$\left|\widehat{\theta}_n - \theta^*\right| \approx 2^{-n}$$

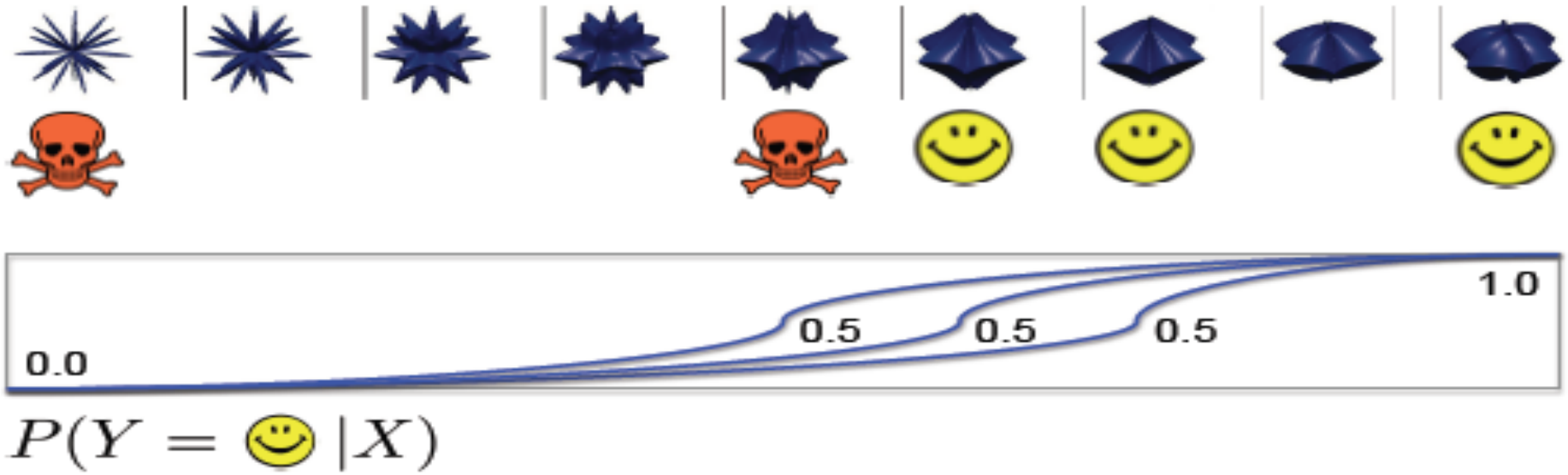
Vergleich typischer Lernkurven

- Exponentielle Verbesserung beim Lernen



Wie wählt man die nötigen Lerndaten?

- Probabilistische Interpretation

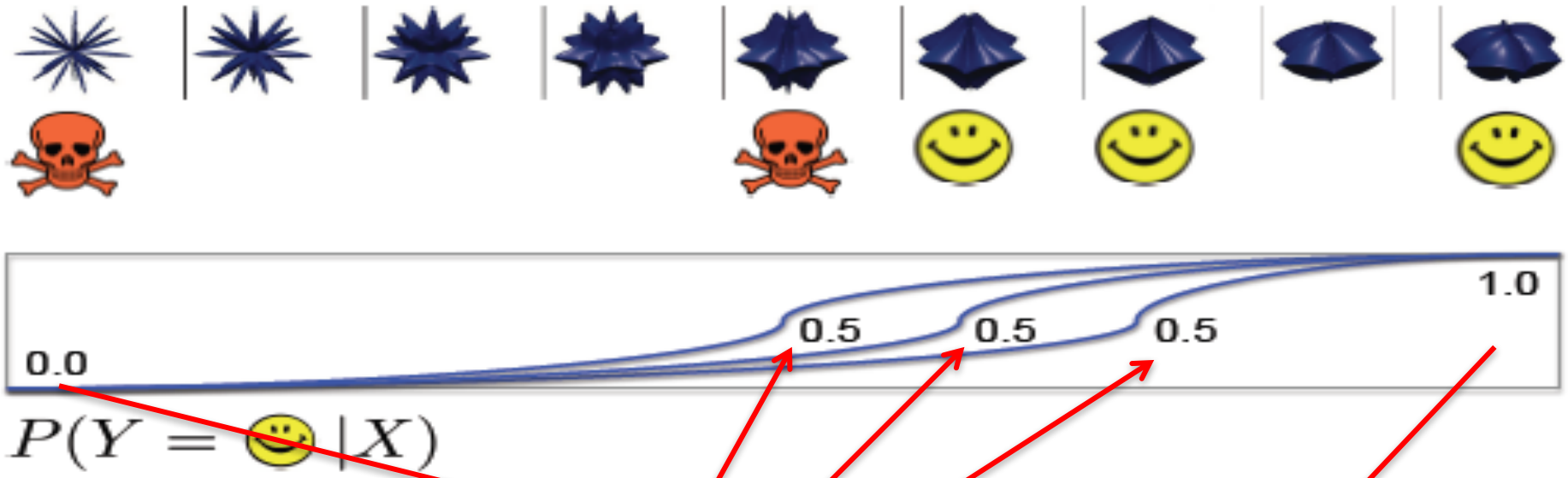


- Wähle die Daten für die Lernmaschine die größte Unsicherheit haben, z.B.:

$$x^*_{LC} = \text{argmax}_x \; 1 - P_\theta(\hat{y}|x) \qquad wobei \quad \hat{y} = \text{argmax}_y \, P_\theta(y|x)$$

Wie wählt man die nötigen Lerndaten?

