Approximate Maximum Margin Algorithms with Rules Controlled by the Number of Mistakes

Seminar Maschinelles Lernen

VORTRAGENDER: SEBASTIAN STEINMETZ

BETREUT VON: ENELDO LOZA MENCÍA

Inhalt

- Vorbedingungen an die **Daten**
- Perceptron-Like Large Margin Classifiers
- Der Neue: **MICRA**
  - Herleitung & **Konvergenzbetrachtungen**
  - Effiziente **Implementierung**
- Das Experiment: **MICRA** gegen den Rest der Welt
- Zusammenfassung

Die Daten

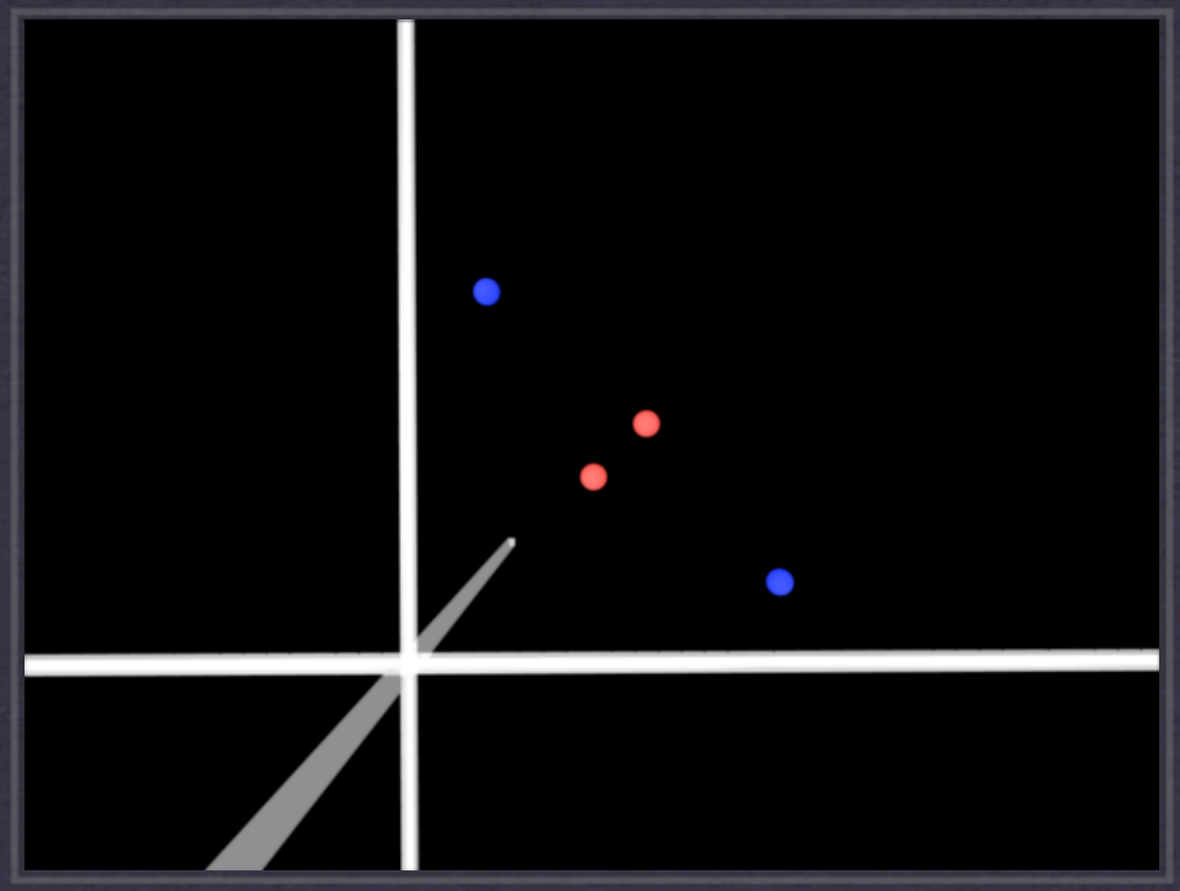
- Linear separierbar
  - wenn nicht -> Transformation in einen höheren Raum („Considered Space“)
- Augmentierung des Raums um eine weitere Dimension.
  - daraus folgt: Trennende Hyperebene verläuft durch den Ursprung.

Die Daten (II)

- Fallunterscheidung (Positiv-/Negativbeispiel)
- Punktspiegelung am Ursprung
- Einheitliche Darstellung in den Formeln

