Mustererkennung mit Dünnen Gittern

Jochen Garcke, Prof. Dr. Michael Griebel
Universität Bonn
Dr. Michael Thess
Prudential Systems
Dr. Ulrich Windheuser
West LB

BMBF-Statusseminar 11.12-12.12.2000
Frankfurt

Das diesem Bericht zugrundeliegende Vorhaben wird mit Mitteln des Bundesministeriums für Forschung und Technologie unter dem Förderkennzeichen 03GRM6BN gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Das Klassifikationsproblem

Wir betrachten eine gegebene Menge bereits klassifizierter Daten

$$S = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i=1}^M.$$

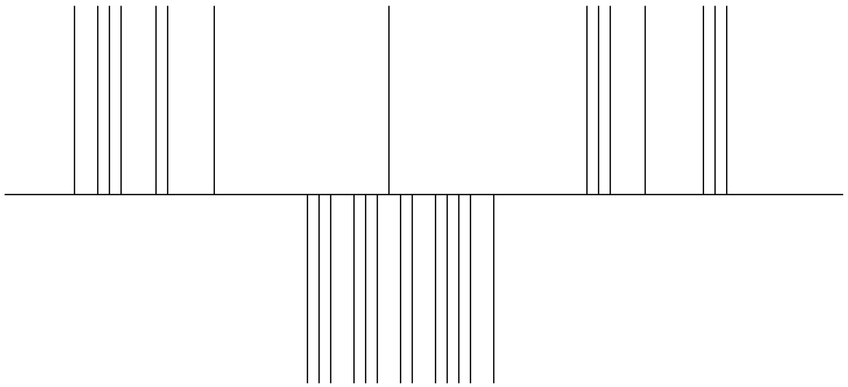
Ziel ist es, eine Funktion, den Klassifikator, zu bestimmen, die zum einen die gegebenen Daten gut approximiert, zum anderen auch auf ungesehenen Daten 'gute' Ergebnisse liefert.

Es ist dabei abzuwägen zwischen der Genauigkeit der Approximation, d.h. der Größe des Datenfehlers und der angenommenen Glattheit der Funktion, bestimmt durch einen Regularisierungsoperator.

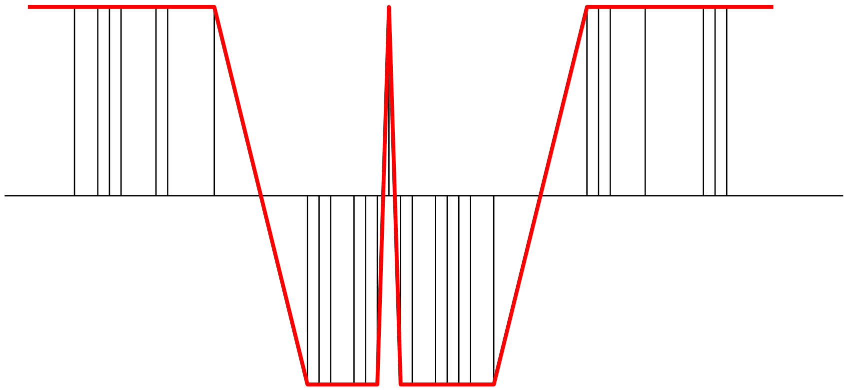
Anwendungen sind z.B. in der Wirtschaft

- Kreditvergabe
- Aufdeckung von Betrugsversuchen
- Auswahl von Adressaten für Werbebriefe