- Probabilistische Interpretation



- Wähle die Daten für die Lernmaschine die größte Unsicherheit haben, z.B.:

$$x^*_{LC} = \text{argmax}_x \; 1 - P_\theta(\hat{y}|x) \qquad wobei \quad \hat{y} = \text{argmax}_y \, P_\theta(y|x)$$

Weitere Unsicherheitsmaße

- Niedrigste Konfidenz (bester Klassenzugehörigkeit) bzw. höchste Unsicherheit

$$\phi_{LC}(x) = 1 - P_\theta(y^*|x)$$

- Kleinster Rand / Margin (bester Klassenzugehörigkeiten)

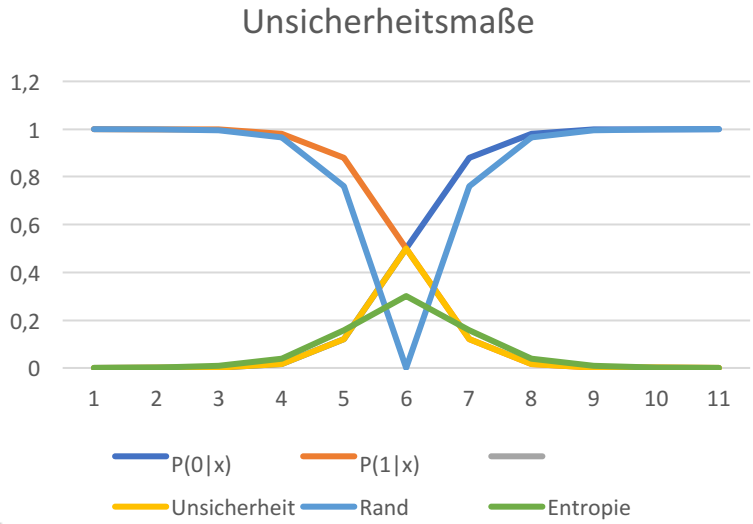
$$\phi_M(x) = P_\theta(y_1^*|x) - P_\theta(y_2^*|x)$$

- Entropie
  - Als Erwartungswert des Informationsgehalts, definiert durch

$$E_y\left[-\log P_\theta(y|x)\right]$$

  - Maß (zu maximieren):

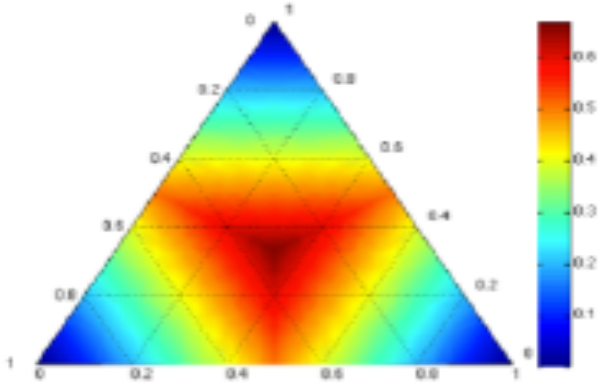
$$\phi_{ENT}(x) = -\sum_y P_\theta(y|x)\log_2 P_\theta(y|x)$$



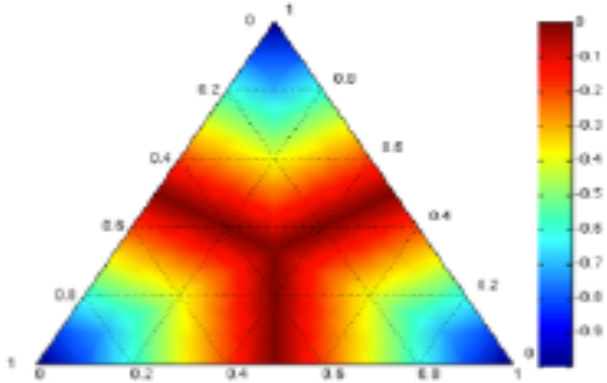
- Für binäre Klassifikation sind diese Maße identisch sonst nicht

Unsicherheitsmaße

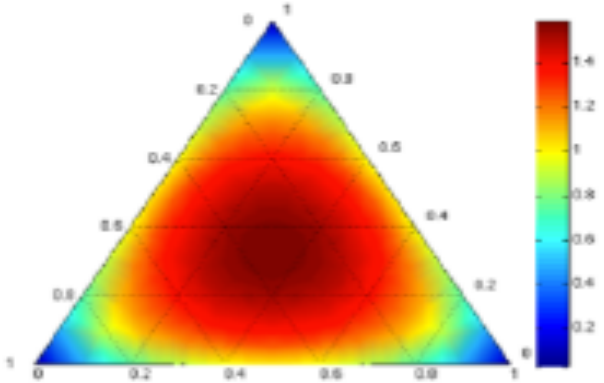
- 3 Klassen (jeweils in den Ecken)



Konfidenz



Margin



Entropie

  - Rot – Bereich der Daten die bevorzugt werden für 3 Klassen Aufgabe – gegeben die a posteriori Wahrscheinlichkeit (Ecken)

- Für binäre Klassifikation sind diese Maße identisch sonst nicht

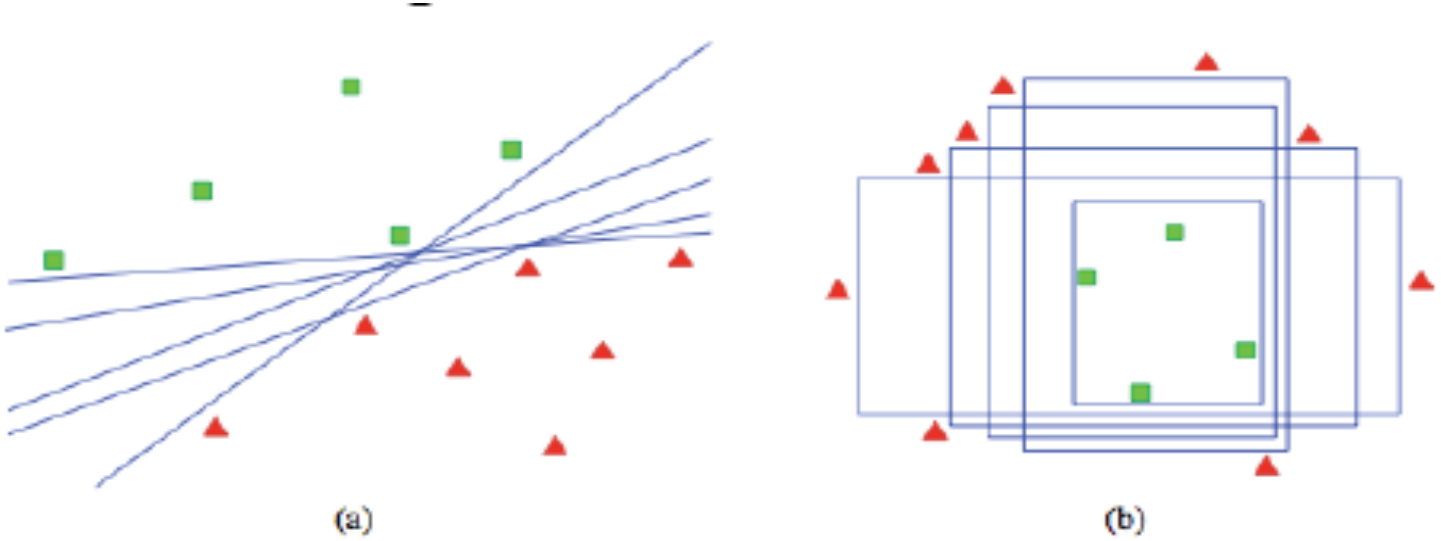
Diskussion und Anwendung in Lernszenarien