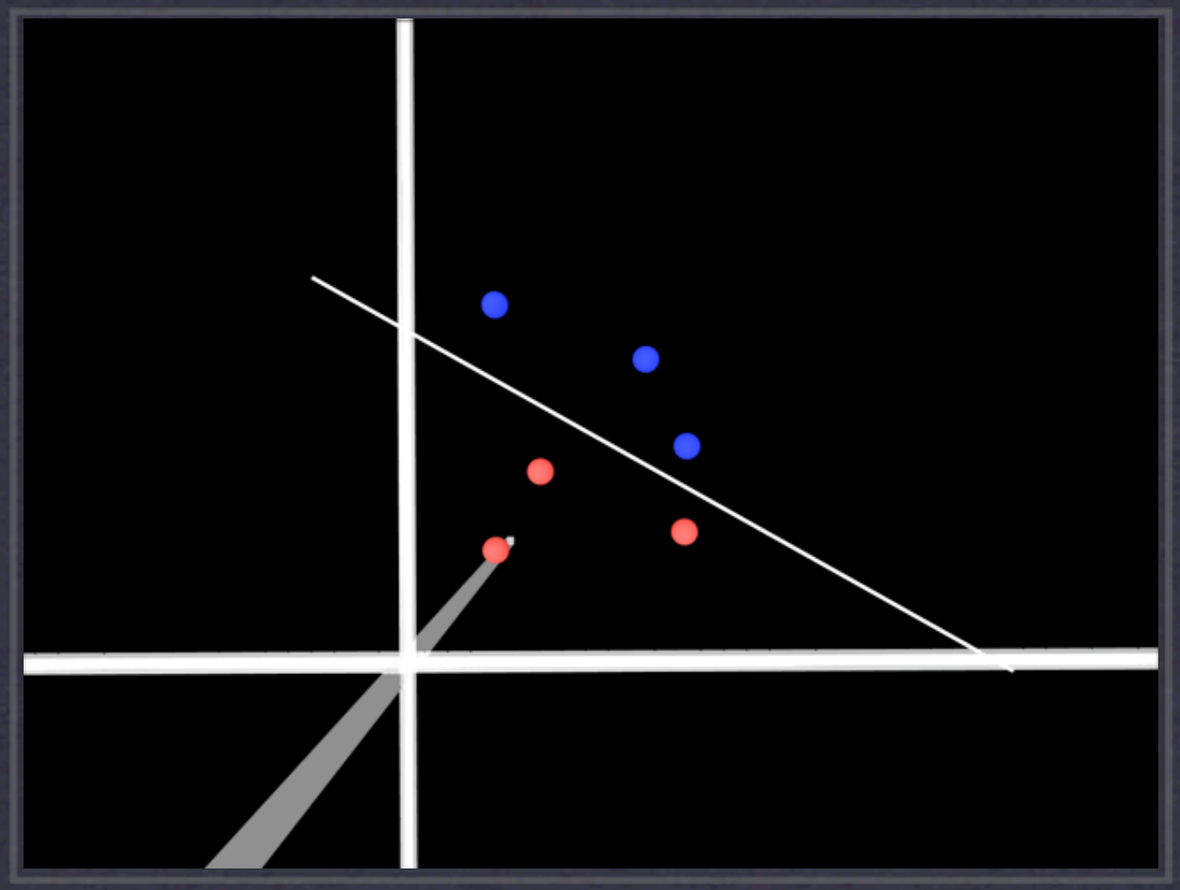
Lineare Separierbarkeit

Was tun, wenn sie nicht gegeben ist?



Augmentierung des Raums

Wie sieht das aus, und welchen Vorteil hat man davon?