1D Beispiel

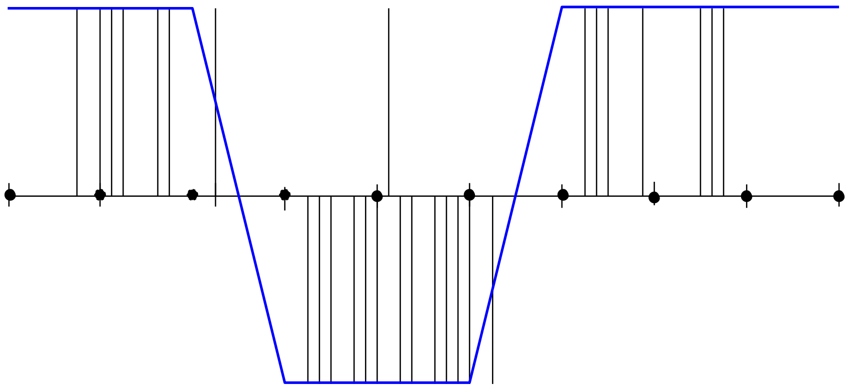


1D Beispiel



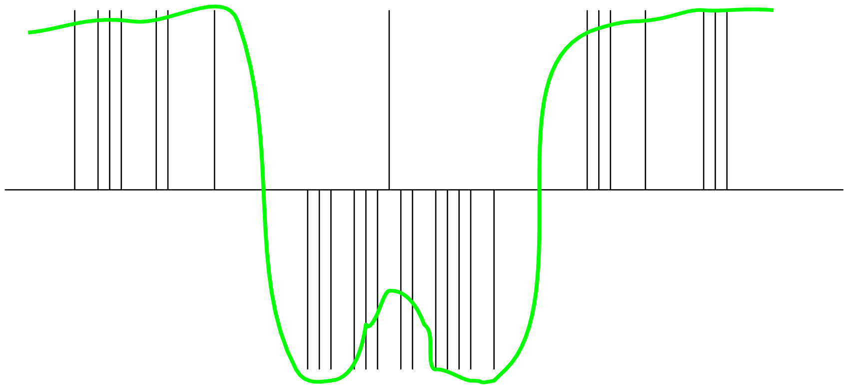
- Fehler an den Daten ist 0
- Aber f ist overfitted, möglicherweise schlechte Verallgemeinerung

Gitterbasierte Basisfunktionen



- Hier nur (1,-1) auf den Gitterpunkten, i.A. kontinuierliche Werte

1D Beispiel



- Beschreibung der Glattheit durch Ableitungen von f
- Praxis: hochdimensionale Probleme

Wahl des Funktionenraum V_h

Weitverbreitet sind Verfahren mit globalen datenzentrierten Basisfunktionen, die in der Berechnung mit der Anzahl der Daten skalieren, wie z.B. radiale Basisfunktionen, neuronale Netze, Support Vektor Maschinen

Wir benutzen ein Gitter zur Diskretisierung des Datenraumes und lokale Basisfunktionen auf den Gitterpunkten

Ein naives Gitter hat N^d Gitterpunkte, bei sinnvollen Größen von N scheitert man am Fluch der Dimension

Abhilfe ist die Verwendung von dünnen Gittern mit $O(N\log(N)^{d-1})$ Gitterpunkten

Regularisierungsnetzwerke für Klassifikation

Wir wollen den Fehler im Quadrat minimieren:

$$R_{reg}(f_h) = \frac{1}{M}\sum_{i=1}^{M}(f_h(x_i) - y_i)^2 + \lambda\|\hat{P}f_h\|^2_{L_2}, \quad f_h \in V_h$$

im Raum V_h, d.h. $f_h = \sum_{j=1}^{r}\alpha_j\varphi_j(x)$. Umformen und setzen von $\hat{P} = \nabla$ ergibt $(k = 1, \ldots, r)$

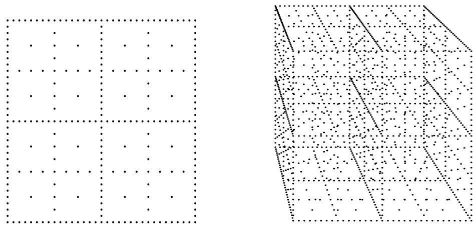
$$\sum_{j=1}^{r}\alpha_j\left[M\lambda(\nabla\varphi_j, \nabla\varphi_k)_{L_2} + \sum_{i=1}^{M}\varphi_j(x_i)\cdot\varphi_k(x_i)\right] = \sum_{i=1}^{M} y_i\varphi_k(x_i)$$

Damit erhalten wir das zu lösende lineare Gleichungssystem

$$(\lambda C + B \cdot B^T)\alpha = By.$$

mit $C_{j,k} = M \cdot (\nabla\varphi_j, \nabla\varphi_k)_{L_2}$
und $B_{j,i} = \varphi_j(\boldsymbol{x}_i), i = 1, \ldots M, j = 1, \ldots r$

