

1	Einleitung: Vom Batch Machine Learning zum Online Machine Learning	1
	Thomas Bartz-Beielstein	
1.1	Datenströme	2
1.2	Nachteile des Batch-Lernens	3
1.2.1	Speicherbedarf	4
1.2.2	Drift	4
1.2.3	Neue, unbekannte Daten	8
1.2.4	Zugänglichkeit und Verfügbarkeit der Daten	9
1.2.5	Weitere Probleme	9
1.3	Inkrementelles Lernen, Online-Lernen und Stream-Lernen	10
1.4	Überführung des Batch Machine Learning in das Online Machine Learning	12
2	Supervised Learning: Klassifikation und Regression	13
	Thomas Bartz-Beielstein	
2.1	Klassifikation	14
2.1.1	Baselinealgorithmen	14
2.1.2	Der naive Bayes-Klassifikator	14
2.1.3	Baumbasierte Verfahren	16
2.1.4	Weitere Klassifikationsverfahren	19
2.2	Regression	20
2.2.1	Online Linear Regression	20
2.2.2	Hoeffding Tree Regressor	20
2.3	Ensemble-Methoden für Online Machine Learning	21
2.4	Clustering	21
2.5	Übersicht: Online Machine Learning-Verfahren	22
3	Drifterkennung und -behandlung	23
	Thomas Bartz-Beielstein	
3.1	Architekturen für Driftbehandlungsmethoden	24
		IX

3.1.1	Adaptive Schätzer	24
3.1.2	Change Detectors	25
3.1.3	Ensemblebasierte Ansätze	25
3.2	Grundlegende Überlegungen zu Fenstertechniken	26
3.3	Populäre Verfahren zur Drifterkennung	27
3.3.1	Statistische Tests zur Drifterkennung und Change Detection	27
3.3.2	Kontrollkarten (Drift Detection Method)	27
3.3.3	Adaptive Windowing	28
3.3.4	Implizite Drifterkennungsalgorithmen	30
3.4	Online Machine Learning Algorithmen mit Drifterkennung: Hoeffding Window Trees	31
3.4.1	Concept-adapting Very Fast Decision Trees	31
3.4.2	Hoeffding Adaptive Trees	33
3.4.3	Übersicht: Hoeffding Window Trees	33
3.4.4	Übersicht: HT in river	33
4	Initiale Auswahl und nachträgliche Aktualisierung von OML-Modellen	37
	Thomas Bartz-Beielstein	
4.1	Initiale Modellauswahl	38
4.2	Modelländerungen	39
4.2.1	Hinzufügen neuer Merkmale	39
4.2.2	Manuelle Modelländerungen als Reaktion auf Drift	39
4.2.3	Sicherstellung der Modellgüte nach einem Modellupdate	40
4.3	Katastrophales Vergessen	40
4.3.1	Definition: Katastrophales Vergessen	41
4.3.2	Methoden gegen das katastrophale Vergessen	41
5	Evaluation und Performanzmessung	43
	Thomas Bartz-Beielstein	
5.1	Auswahlmethode	44
5.1.1	Holdout	44
5.1.2	Progressive Validierung: Interleaved test-then-train	45
5.1.3	Maschinelles Lernen im Batch-Verfahren mit einem Vorhersagehorizont	46
5.1.4	Landmark Batch Machine Learning mit einem Vorhersagehorizont	47
5.1.5	Window Batch Machine Learning mit einem Vorhersagehorizont	47
5.1.6	Online Machine Learning mit einem Vorhersagehorizont	48
5.1.7	Online-Maschinelles Lernen	48