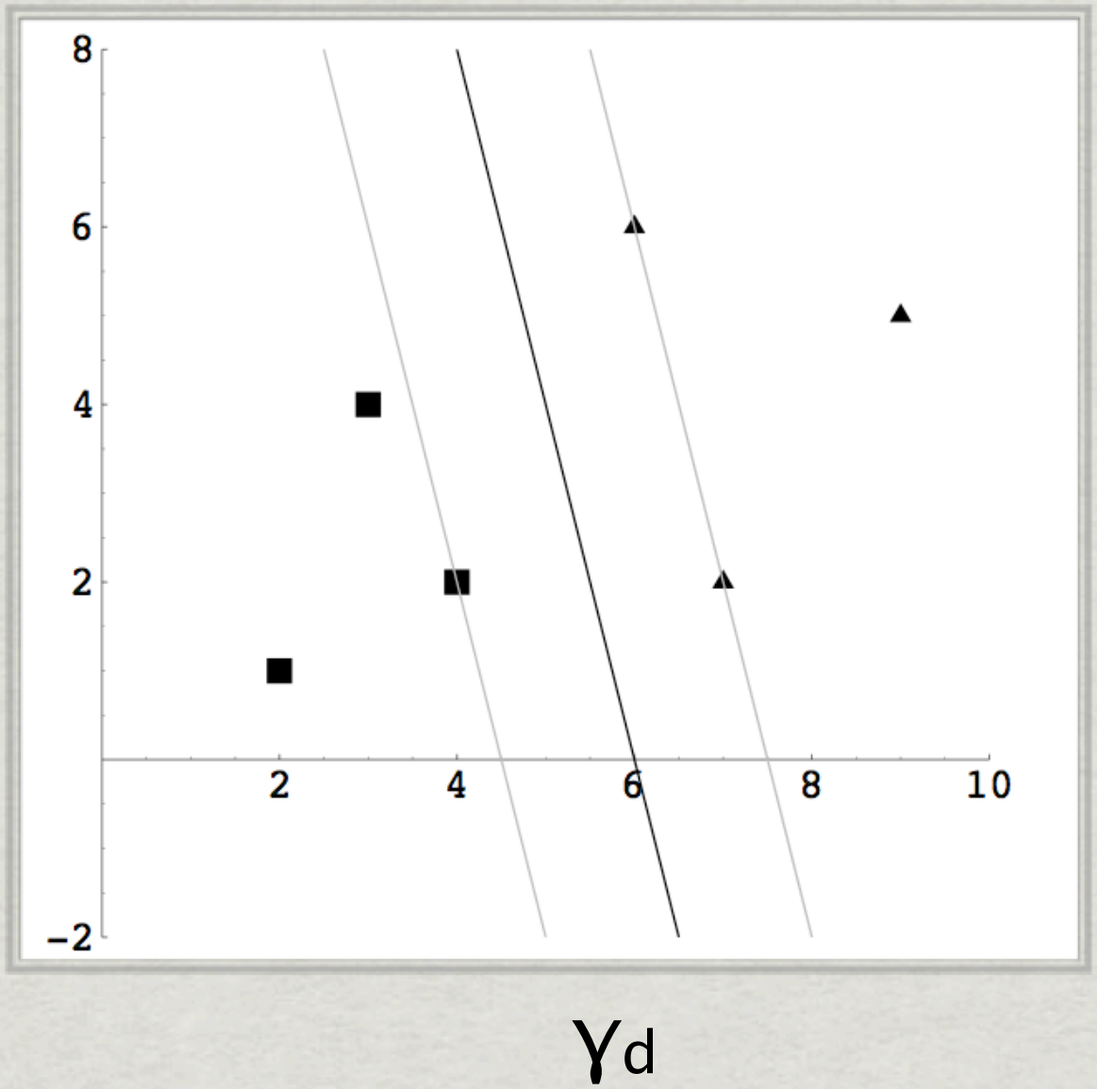


Inhalt

- ✓ Vorbedingungen an die **Daten**
- Perceptron-Like Large Margin Classifiers
  - Der Neue: **MICRA**
    - Herleitung & **Konvergenzbetrachtungen**
    - Effiziente **Implementierung**
  - Das Experiment: **MICRA** gegen den Rest der Welt
  - Zusammenfassung

Large Margin Klassifikation

- Gesucht: Die Lineare Hyperebene, welche genau zwischen den Clustern liegt bei maximalem Abstand.
- Beispiel: Support Vector Machines (SVM) oder Perceptron Like Algorithms (PLA)



PLAs - formal

- Update-Regel, Aktiviertheit

$$\boldsymbol{a}_{t+1} = (\boldsymbol{a}_t + \eta_t f_t \boldsymbol{y}_k)\, N_{t+1}^{-1}$$

- Update-Regel, „Weight-Vectors“

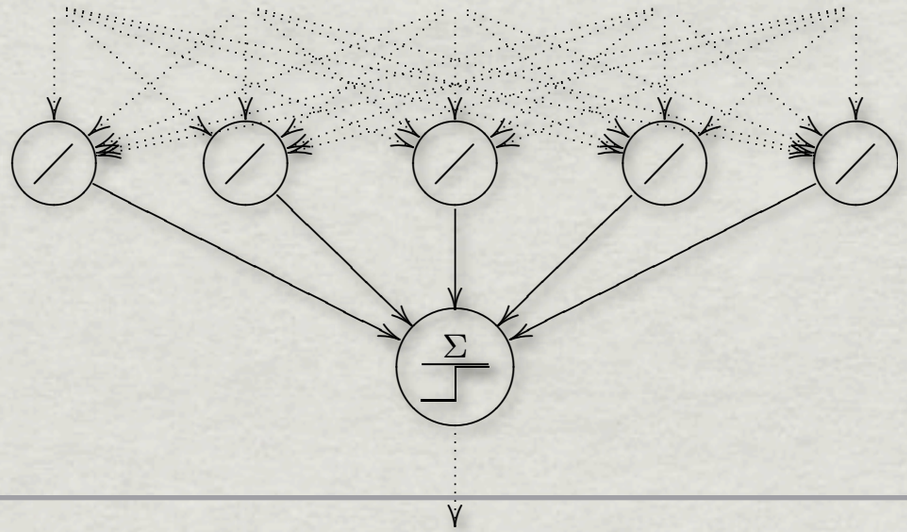
$$\boldsymbol{u}_{t+1} = \frac{\boldsymbol{u}_t + \eta_{\mathrm{eff}\,t} f_t \boldsymbol{y}_k / R}{\|\boldsymbol{u}_t + \eta_{\mathrm{eff}\,t} f_t \boldsymbol{y}_k / R\|}$$

- u ist nur normiertes a

$$\boldsymbol{u}_t \equiv \boldsymbol{a}_t / \|\boldsymbol{a}_t\|$$

- Effektive Lernrate:

$$\eta_{\mathrm{eff}\,t} \equiv \eta_t R \|\boldsymbol{a}_t\|^{-1}$$



PLAs - formal (II)

- Misclassification Condition

$$\boldsymbol{u}_t \cdot \boldsymbol{y}_k \leq C(t)$$

- C(t) -> 0 für große t

- Dann erhält man einen möglichst großen Korridor

$$\boldsymbol{u} \cdot \boldsymbol{y}_k \geq \gamma_{\mathrm{d}} \equiv \max_{\boldsymbol{u}':\|\boldsymbol{u}'\|=1} \min_i \{\boldsymbol{u}' \cdot \boldsymbol{y}_i\} \qquad \forall k$$

Inhalt

- ✓ Vorbedingungen an die **Daten**
- ✓ Perceptron-Like Large Margin Classifiers (**PLAs**)
- Der Neue: **MICRA**
  - Herleitung & **Konvergenzbetrachtungen**
  - Effiziente **Implementierung**
- Das Experiment: **MICRA** gegen den Rest der Welt
- Zusammenfassung

MICRA$^{\epsilon,\zeta}$

- Update-Rule:

$$\boldsymbol{u}_{t+1} = \frac{\boldsymbol{u}_t + \eta_{\text{eff}\,t} f_t \boldsymbol{y}_k / R}{\|\boldsymbol{u}_t + \eta_{\text{eff}\,t} f_t \boldsymbol{y}_k / R\|}$$

$$\boldsymbol{u}_{t+1} = \frac{\boldsymbol{u}_t + \frac{\eta \boldsymbol{y}_k}{t^\zeta R}}{||\boldsymbol{u}_t + \frac{\eta \boldsymbol{y}_k}{t^\zeta R}||}$$