- Grundsätzlich sind einfache Verfahren, weit verbreitet
  - Pool based Lernen, Grundalgorithmus z.B.:
    - Evaluiere alle Instanzen x
    - Ranking nach Unsicherheit
    - Abfrage der unsichersten k Instanzen
    - Neu Lernen, Iterieren
  - Selektive Entnahme, Grundalgorithmus z.B.:
    - Festlegen einer Unsicherheitsregion z.B. [0.2, 0.8]
    - Beobachte neue Instanzen
    - Abfrage der Instanzen, die in die Unsicherheitsregion fallen
    - Neu Lernen, iterieren

- Problematisch
  - nur die Konfidenz weniger möglichen Hypothesen (Klassifikatoren) wird betrachtet– diese können auch bezüglich **wichtiger** unbekannter Daten sicher sein ➔ schlechtes Ergebnis
  - Bessere Lösung?

Version Space (konsistenter Hypothesenraum)? Wdh?!

- VS: Die Menge aller Hypothesen die konsistent sind mit den Daten



- Annahme: Um so größer der Version Space $\mathcal{V}$ ist um so schlechter ist jede mögliche Hypothese (Klassifikator)
- Ziel beim aktiven Lernen: Reduktion des Version Space

Gedankenexperiment – Version Space

- Binäre Klassifikation

- Mögliche Hypothesen: 8 → 4 → 2 → 1

Simpler (naiver) Version Space Algorithmus

- Bestimme alle konsistenten Hypothesen
  - Oder bestimme $|\mathcal{V}|$ analytisch
- Optimales neues x reduziert die „Größe“ von $\mathcal{V}$ am stärksten
  - als Erwartungswert über y (weil Label zunächst unbekannt)
  - über alle Lerndaten inklusive der neuen Daten $\mathcal{L} \cup \langle x, y \rangle$

$$x^*_{VS} = \arg\min_x E_y \left| \mathcal{V}^{\mathcal{L} \cup \langle x, y \rangle} \right|$$

- Diskussion
  - Idealerweise lässt sich der Version Space halbieren
  - Binäre Suche implementiert dies in 1D
  - Problem – effiziente Realisierung
    - $\mathcal{V}$ kann sehr groß werden oder ist analytisch nicht beschreibbar
    - Idee: „Extremen“ des Hypothesenraums betrachten, wenn die Modelle sich „stark“ widersprechen ➔ Daten (mit hoher Unsicherheit) reduzieren $\mathcal{V}$
    - Allgemeiner Ansatz: Query-by-Committee

Query – by – Committee QBC

- Allgemeiner Ansatz
  - Trainiere eine Menge $\mathcal{C}$ von Maschinen (Klassifikatoren)
    - $\mathcal{C}$ kann beliebiger Kardinalität sein
    - Wähle neue Daten wenn die Hypothesen (Klassifikatoren) widersprüchlich sind

- Selektive Entnahme
  - ...
  - Beobachte neue Instanzen (Auswerten)
  - Abfrage falls Widerspruch
  - Neutrainieren, Iterieren

- Pool-based Lernen
  - ...
  - Messung des Widerspruchs für alle Instanzen x
  - Ranking
  - Abfrage der k widersprüchlichsten Instanzen
  - Neutrainieren, Iterieren

Query – by – Committee QBC

- Design:
  - Wahl des Ausschusses $\mathcal{C}$, z.B.:
    - „sampling“ von zulässigen Modellen geg. Lerndaten entsprechend $P(\theta|\mathcal{L})$
    - Lernen der Modelle ← Datenabhängig
  - Bestimmung des Widerspruchs
    - Einfach – z.B. XOR
    - Korrekt - Betrachten der Einzelentscheidungen als Wahrscheinlichkeitsverteilung und Unsicherheitsmaß darauf anwenden, z.B. Entropie:
      - Bayes‘sche Interpretation eines Ensembles

$$P_{\mathcal{C}}(y|x) = \sum_{\theta \in \mathcal{C}} P_\theta(y|x) P(\theta)$$

Theoretisch gut

Geeignete Realisierung ?

      - Unsicherheitsmaß über als Entropie

$$\phi_{VE}(x) = -\sum_{y} \sum_{\theta \in \mathcal{C}} \left[P_\theta(y|x) P(\theta) \right] \log \left[P_\theta(y|x) P(\theta) \right]$$

Ausreißerproblem

- Problem:
  - Eine Instanz kann widersprüchlich sein weil es sich um einen Ausreißer handelt
  - Ausreißer sind nicht geeignete Lerndaten
- Mögliche Lösung
  - Gewichten der Unsicherheit einer Instanz x anhand der Dichte im Datenraum

$$\phi_{ID}(x) = \phi_{VE}(x) \times \left(\frac{1}{U} \sum_{u \in \mathcal{U}} \mathrm{sim}(x, u) \right)^{\beta}$$