5.2	Bestimmung des Training- und Testdatensatzes im Paket spotRiver	50
5.2.1	Methoden für Batch Machine Learning und Online Machine Learning	50
5.2.2	Methoden für Online Machine Learning (River)	52
5.3	Performanz des Algorithmus/Modells	56
5.4	Datenstrom- und Driftgeneratoren	58
5.4.1	Data-Stream-Generatoren in den scikit-Paketen	58
5.4.2	SEA-Drift-Generator	58
5.4.3	Friedman-Drift-Generator	59
5.5	Zusammenfassung	59
6	Besondere Anforderungen an OML-Verfahren	61
	Thomas Bartz-Beielstein	
6.1	Fehlende Daten, Imputation	62
6.2	Kategorische Attribute	63
6.3	Ausreißer	63
6.3.1	Weitere Anomalieerkennungsverfahren für Zeitreihen	64
6.3.2	One-Class Support Vector Machine zur Anomalieerkennung	64
6.3.3	Verfügbare Algorithmen zur Anomalieerkennung in river	64
6.4	Unbalancierte (unausgewogene) Daten	65
6.5	Große Anzahl an Features (Attributen)	65
6.6	FAIR, Interpretierbarkeit und Erklärbarkeit	66
7	Praxisanwendungen	69
	Steffen Moritz, Florian Dumpert, Thomas Bartz-Beielstein und Eva Bartz	
7.1	Anwendungen und Anwendungsperspektiven in der amtlichen Statistik	70
7.1.1	Potenziale und Herausforderungen	71
7.1.2	Vereinbarkeit mit Qualitätskriterien	76
7.1.3	Einbettung in den Statistikproduktionsprozess	78
7.1.4	(Online-)Machine-Learning-Anwendungen in Statistikinstitutionen	79
7.2	Andere Anwendungen mit Bezug zur amtlichen Statistik	81
7.2.1	Immobilienpreise	82
7.2.2	Pandemie-Vorhersagen	82
7.2.3	Stimmungsvorhersagen für Wahlen	83
7.2.4	Nowcasting für Wirtschaftsindizes	84
7.3	Aspekte bezüglich des Praxiseinsatzes	85

7.3.1	Unterschiede im Deployment-Prozess bei Batch Machine Learning- und Online Machine Learning-Ansätzen	85
7.3.2	Personalaufwand	86
8	Open-Source-Software für Online Machine Learning	89
	Thomas Bartz-Beielstein	
8.1	Übersicht und Beschreibung der Softwarepakete	90
8.1.1	Massive Online Analysis	90
8.1.2	Massive Online Analysis in R	90
8.1.3	stream	91
8.1.4	river	91
8.2	Softwareumfang	93
8.3	Vergleich der Programmiersprachen	95
9	Ein experimenteller Vergleich von Batch- und Online-Machine-Learning-Algorithmen	97
	Thomas Bartz-Beielstein	
9.1	Studie: Bike Sharing	98
9.1.1	Modellübersicht	101
9.1.2	Lineare Regression	101
9.1.3	Gradient Boosting	105
9.1.4	Hoeffding-Regressionsbäume	108
9.1.5	Abschließender Vergleich der Bike-Sharing-Experimente	110
9.1.6	Zusammenfassung: Bike-Sharing-Experimente	110
9.2	Studie: Sehr große Datensätze mit Drift	112
9.2.1	Der Friedman-Drift-Datensatz	112
9.2.2	Algorithmen	113
9.2.3	Ergebnisse	113
9.3	Zusammenfassung	114
10	Hyperparameter-Tuning	117
	Thomas Bartz-Beielstein	
10.1	Hyperparameter-Tuning: Eine Einführung	118
10.2	Die Hyperparameter-Tuning-Software Sequential Parameter Optimization Toolbox	119
10.3	Studie: Hyperparameter-Tuning des Hoeffding Adaptive Tree Regressor-Algorithmus auf den Friedman-Drift-Daten	120
10.3.1	Laden der Daten	120
10.3.2	Spezifikation des Vorverarbeitungsmodells	121
10.3.3	Auswahl des zu tunenden Algorithmus und der Defaulthyperparameter	122
10.3.4	Modifikation der Defaultwerte für die Hyperparameter	122
10.3.5	Auswahl der Zielfunktion (Loss-Funktion)	124

10.3.6 Aufruf des Hyperparameter Tuners Sequential Parameter Optimization Toolbox	125
10.3.7 Ergebnisse des Hoeffding Adaptive Tree Regressor-Tunings	125
10.3.8 Erklärbarkeit und Verständnis	128
10.4 Zusammenfassung	132
11 Zusammenfassung und Ausblick	133
Thomas Bartz-Beielstein und Eva Bartz	
11.1 Notwendigkeit für OML-Verfahren	133
11.2 Empfehlungen für die Online Machine Learning-Praxis	134
A Definitionen und Erläuterungen	137
A.1 Gradientenabstieg	137
A.2 Satz von Bayes	137
A.3 Hoeffding-Schranke	138
A.4 Kappa-Statistiken	138
Zusatzmaterial	141
B.1 Notebooks	141
B.2 Software	142
Glossary	143
Literatur	145
Stichwortverzeichnis	151