- Misclassification Condition:

$$\boldsymbol{u}_t \cdot \boldsymbol{y}_k \leq C(t)$$

$$\boldsymbol{u}_t \cdot \boldsymbol{y}_k \leq \frac{\beta}{t^\epsilon}$$

Konvergenz

- für $\zeta \leq 1$ Konvergenz in endlicher Schrittzahl
- wenn $\eta = \eta_0(\beta/R)^{-\delta}$ konvergiert der „Margin“ für $\beta/R \rightarrow \infty$ gegen den maximalen Margin γ_d vorausgesetzt $0 < \epsilon\delta+\zeta < 1$
- Für $\zeta + 2\epsilon = 1$ mit $\zeta > 1/2$
  - obere Schranke für die Anzahl der benötigten Updates (t_b)
  - untere Schranke für den Anteil, des Margins, den der Algorithmus erzielt (f_b)

Algorithm 1 MICRA$^{\epsilon,\zeta}$

Input: A linearly separable augmented set with reflection assumed $S = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k, \ldots, \boldsymbol{y}_m)$

Fix: $\eta, \ \beta$

Define: $R = \max_k \|\boldsymbol{y}_k\|, \ \ q_k = \|\boldsymbol{y}_k\|^2, \ \ \bar{\eta} = \eta / R$

Initialise: $t = 1, \ \ \boldsymbol{a}_1 = \boldsymbol{y}_1, \ \ \|\boldsymbol{a}_1\| = \|\boldsymbol{y}_1\|,$
$\eta_1 = \|\boldsymbol{a}_1\| \bar{\eta}, \ \ \beta_1 = \|\boldsymbol{a}_1\| \beta$

repeat
  for $k = 1$ **to** m **do**
    $p_{tk} = \boldsymbol{a}_t \cdot \boldsymbol{y}_k$
    - **if** $p_{tk} \leq \beta_t$ **then**
      - $\boldsymbol{a}_{t+1} = \boldsymbol{a}_t + \eta_t \boldsymbol{y}_k$
      - $\|\boldsymbol{a}_{t+1}\| = \sqrt{\|\boldsymbol{a}_t\|^2 + \eta_t \left(2 p_{tk} + \eta_t q_k\right)}$
      - $t \leftarrow t + 1$
        $\eta_t = \|\boldsymbol{a}_t\| \bar{\eta} t^{-\zeta}, \ \ \beta_t = \|\boldsymbol{a}_t\| \beta t^{-\epsilon}$
    end if
  end for
until no update made within the **for** loop

misclassification condition

update-rule

gilt jetzt (Kommentar)