Dünne Gitter



Dünnes Gitter in 2D und 3D mit Tiefe $n = 5$

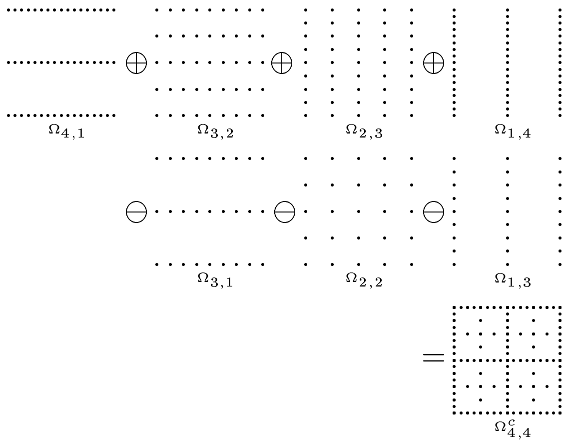
	volles Gitter	dünnes Gitter
Anzahl Punkte	$O(N^d)$	$O(N\log(N)^{d-1})$
Approxim.-güte	$O(N^{-2})$	$O(N^{-2}\log(N)^{d-1})$
Glattheitsvoraus.	$u \in H^2$	$\|\frac{\partial^{2d}u}{\partial x_1^2 \ldots \partial x_d^2}\|_2 < \infty$

Beispiel: n=6, d.h. N = 65, d=6
Volles Gitter: 75 418 890 625 Punkte
Dünnes Gitter: 483 201 Punkte

Dünne Gitter mittels der Kombinationstechnik

$$f_n^{(c)}(\boldsymbol{x}) := \sum_{q=0}^{d-1}(-1)^q \binom{d-1}{q} \sum_{l_1+\ldots+l_d=n+(d-1)-q} f_{\mathbf{l}}(\boldsymbol{x})$$

mit den Ergebnissen $f_{\mathbf{l}}(\boldsymbol{x}) = \sum_{\mathbf{j}} \alpha_{\mathbf{l},\mathbf{j}}\phi_{\mathbf{l},\mathbf{j}}(\boldsymbol{x}) \in V_{\mathbf{l}}$
der verschiedenen Gitter $\Omega_{\mathbf{l}}$
Beispiel in zwei Dimensionen



Zu berechnende diskrete Probleme

Diskretisiere und löse das Problem

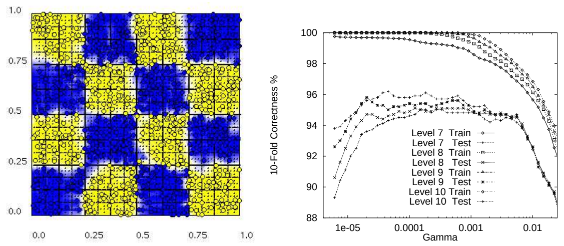
$$R_{reg}(f_{\mathbf{l}}) = \frac{1}{M}\sum_{i=1}^{M}(f_{\mathbf{l}}(x_i) - y_i)^2 + \lambda\|\hat{P}f_{\mathbf{l}}\|^2_{L_2}, \quad f_{\mathbf{l}} \in V_{\mathbf{l}}$$

nun auf $\Omega_{\mathbf{l}}$, $|\mathbf{l}|_1 = n + (d-1) - q, \quad q = 0, \ldots, d-1$

Die Anzahl der Probleme hat die Ordnung $O(dn^{d-1})$

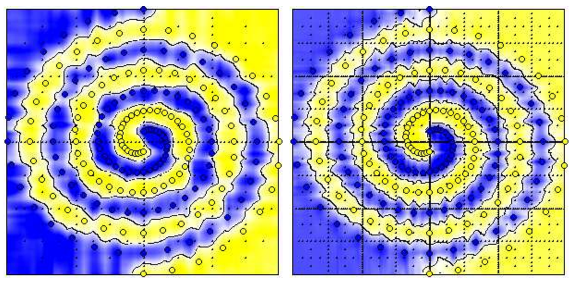
Die Dimension jedes Teilproblems dim($V_{\mathbf{l}}$) hat die Ordnung $O(N) = O(2^n)$, d.h. klein genug für den Hauptspeicher einer Workstation

Testbeispiel Checkerboard (2D)



Links: Kombinationstechnik mit Level 10, $\lambda = 4.53999 \cdot 10^{-5}$, Rechts: Plot der Abhängigkeit von λ (in log.-skalierung) und Level

Testbeispiel Spirale (2D)