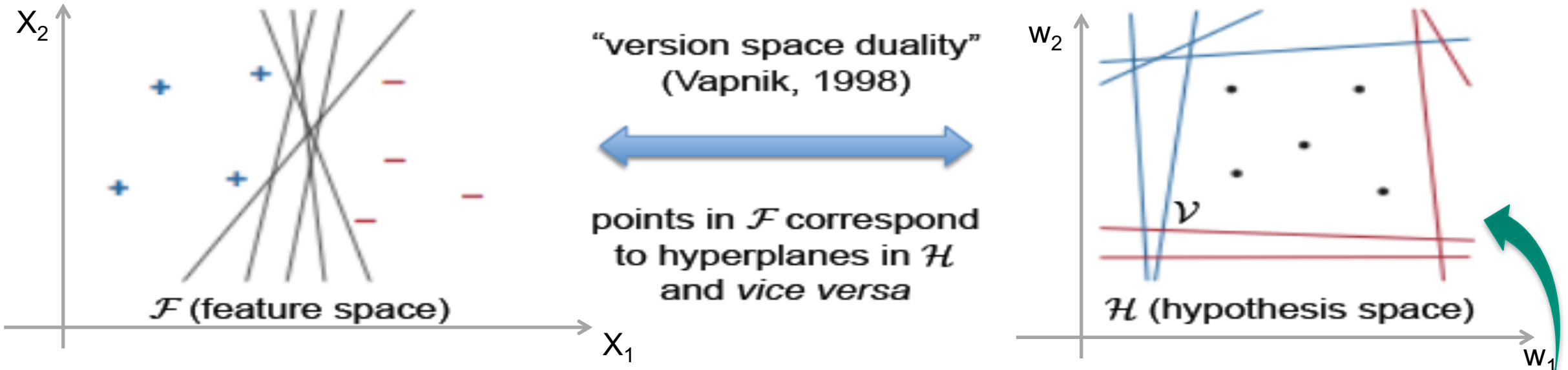
Entropie

Dichte im Datenraum

  - Verwende Datenpool $\mathcal{U}$ um $P(x)$ zu approximieren
  - Je größer die Ähnlichkeit der Instanzen um so eher kein Ausreißer

Version Space für SVM

- Besondere Eigenschaft: Dualität Merkmal- u. Hypothesenraum

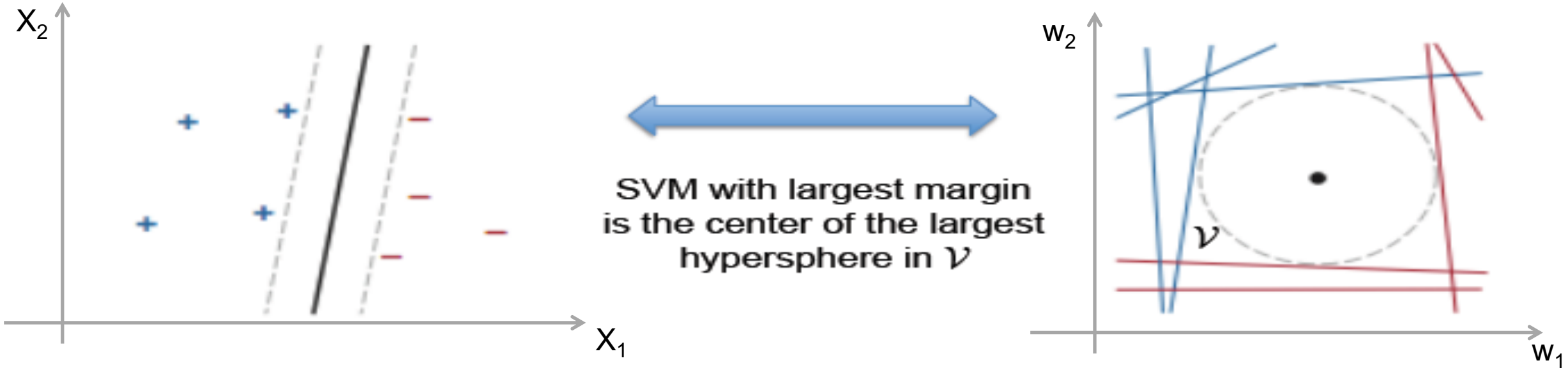


- Randbedingung korrekter Klassifikation $y_i\left(\vec{w}\vec{x}_i + b\right) \geq 1 \quad i = 1 \ldots n$

- D.h. aber auch, dass jeder Datenpunkt $x_i \quad i = 1 \ldots n$ eine Hyperebene im Hypothesenraum definiert s.d. gültige $\vec{w}$ jeweils in einem Halbraum = reduzierter Hypothesenrum sind

Version Space für SVM

- Größter Rand



- Heißt gleichzeitig dass wir nach dem $\vec{w}$ suchen, dass den maximalen Abstand zu allen Hyperebenen der Datenpunkte hat ➔ Mittelpunkt der Hyperkugel