t++

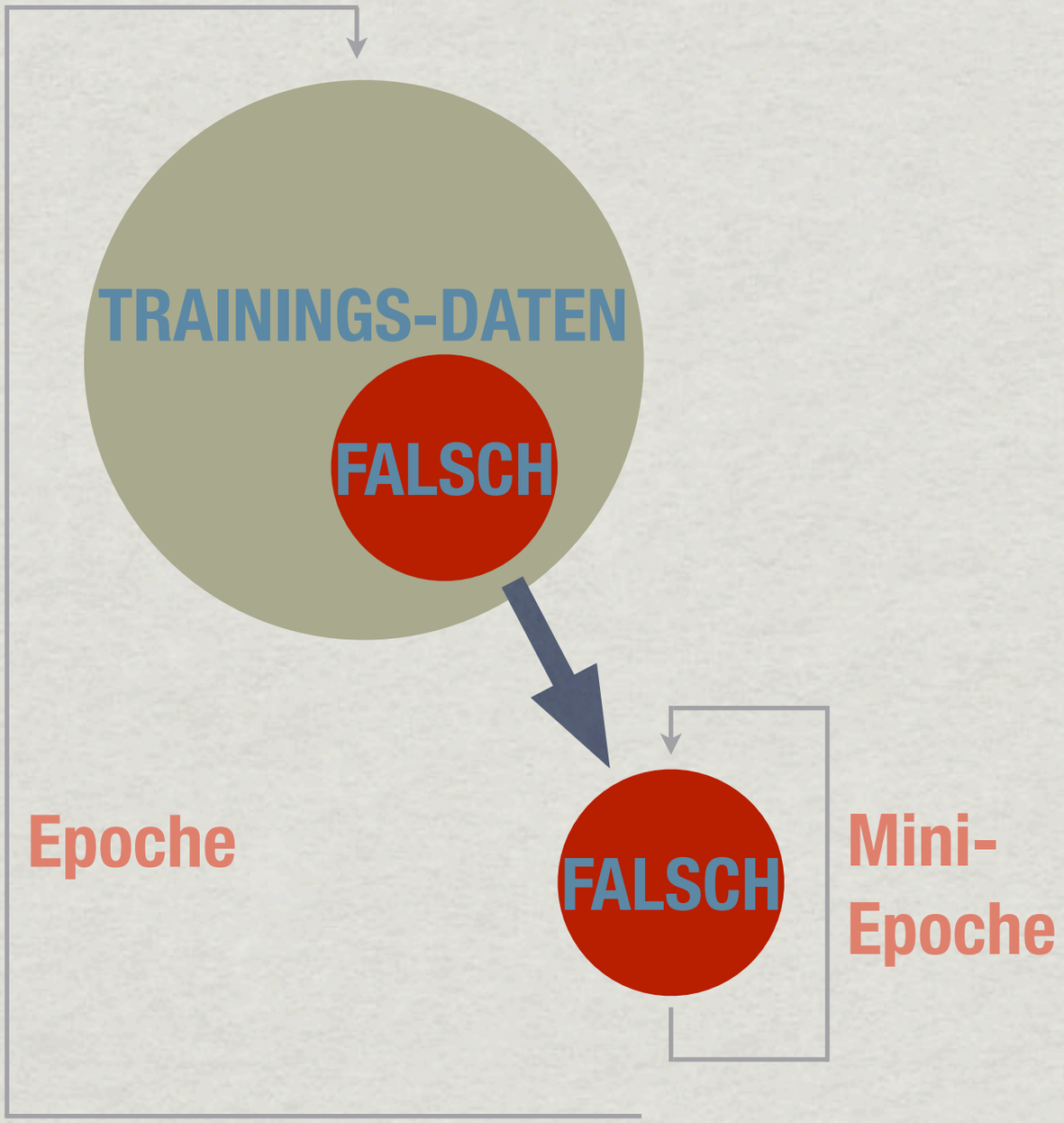
ALGORITHMUS

MICRA$^{\epsilon,\zeta}$

IMPLEMENTIERUNG IN PSEUDOCODE

Weitere Optimierung-möglichkeiten

- Epoche: einmal alle Traningsbeispiele dem Algorithmus zeigen
- bilden eines reduzierten „**active set**“
- enthält nur, in der aktuellen Epoche, falsch klassifizierte Beispiele
- dieser Algorithmus wird **red-MICRA** genannt



Inhalt

- ✓ Vorbedingungen an die **Daten**
- ✓ Perceptron-Like Large Margin Classifiers (**PLAs**)
- ✓ Der Neue: **MICRA**
  - ✓ Herleitung & **Konvergenzbetrachtungen**
  - ✓ Effiziente **Implementierung**
- ▸ Das Experiment: **MICRA** gegen den Rest der Welt
- ❋ Zusammenfassung

MICRA vs. PLAs

- MICRA wird mit „sinnvoll gewählten“ Parametern gegen agg-ROMMA und den normalen Perceptron-Algorithmus getestet.
- ROMMA: Relaxed Online Maximum Margin Algorithm
- Testdaten aus dem Machine Learning Repository der UCI (University of California - Irvine)

Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

Experiment 1 (sonar classification prob.) linear

- Sonar-Signale von zylindrischen Metall-gegenständen

vs.

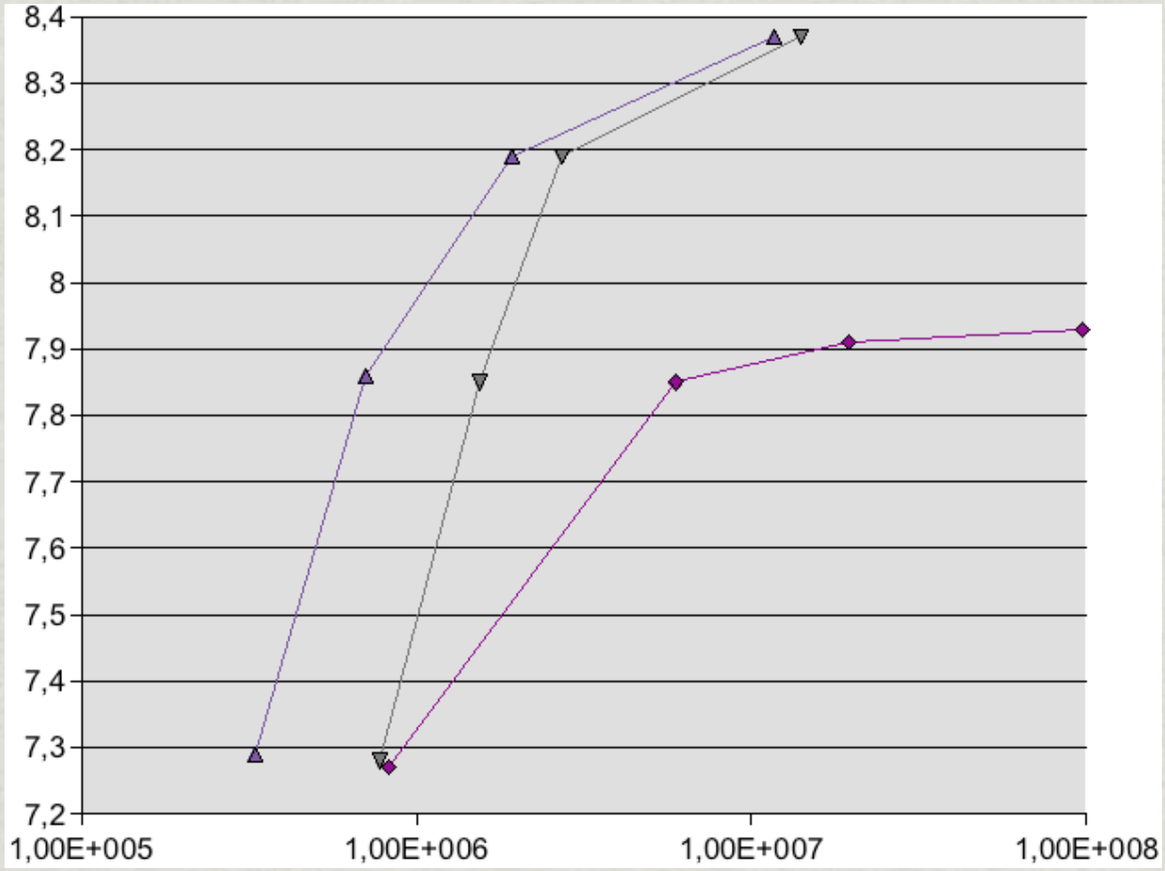
- Sonar-Signale von Zylindrischen Steinen

Instanzen	208
Trainingsinstanzen	104
Attribute	60
Fehlerhafte Daten	0
Entfernte Daten	0
gewähltes ρ	1
➞ R ≃	3,8121
➞ γ_d ≃	0,00841

Experiment 1 (sonar classification prob.) linear

Perceptron		agg-ROMMA		MICRA$^{0.05,0.9}$	
$10^3\gamma'_d$	upds	$10^3\gamma'_d$	upds	$10^3\gamma'_d$	upds
7,27	820.261	7,28	778.412	7,29	327.468
7,85	5.930.214	7,85	1.546.595	7,86	706.274
7,91	19.599.882	8,19	2.716.711	8,19	1.932.165
7,93	97.717.549	8,37	14.079.715	8,37	11.610.899

$\eta = 50$



Experiment 2.1 (wisconsin breast cancer)

- Gutartiger Tumor

vs.

