
Index

- Accuracy-Weighted Ensembles, **129**, 209
- AccuracyUpdatedEnsemble, 130, 209
- AccuracyWeightedEnsemble, 130, 209
- active learning, 13, **117**
 - Fixed Uncertainty Strategy, 119
 - in MOA, 211
 - Random Strategy, 119
 - Uncertainty Strategy with Randomization, 121
 - Variable Uncertainty Strategy, 119
- ActiveClassifier, 211
- Adaboost, 135
- AdaGraphMiner algorithm, **179**, 189
- AdaHoeffdingOptionTree, 209
- ADAMS project, 190
- adaptive bagging, *see* ADWIN Bagging
- Adaptive Random Forests, **137**
- Adaptive-Size Hoeffding Trees, **138**, 209
- AddNoiseFilter, 206
- ADWIN Bagging, 17, **133**, 200, 209
- ADWIN sketch, **79**, 82, 108, 179, 211
- AgrawalGenerator, 206
- Agresti-Coull bound, 39
- AMRules, **147**, 200
- AMS (Alon-Matias-Szegedy) sketch, 57
- Android operating system, 190
- Apex, 196
- approximation, 36
 - absolute, 36
 - (ϵ, δ) -approximation, **36**, 37, 62, 64
 - relative, 36
- Apriori algorithm, 19, **168**
- Area under the curve (AUC), 90
- ARFF files, 22, 203
- ArffFileStream, 204
- ARL, Average Run Length, 75
- attributes, 85
- AUC, 90

- bagging, 17, **133**
- Bayes' theorem, 95
- Bernstein's inequality, 39
- bias (in classifiers), 94
- BICO algorithm, **154**
- Big Data, 3
 - challenges, 6
 - hidden, 7
 - Three V's, 3
 - visualization, 7, 212
- BIRCH algorithms, **152**
- Bloom filter, 43
- boosting, **135**
- bootstrap, 133

- C++ language, 195

- C4.5, 101, 117
- CART, 101
- centers (clustering), 149
- centroids (clustering), 149
- CF trees, 153
- change in data streams, *see* drift
- CHARM algorithm, 170, 178
- Chebyshev's inequality, **38**, 46, 62, 92
- Chernoff's bound, **38**, 92
- classification, 11, **85**
 - comparing classifiers, 92
 - concept evolution, 121
 - CVFDT, 105
 - decision stump, 208
 - decision trees, 99, 208
 - delayed feedback, 13
 - ensembles, 71, 82, *see also* ensembles
 - evaluation, 86
 - Hoeffding Adaptive Tree, 108
 - Hoeffding Tree, 102
 - in MOA, 190, 201, 208–210
 - k -NN, 114, 190
 - lazy learning, *see* k -NN (nearest neighbors)
 - Majority Class classifier, 94
 - missing feedback, 13
 - multi-label, 115
 - Multinomial Naive Bayes, 98
 - Naive Bayes, 95
 - No-change classifier, 94
 - perceptron, 113
 - UFFT, 107
 - VFDT, 104
 - VFDTc, 107
- closed pattern, 169
- CloseGraph algorithm, 170, 179, 182
- cluster mapping measure (CMM), 151
- clustering, 11, 17, **149**
 - BICO, 154
 - BIRCH, 152
 - centroids or centers, 149
 - CluStream, 154
 - ClusTree, 156
 - CobWeb, 212
 - cost functions, 149
 - DBSCAN, 155
 - Den-Stream, 155
 - density-based, 155
 - distance function, 149
 - distributed, 200
 - evaluation, 150
 - in MOA, 160, 211
 - k -means, 18, 151
 - k -means++, 152
 - microclusters, 152
 - other methods, 159

- similarity, 149
 - StreamKM++, 158, 212
 - surveys, 159
- CluStream algorithm, **154**, 212, 213
- ClusTree algorithm, **156**, 212
- CM-sketch, *see* Count-Min sketch
- CMM (cluster mapping measure), 151
- CobWeb algorithm, 212
- Cohen's counter, **44**, 60
- cohesion measure (clustering), 150
- communities, 18
- comparing classifiers, 92
- concentration inequalities, **37**, 101
- concept drift, *see* drift
- concept evolution, 121
- ConceptDriftRealStream, 205
- ConceptDriftStream, 204
- confidence intervals, 37, 92
- confusion matrix, 91
- coresets
 - coreset tree, 158
 - in clustering, 158
 - in pattern mining, 172, 178, 182
- cost measures, 93
- Count-Min sketch, **51**, 60, 81, 82
- counting
 - distinct or unique items, 40, **42**, 48
 - items, **41**
- CountSketch, **54**
- cross-validation, 87, 204
 - distributed, 88
- CUSUM test, **75**, 82, 211
- CVFDT, **105**, 110
- data streams, **35**
 - adversarial vs. stochastic, 35, 69
 - change, *see* drift
 - definition, 8, **11**
 - distributed, 61, 88, 197
 - frequency moments, 56
 - in computer security, 9, 121
 - in disaster management, 9
 - in e-commerce, 9
 - in healthcare, 9
 - in marketing, 9
 - in social media, 9, 189, 190
 - in utilities, 9
 - items, **36**
 - Markovian, 69
 - scenarios, 8, 85, 121, 143
- dataset shift, 68
- DBSCAN algorithm, **155**
- DDM, Drift Detection Method, **78**, 82, 83, 107, 211
- decay factor, 73
- decision rules, **146**, 200
- Decision Stump classifier, 208
- decision trees, 16, **99**, 208
 - split criteria, 101
- delayed feedback, 13
- δ , confidence parameter, 37
- Δ -support, 178, 183
- Den-Stream algorithm, **155**, 212
- density-based clustering, **155**
- discretization, **109**, 190
- distinct items, *see* counting
- distributed evaluation, 88
- drift, **67**
 - gradual, 69
 - in MOA, 190, 210
 - recurrent concepts, 69, 139
 - shift, 69
 - simulating in MOA, 22, 25, 204
 - strategies to manage, 70
 - types of, 69
- Eclat algorithm, 19, 169
- ensembles, 17, 71, 82, **129**
 - Accuracy-Weighted, 129
 - Adaboost, 135
 - Adaptive Random Forests, 137
 - Adaptive Size Hoeffding Tree, 138
 - ADWIN Bagging, 17, **133**
 - bagging, 17, **133**
 - boosting, 135
 - exponentiated gradient, 132
 - Hoeffding Option Tree, 136
 - in MOA, 209
 - Leveraging Bagging, 134
 - Online Bagging, 133
 - Online Boosting, **135**
 - random forests, 136
 - stacking, 132, 137
 - Weighted Majority, 130
- entropy, 101, 117
- ϵ , accuracy parameter, 36
- Equal-frequency discretization, 109
- Equal-width discretization, 109
- error-correcting output codes, 134
- estimators, **72**
- evaluation, 14, **86**
 - AUC, 90
 - cross-validation, *see* cross-validation
 - distributed, *see* distributed evaluation
 - holdout, *see* holdout evaluation
 - in clustering, **150**
 - in MOA, 22–31, **203**
 - interleaved chunks, *see* interleaved chunks evaluation