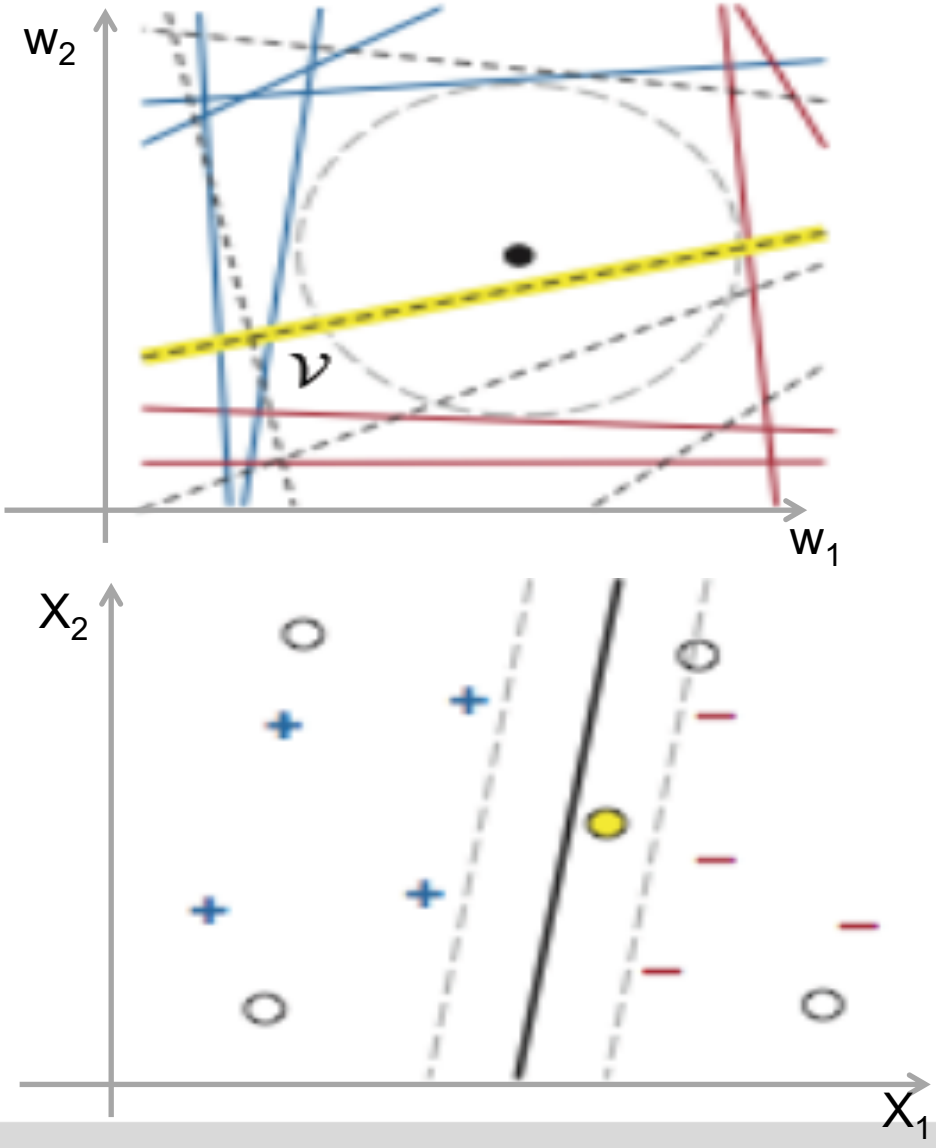
Aktives Lernen mit SVM – Version Space

- Gegeben ungelabelte Instanzen (➔Hyperebenen in $\mathcal{H}$) suchen wir diejenigen, die den VS maximal verringern

- Einfache Lösung: „Simple Margin“

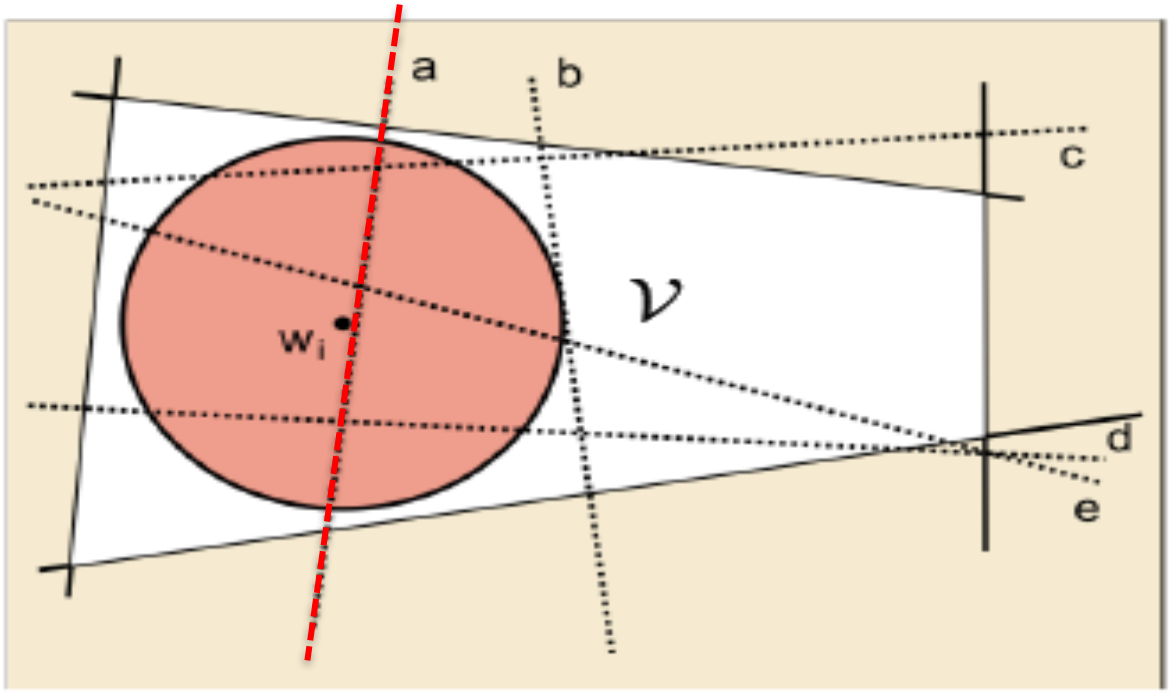
  ➔ Daten deren entsprechende Hyperebene die Hyperkugel gültiger Gewichtsvektoren möglichst zentral schneiden

  ➔ dies sind die Daten die im nächsten zur Trennhyperebene im Merkmalsraum liegen



Aktives Lernen mit SVM – Version Space

- Simple Margin
  - Wenn $\mathcal{V}$ nicht symmetrisch ist und der Gewichtsvektor nicht zentral – langsame Konvergenz
  - in der Skizze:
    Instanz a ➔ sehr langsame Konvergenz
    Instanz b ➔ wäre besser