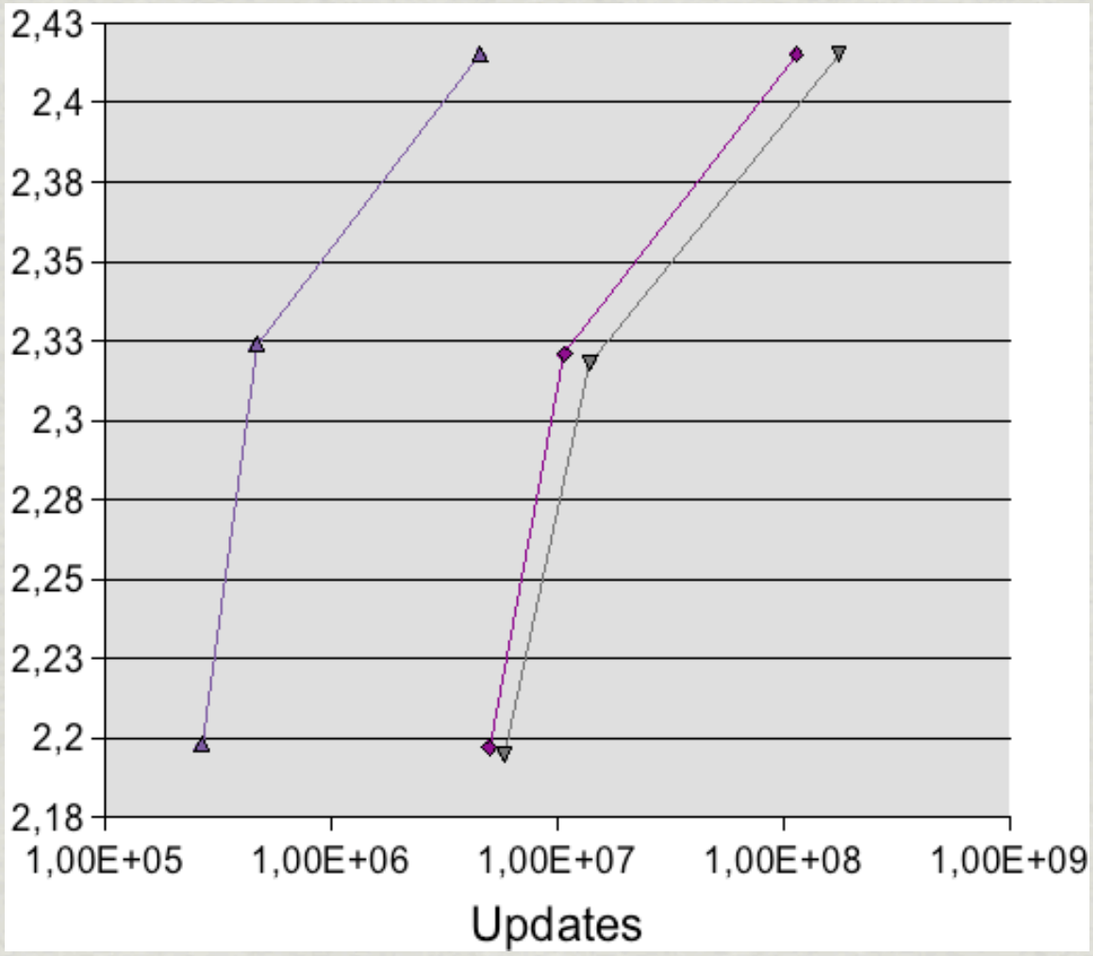
- Bösartiger Tumor

- Zwecks linearer Separierbarkeit, entfernen von 11 Datenpunkten

Instanzen	699
Trainingsinstanzen	672
Attribute	9
Fehlerhafte Daten	16
Entfernte Daten	(11)
gewähltes ρ	30
→ $R \simeq$	41,4246
→ $\gamma_d \simeq$	0,0243

Experiment 2.1 (WBC_{-11}) linear

Perceptron		agg-ROMMA		MICRA$^{0.1,0.8}$	
$10^3\gamma'_d$	upds	$10^3\gamma'_d$	upds	$10^3\gamma'_d$	upds
2,197	4.980.423	2,195	5.784.868	2,198	267.145
2,321	10.761.773	2,318	13.931.792	2,324	467.369
2,415	113.406.210	2,415	174.388.827	2,415	4.533.155



$\eta = 2,3$

Experiment 2.2 (wisconsin breast cancer)

- Gutartiger Tumor

vs.

