

Abteilung für Mathematik
Seminar für Statistik

Wintersemester 1999/2000

Diplomarbeit

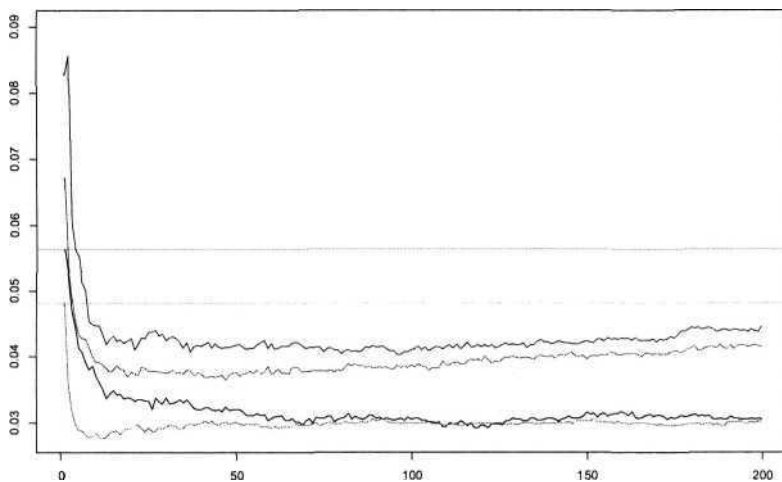
Josef Vogt

Diplomprofessor: **Prof. Peter Bühlmann**

Ausgabe: 1. November 1999

Abgabe: 7. März 2000

Bagging, Boosting und verwandte Methoden



Inhaltsverzeichnis

Abbildungsverzeichnis	viii
Tabellenverzeichnis	x
1 Bagging	1
1.1 Einführung	1
1.2 Mitteln	1
1.3 Vom Bootstrappen zum Bagging	2
2 Classification and Regression Trees (CART)	3
2.1 Einleitung	3
2.2 Die Bayes-Regel	4
2.3 Konstruktion eines Klassifizierungsbaumes	4
2.3.1 Wahl des richtigen Splits	6
2.3.2 Eine Stop-Splitting Regel	7
2.3.3 Zuweisungsregel für Endknoten	7
2.4 Baumgrösse und Schätzungen für $R^*(T)$	8
2.4.1 Minimal Cost-complexity pruning	8
2.4.2 Schätzung für $R^*(T)$	8
2.5 Regression Trees	9
2.5.1 Konstruktion des Regressionsbaumes	10
2.5.2 Pruning	11
3 Multivariate Adaptive Regression Splines (MARS)	13
3.1 Einführung	13
3.2 Parametrische Modelle	13
3.3 Regression durch rekursives Unterteilen in Subregionen	14
3.4 Adaptive Regression Splines	15
3.4.1 Vorwärtsalgorithmus	15
3.4.2 Rückwärtsalgorithmus	16
3.4.3 Stetigkeit	16
3.4.4 Das MARS-Verfahren	16
4 Bagging regression trees	17
4.1 Bringt Bagging überhaupt irgendwas?	17
4.1.1 Die Friedman#1 Daten	17
4.1.2 Vorhersage an einem bestimmten Punkt	17
4.1.3 Lage des Bias	19
4.1.4 m out of n mit und ohne Zurücklegen	19
4.2 Friedman#2 Simulation	22
4.3 Friedman#3 Simulation	22
4.4 Ozon Daten	22
4.5 Zusammenfassung	25
5 Bagging classification trees	27
5.1 Glasdaten	27
5.1.1 Simulation mit 'voting'	28

5.1.2	Simulation ohne 'voting'	28
5.2	Brustkrebsdaten.	30
5.3	Zusammenfassung.	30
6	Bagging MARS	33
6.1	Friedman#1 Simulation	33
6.2	Friedman#2 Simulation.	35
6.3	Friedman#3 Simulation.	36
6.4	Ozon Daten.	36
6.5	Warum funktioniert Bagging bei MARS nicht so gut?.	37
6.6	Vergleich von Bagging MARS und Bagging CART.	38
7	Bagging bei linearer Regression mit Modelselection	41
7.1	Modelselection mit AIC.	41
7.2	Modelselection bei Boston-Housing Daten.	42
7.3	Modelselection bei Ozon Daten.	44
7.4	Modelselection bei simulierten Daten.	44
7.5	Zusammenfassung.	45
8	Bagging Stumps	47
8.1	Erzeugen der Daten.	47
8.2	Die Bayes Missklassifikationsrate.	47
8.3	Durchführen der Simulation.	48
9	Boosting	51
9.1	Einleitung.	51
9.2	Diskreter AdaBoost.	51
9.3	Real AdaBoost.	52
9.4	AdaBoost - ein additives logistisches Regressionsmodell.	53
9.4.1	Ein exponentielles Kriterium.	53
9.4.2	Warum ein exponentieller Verlust.	54
9.5	Optimierung durch adaptive Newtonschritte.	55
9.6	Mehrklassen-Methoden.	55
9.7	L_2 -Boosting.	56
10	Simulationen mit dem Boosting-Algorithmus	57
10.1	Boosting mit simulierten Daten.	57
10.1.1	Bag-Boosting.	58
10.1.2	Bayes-MCR und optimaler Verlust.	58
10.1.3	Ergebnisse der Simulation.	59
10.2	Boosting mit den Brustkrebsdaten.	60
11	Bagging und Boosting bei Versicherungsdaten	65
11.1	Einleitung.	65
11.2	Klassifizierung im Versicherungs-Datensatz.	66
11.3	Bagging bei Versicherungsdaten.	67
11.3.1	Bringt Bagging bei Versicherungsdaten etwas?.	67
11.3.2	Subbagging $m = 1/8n$	68
11.4	Variation des Verlustes l_0	68
11.4.1	Verlust $l_0 = 0.0025$	69

11.4.2	Verlust $l_0 = 0.01$.	70
11.5	Zusammenfassung des Bagging bei Versicherungsdaten.	70
11.6	Boosting bei Versicherungsdaten.	71
11.7	L2-Boosting bei Versicherungsdaten.	72
A	S-Plus Programme	75
A.1	Bagging Regression trees.	75
A.1.1	Friedman# 1,2,3.	75
A.1.2	Ozon-Daten Simulation.	77
A.2	Bagging Classification trees.	78
A.2.1	Glas-Daten.	78
A.2.2	Brustkrebs-Daten.	79
A.3	Bagging MARS.	80
A.3.1	Friedman# 1,2,3 Simulation.	80
A.3.2	Ozone-Daten Simulation.	81
A.4	Bagging bei linearer Regression mit Modelselection.	82
A.4.1	Boston-housing Daten.	82
A.4.2	Ozon-Daten.	83
A.4.3	Simulierte Daten.	84
A.5	Bagging stumps.	86
B	Eigene S-Plus Funktionen	89
B.1	Friedman#1,2,3 Funktionen.	89
B.2	Eigene Funktionen.	90
C	R Programme	91
C.1	Boosting Sinus-Daten.	91
C.2	Bag-Boosting Sinus-Daten.	92
C.3	Boosting Krebsdaten.	93
C.4	Bag-Boosting Krebsdaten.	95
	Literaturverzeichnis	97