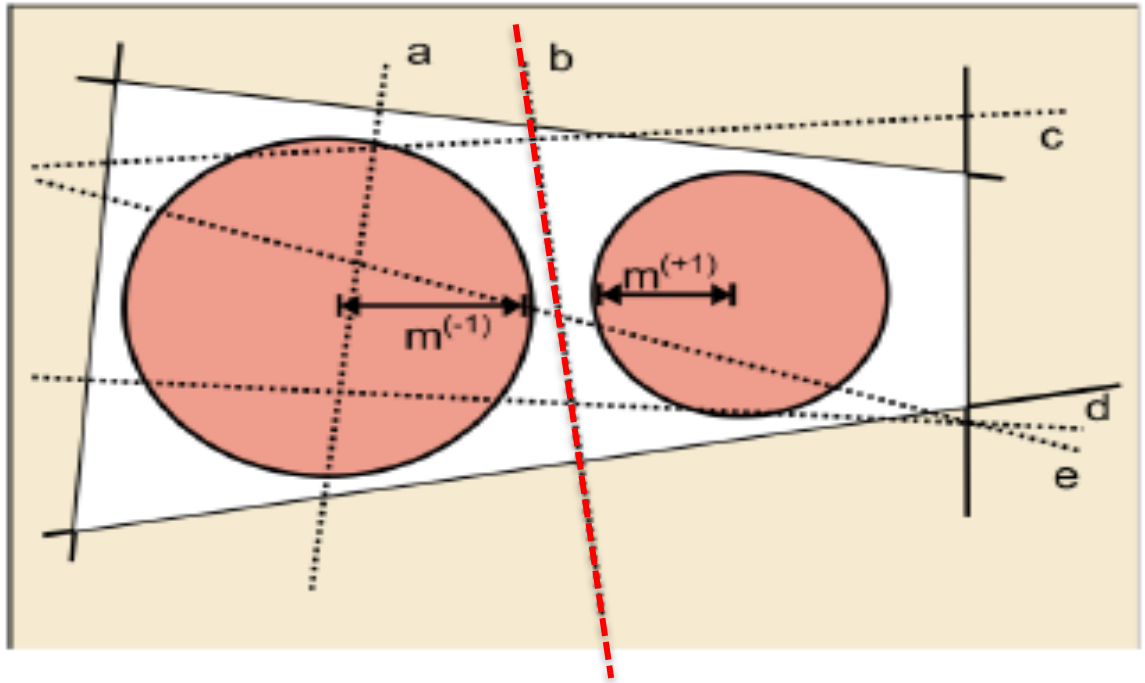
- Besser: Nutzen der Eigenschaft, dass der SVM Rand (~Radius Hyperkugel) proportional zur Version Space Fläche ist

Aktives Lernen mit SVM – Version Space

- MaxMin Margin
  - Für jeden Datenpunkt berechne den Rand $m+$ und $m-$ nach potentieller Teilung in $\mathcal{V}+$ bzw. $\mathcal{V}-$
  - Abfragen der Instanz (Datenpunkt)

$$x = \arg\max_x \ \min(m+, m-)$$

  - in der Skizze:
    Instanz b ➔ kleinste Rand ist anschließend maximal

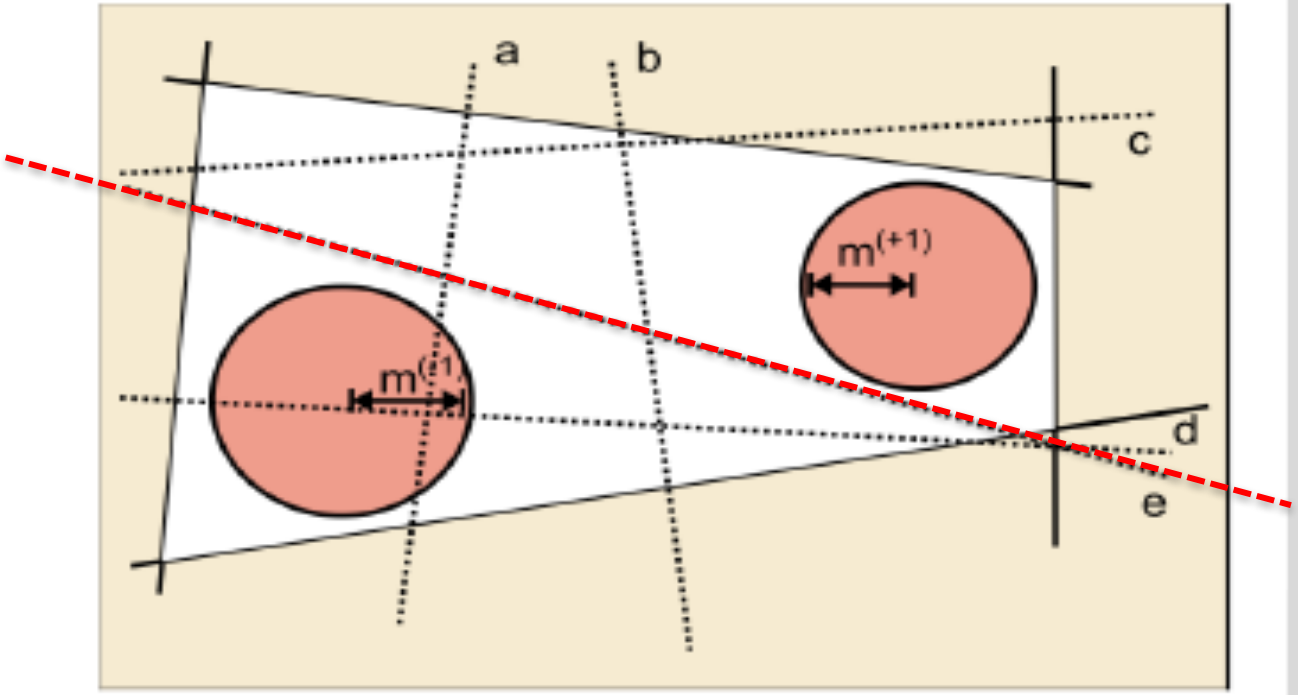


Aktives Lernen mit SVM – Version Space

- Ratio Margin
  - Für jeden Datenpunkt berechne den Rand $m+$ und $m-$ nach potentieller Teilung in $\mathcal{V}+$ bzw. $\mathcal{V}-$
  - Abfragen der Instanz (Datenpunkt)