- Bösartiger Tumor

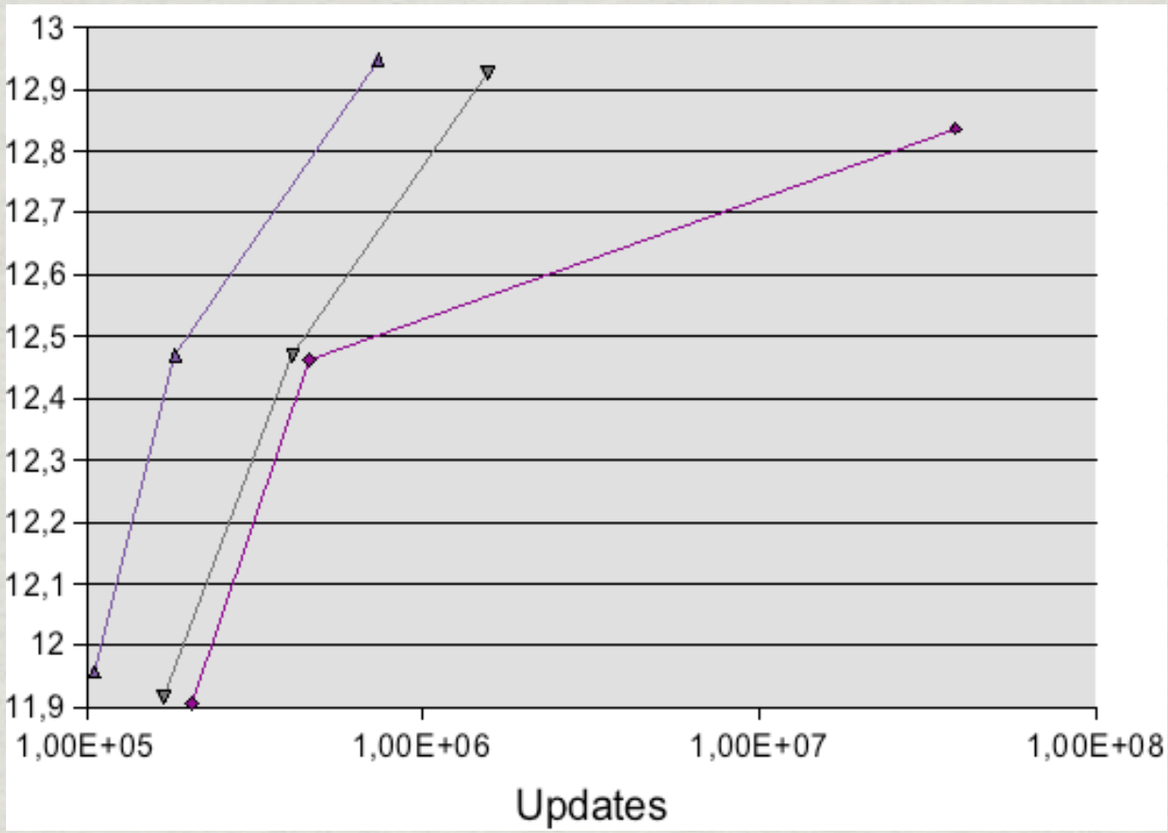
- Vollständiges WBC-Set (-> linear nicht separierbar)

Instanzen	699
Trainingsinstanzen	672
Attribute	9
Fehlerhafte Daten	16
Entfernte Daten	0
gewähltes ρ	10
→ $R \simeq$	30,282
→ $\gamma_d \simeq$	0,13033

Experiment 2.2 (WBC) non-linear

Perceptron		agg-ROMMA		MICRA$^{0.05,0.9}$	
$10^3\gamma'_d$	upds	$10^3\gamma'_d$	upds	$10^3\gamma'_d$	upds
11,905	206.469	11,916	169.588	11,957	105.964
12,462	457.334	12,468	409.956	12,47	183.643
12,837	38.336.601	12,928	1.554.492	12,949	734.629

$\eta = 20$



MICRA vs. SVMs

- PLAs konvergieren in der nähe der optimalen Hyperebene extrem langsam
- Deswegen: Anforderung an γ lediglich 99% des maximalen margins
- Vergleich nur auf Prozessorzeit-Ebene, da SVMs nicht Epochen-Basiert arbeiten.

MICRA vs. SVMs

- LIBSVM und SVM^{light}
- SVMs erlauben weiche Ränder, feature space ist aber hart separierbar
- Dekompositions-Basierte SVMs. Viel schneller als Standard-SVMs
- MICRA wird vertreten durch red-MICRA, also Training mit Hilfe von Micro-Epochen
- Stop von red-MICRA, wenn

  $\gamma_M > \gamma_{S<}$ UND $\gamma_M > \gamma_{S>} \cdot 0{,}99$

data set	Δ	LIBSVM				SVM^{light}				red − MICRA$^{0.05,0.9}$					
		$10^2\gamma'$	Secs	$10^2\gamma'$	Secs	$10^2\gamma'$	Secs	$10^2\gamma'$	Secs	ρ	η	N	$10^5\frac{\beta}{R}$	$10^2\gamma'$	Secs
sonar	0	0.8451	0.17	0.8405	0.10	0.8460	6.85	0.8388	4.84	1	45	80	462.2	0.8406	3.60*
ionosphere	1	10.554	0.06	10.389	0.05	10.551	0.30	10.448	0.19	1.5	10	10	2929	10.449	0.07
votes	1	16.846	0.02	16.708	0.02	16.841	0.18	16.690	0.11	1	5	20	6385	16.718	0.02
WBC	1	13.034	0.12	12.848	0.09	13.033	0.81	12.929	0.45	2	25	20	837.6	12.932	0.35
tic-tac-toe	1	10.300	0.47	10.183	0.27	10.295	3.35	10.185	1.35	0.5	8	20	5334	10.203	0.05
german	25	95.361	0.62	94.055	0.45	95.332	2.96	94.217	1.82	8	30	50	908.9	94.415	0.36
mushroom	0	36.551	0.58	35.988	0.33	36.538	0.17	36.103	0.11	0	4.5	50	12535	36.212	0.10