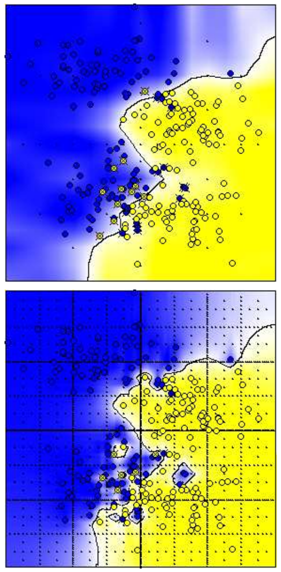
Dünnes Gitter mit Level 6 (links) und 8 (rechts), $\lambda = 0.001$

Testbeispiel BUPA Liver (6D)

	SVM		Dünngitter-Kombinationsmethode			
	SSVM	FSV	Level 1 $\lambda = 0.0001$	Level 2 $\lambda = 0.1$	Level 3 $\lambda = 0.0625$	Level 4 $\lambda = 0.625$
10-fold train. %	70.37	68.18	83.28	79.54	90.20	88.66
10-fold test. %	70.33	65.20	66.89	66.38	69.23	68.74

Ergebnisse für den BUPA liver Datensatz im Vergleich zu Support-Vektor-Maschinen.

Testbeispiel Ripley (2D)



Kombinationstechnik mit Level 4, $\lambda = 0.00248$ (links) und Level 8, $\lambda = 0.00166$ (rechts)

	Ten-Fold		Best
Level	10-Fold Test. %	Test Data %	Testing %
1	84.8	89.9	90.4
2	85.2	90.8	91.0
3	88.4	91.0	91.3
4	87.6	91.3	91.5
5	87.6	90.8	91.4
6	86.4	90.7	91.3
7	86.4	89.7	91.1
8	88.0	90.6	91.0
9	88.4	90.5	91.1
10	88.4	90.9	91.2

Sehr großer synthetischer 6D Datensatz

Tiefe	# der Punkte	Trainings-Genauig.	Test-Genauig.	gesamte Zeit (sek)	Datenmatrix Zeit (sek)	# der Iterat.
1	50 000	87.9	88.1	158	152	41
	500 000	88.0	88.1	1570	1528	44
	5 Million	88.0	88.0	15933	15514	46
2	50 000	89.2	89.3	1155	1126	438
	500 000	89.4	89.2	11219	11022	466
	5 Million	89.3	89.2	112656	110772	490

Sehr großer synthetischer 2D Datensatz

	# der Punkte	Trainings-genauigkeit	Test-genauigkeit	Zeit (sek)	# der Iterationen
Level 1	1 Million	95.1	95.0	26	9
	2 Million	95.1	95.0	54	9
	5 Million	95.1	95.0	132	9
Level 2	1 Million	94.7	94.7	50	41
	2 Million	94.7	94.7	103	41
	5 Million	94.7	94.7	253	41
Level 3	1 Million	94.9	94.8	75	115
	2 Million	94.9	94.8	150	117
	5 Million	94.9	94.8	373	117
Level 4	1 Million	94.9	94.8	99	283
	2 Million	94.9	94.8	199	285
	5 Million	94.9	94.8	496	287
Level 5	1 Million	94.9	94.8	124	666
	2 Million	94.9	94.8	247	669
	5 Million	94.9	94.8	620	676

Testbeispiel PIMA Indians (8D)

	SVM		Dünngitter-Kombinationsmethode		
	SSVM	RLP	Level 1	Level 2	Level 3
			$\lambda = 0.0001$	$\lambda = 0.1$	$\lambda = 1.0$
10-fache Train. genau., %	78.11	76.48	82.69	88.48	93.29
10-fache Test. genau., %	78.12	76.16	77.47	75.01	72.93

Ergebnisse für den PIMA Indians Datensatz im Vergleich zu Support-Vektor-Maschinen.

Zusammenfassung

- gitterbasiertes statt datenzentriertes Verfahren
- dünne Gitter durchbrechen den Fluch der Dimension
- gute erste Ergebnisse
- skaliert linear mit der Zahl der Daten, gut für große Datenmengen

Ausblick

- Weiterentwicklung des Verfahrens
- Realisierung der Parallelisierung der Kombinationstechnik
- Vergleich mit Neuronalen Netzen, Support Vektor Maschinen
- Nachbearbeitung des Klassifikators mittels Wavelet-Tresholding
- Untersuchung der Glattheit typischer Daten (Fehlerabschätzung)
- Anwendung der Erkenntnisse auf die Daten der Projektpartner