$$x = \text{argmax}_x \ \min\left(\frac{m-}{m+}, \frac{m+}{m-}\right)$$

  - in der Skizze:
    Instanz e führt zu gleich großen Hypothesenräumen



Aktive SVM Diskussion

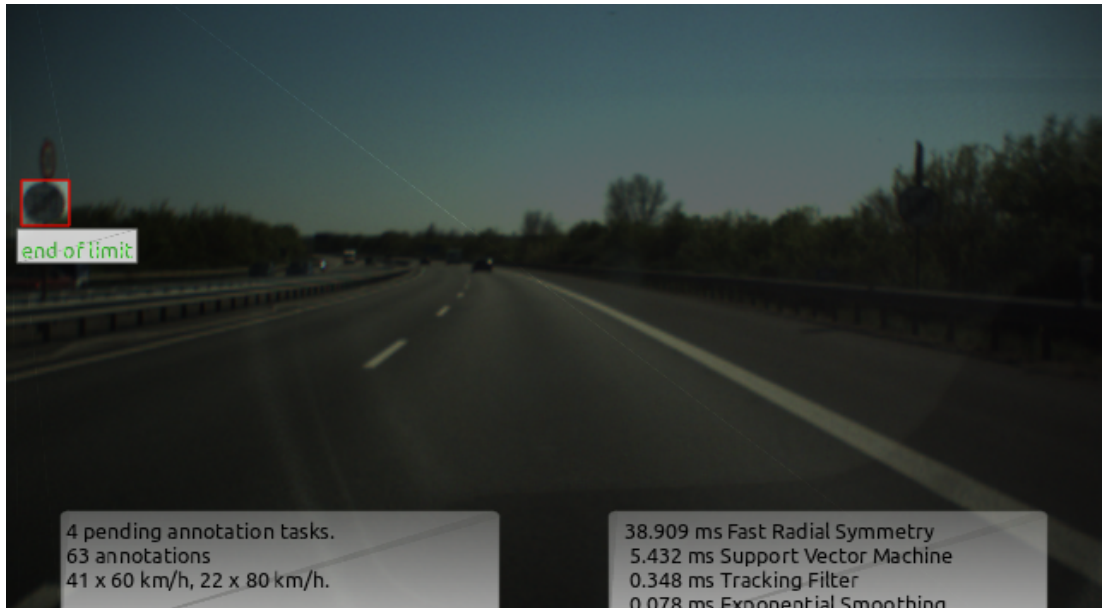
- Vorteile
  - Anwendbar wenn SVM anwendbar
  - Klar formuliertes mathematisches Rahmenwerk
  - Berechnung des Randes jeweils nach Trainieren der SVM möglich
  - Praktische Ergebnisse zeigen, dass aktive SVM besser als passive SVM

- Nachteile
  - MinMax und Ratio sind aufwändig in der Berechnung

Aktives Lernen Verkehrszeichenerkennung

- Probleme – Sehr viele (insbesondere negative) ungelabelte Beispiele vorhanden >> 100000
- Welche Beispiele sollen verwendet werden?
- Rechenzeit?
- Labels? Aufwand?
- Güte maximieren?
- Konvergenz?

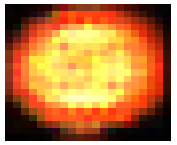
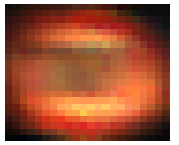
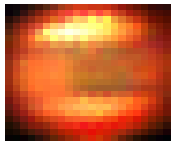
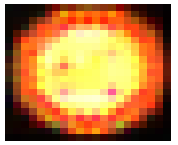
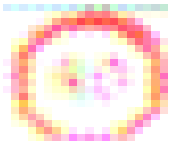
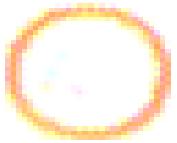
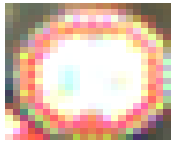
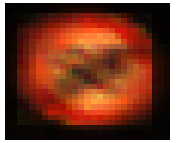
➔Aktives Lernen mit
SVM und s.g. uncertainty sampling
(~simpe margin)



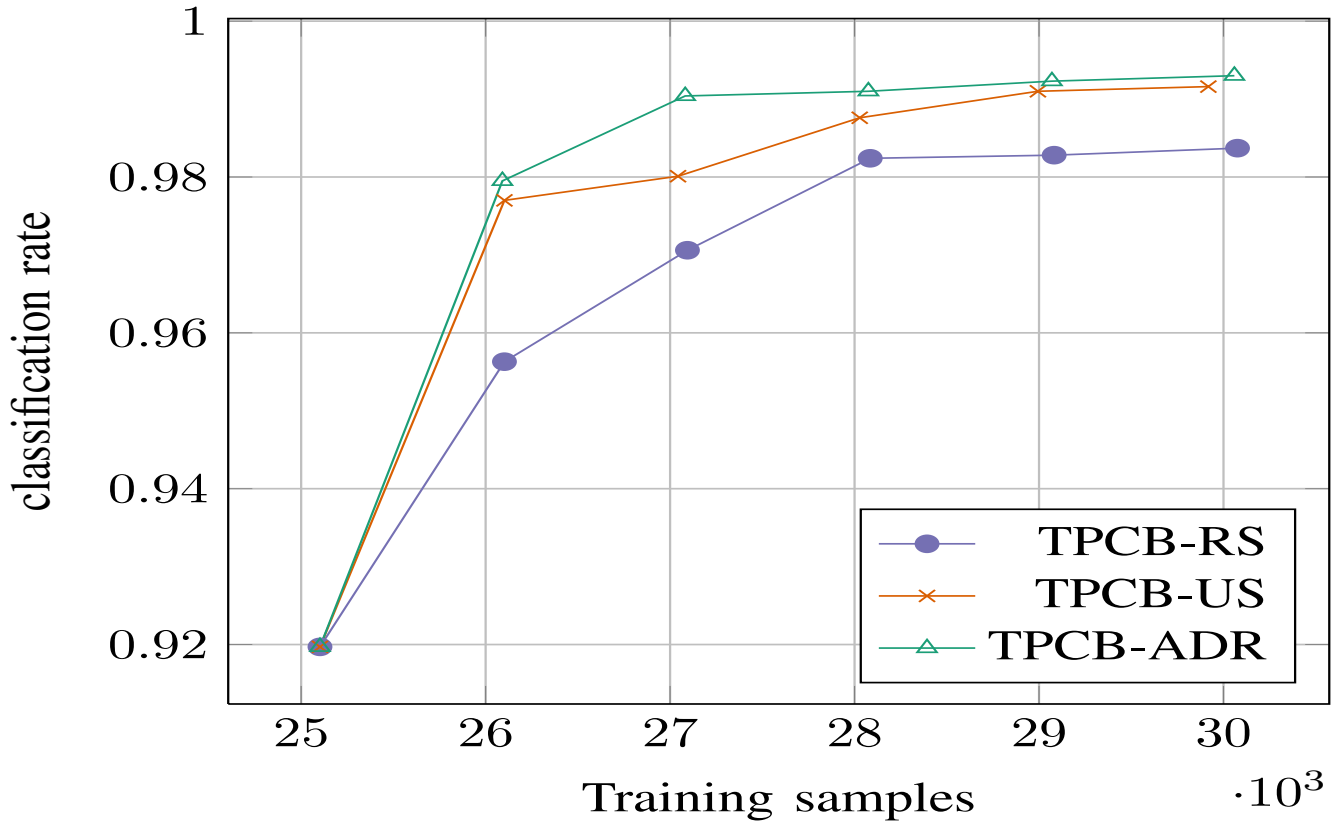
[Nienhüser, Zöllner, IV2013]