Margins der Algorithmen: $\epsilon = 0{,}001$; $\epsilon > 0{,}001$; anhalte-margin

EXPERIMENT

3: MICRA vs. SVMs

VERSCHIEDENE UCI-DATENSÄTZE

Experiment 4 (subsets des Adult-Datensatzes)

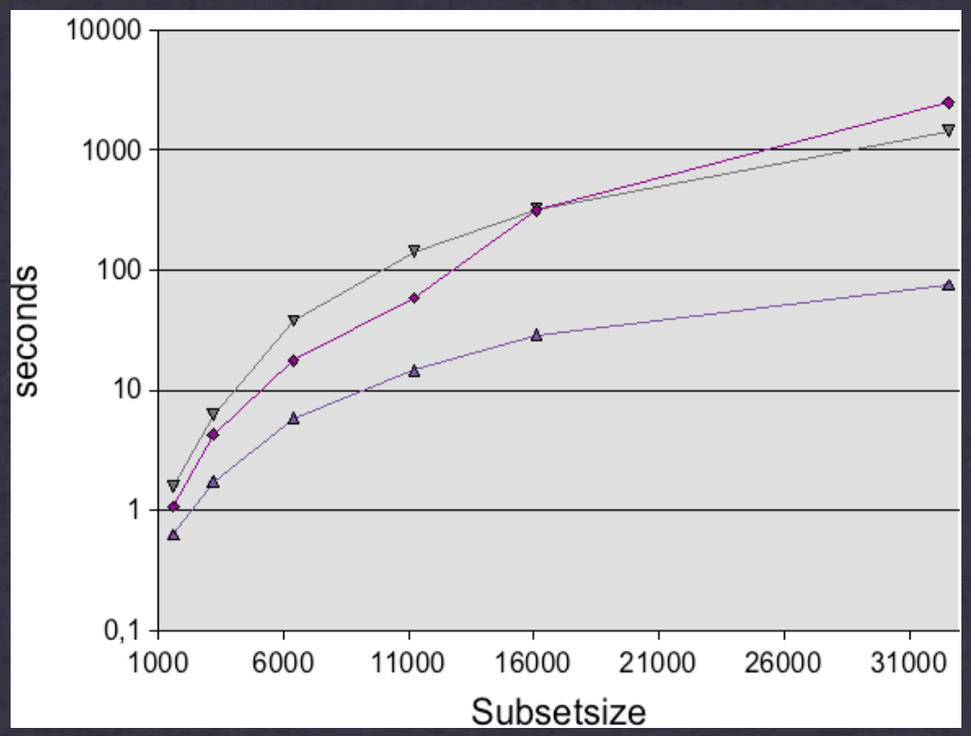
- Einkommen
  < $ 50k / Jahr

vs.

- Einkommen
  > $ 50k / Jahr

- Augmentierung nicht notwendig -> ρ=0

Instanzen	48842
Trainingsinstanzen	1.605-32.561
Attribute (binär)	14 (123)
Fehlerhafte Daten	0
Entfernte Daten	0
gewähltes ρ	0



subset size	LIBSVM				SVM^{light}				$red-MICRA^{0.05,0.9}$				
	$10^2\gamma'$	Secs	$10^2\gamma'$	Secs	$10^2\gamma'$	Secs	$10^2\gamma'$	Secs	η	N	$10^2\frac{\beta}{R}$	$10^2\gamma'$	Secs
1605	3.9383	1.41	3.9022	1.07	3.9375	3.02	3.8877	1.58	20	100	1.918	3.9038	0.63
3185	2.7437	5.55	2.7187	4.29	2.7434	11.3	2.7093	6.23	25	100	1.400	2.7187	1.73
6414	1.9292	22.5	1.9094	17.6	1.9290	71.3	1.9097	37.7	45	300	1.025	1.9111	5.83
11220	1.4499	73.2	1.4348	58.6	1.4497	283.4	1.4342	141.7	65	300	0.798	1.4356	14.7
16100	1.2069	389.7	1.1927	312.3	1.2068	638.2	1.1923	318.6	80	500	0.673	1.1950	28.7
32561	0.8526	3902.3	0.8424	2484.5	0.8525	2733.8	0.8432	1439.4	105	600	0.492	0.8441	75.0

Margins der Algorithmen: $\epsilon = 0{,}03$; $\epsilon = 0{,}025$; anhalte-margin

4: MICRA vs. SVMs

ADULT-DATENSATZ

Experiment 5 (subsets des Web-Datensatzes)

- Fichten-/Tannenwald

vs.

- Jede andere Art bewaldung

- Linear nicht separierbar, Δ=10

- Für SVM^{light} genauigkeit: ϵ=0,01

data size	SVM^{light}		red − MICRA$^{0.05,0.9}$				
	$10^3\gamma'$	Secs	η	N	$10^5\frac{\beta}{R}$	$10^3\gamma'$	Secs
581012	15.774	47987.7	70	400	336	15.789	4728.0

Instanzen	581012
Trainingsinstanzen	581012
Attribute	54
Fehlerhafte Daten	0
Entfernte Daten	0
gewähltes ρ	2

Zusammenfassung

- MICRA ist ein **schnell** konvergierender Perceptron-Like-Large-Margin-Classifier
- **Geringer** Speicherbedarf

	Perceptron	MICRA	SVM
Genauigkeit	o	+	++
Geschwindigkeit	o	++	o+
Speicherbedarf	++	++	o

Vielen Dank für Ihre Aufmerksamkeit