Aktives Lernen Verkehrszeichenerkennung

- Vorgeschlagene Beispiele und Experten Klassifikation
- Komplexe Beispiele

positive	 80 km/h	 80 km/h	 100 km/h
unsure	 100 km/h	 60 km/h	 80 km/h
negative	 40 km/h	 max width restriction	 80 km/h (suspended)

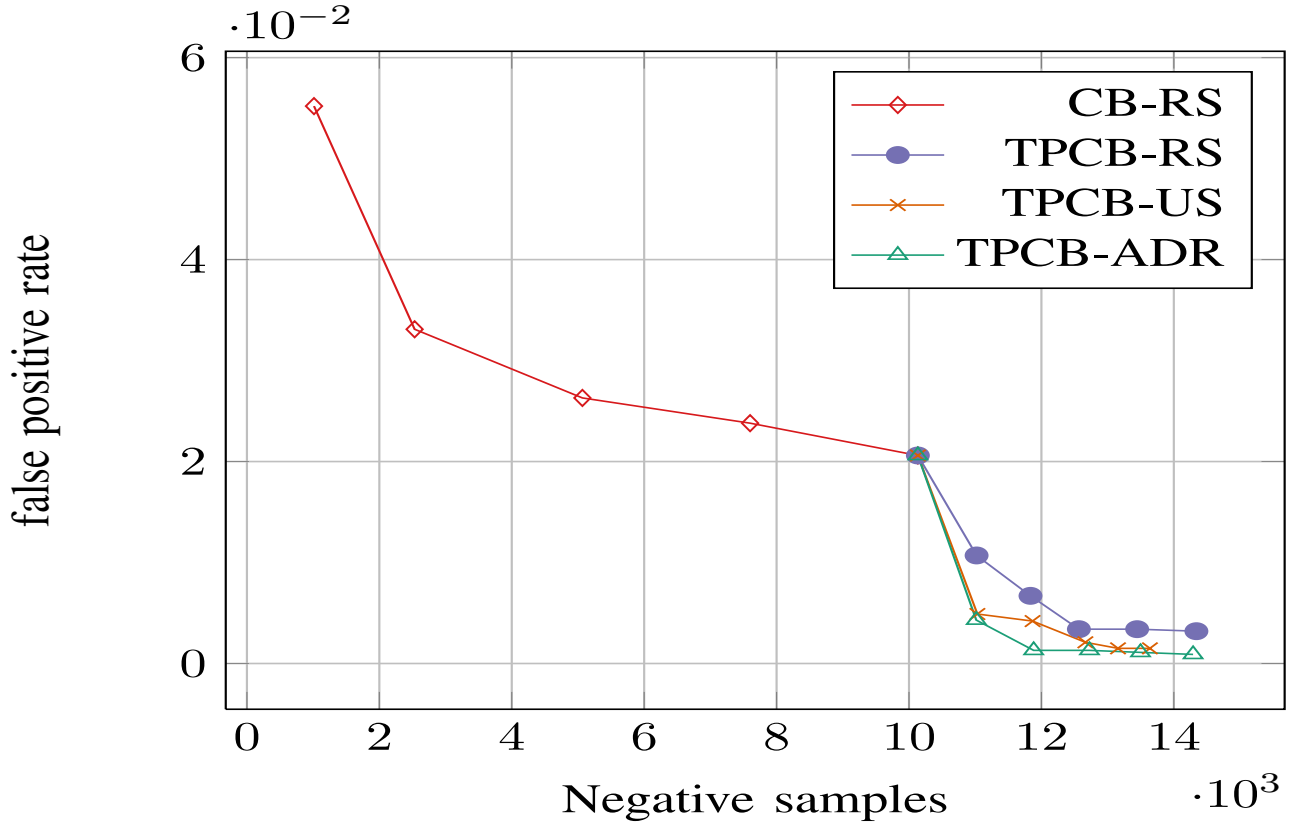
Aktives Lernen Verkehrszeichenerkennung



(RS – random sampling, ADR – Aktive Suche)

- Neue Beispiele → Klassifikationsrate steigt sehr schnell

Aktives Lernen Verkehrszeichenerkennung



(RS – random sampling, ADR – Aktive Suche)

- Neue Negative Beispiele → Falsch Positiv Rate fällt sehr schnell

Literatur

Chapelle, Schöllkopf, Zien: „Semi – Supervised Learning“,

- MIT – Press, 2010

X. Zhu: „Semi-Supervised Learning Literature Survey“

- http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf

K-R Müller, A.Zien: „Semi-Supervised Learning“

- Folien: https://ml01.zrz.tu-berlin.de/wiki/Main/SS09_MaschinellesLernen2

B. Schiele: Vorlesung ML

- Folien: http://www.mis.tu-darmstadt.de/ml2

Literatur

Burr Settles: „Active Learning Literature Survey “,

- Computer Sciences Technical Report 1648, University of Wisconsin–Madison

CMU: Vorlesung ML 2010/2011

- Folien: http://www.cs.cmu.edu/~tom/10701_sp11/lectures.shtml
- Folien: http://www.cs.cmu.edu/~epxing/Class/10701/lecture.html