- prequential, *see* prequential evaluation
 - statistical significance, 92
 - test-then-train, *see* test-then-train evaluation
- EWMA estimator, **73**, 82, 151, 211
- exhaustive binary tree, **110**, 146
- Exponential Histograms, **57**, 61, 64, 73, 80
- exponentiated gradient algorithm, 132

- Facebook graph, 48
- fading factor, 73
- Fayyad and Irani's discretization, **109**
- feature extraction, 10
- features, *see* attributes
- FilteredStream, 205
- FIMT-DD, **146**
- Flajolet-Martin counter, **45**, 60
- Flink, 6, 196
- FP-Growth algorithm, 19, 168, 175
- FP-Stream algorithm, 175
- FP-Tree, 168
- frequency moments (in streams), 56
- frequency problems, **48**
- frequent elements, *see* heavy hitters
- frequent pattern, *see* pattern mining
- Frequent sketch, 49
- FrugalStreaming sketch, 54

- Gaussian distribution, 38, 111
- Gini impurity index, **101**
- gnuplot, 219
- GPU computing, 137
- graph mining, 10, 178
- graphical models, 94
- GraphX, 6

- Hadoop, 5, 196
- hash functions, 43, 44, **61**
 - families of random, 61
 - fully independent, 61
 - in practice, 62
 - pairwise independent, 61
- HDFS, 5
- heavy hitters, **49**, 64
 - by sampling, 49
 - in itemset mining, 174
 - in pattern mining, 174
 - surveys, 49
- Hoeffding Adaptive Tree classifier, 17, **108**, 209
- Hoeffding adaptive tree classifier, 195
- Hoeffding Option Tree classifier, **136**, 146, 209
- Hoeffding Tree classifier, 16, **102**, 190, 208
 - multi-label, 117
 - vertical, 200
- Hoeffding's bound, **38**, 46, 63, 65, 81, 82, 92, 101, 102, 172, 177
- holdout evaluation, 14, **87**, 204
- Huawei, 195
- HyperANF counter, 47
- HyperLogLog counter, **46**, 47
- HyperplaneGenerator, 206
- hypothesis testing, *see* statistical tests

- IBLStreams, **145**, 189
- iceberg queries, 49
- IID assumption, 69, **86**, 91
- IncMine algorithm, 19, 176, 183, 189
- information gain, **101**, 101, 117
- interleaved chunks evaluation, **88**, 204
- Internet of Things, 3, 8
- items, **36**
- itemset, **165**

- Java language, 187, 188, 195, 196, 221, 227
 - good practices, 238

- Kalman filter estimator, 74
- Kappa architecture, 6
- Kappa M statistic, **90**
- Kappa statistic, **90**
- Kappa temporal statistic, **91**
- kernel methods, 94, 148
- k -grams, counting, 42
- k -means algorithm, 18, **151**
- k -means++ algorithm, **152**
- k -NN (nearest neighbors), 15, 190
 - for classification, **114**, 122
 - for regression, **145**

- Lambda architecture, 6
- Laplace correction, **97**, 99
- large-deviation bounds, *see* concentration inequalities
- lazy learning, *see* k -NN (nearest neighbors)
- learning rate, 114
- LEDGenerator, 206
- LEDGeneratorDrift, 207
- Leveraging Bagging, **134**, 210
- LimAttClassifier, 138, 210
- Linear counting, **43**, 60
- linear estimator, 73
- linear regression, **143**
- Lossy Counting sketch, 49, 174

- Mahout, 6
- Majority Class classifier, 15, **94**, 210
- Markov's inequality, **38**, 53, 92
- maximal pattern, 169

- McDiarmid's inequality, 39, 101
- McNemar's test, **93**
- MDL, Minimum Description Length, 109
- MDR, Missed Detection Rate, 75
- MEKA project, 193
- Mergeability, **60**
- Merging sketches, **60**
- microclusters, 18, **152**, 154, 200
- Milgram's degrees of separation, 48
- Misra-Gries counter, 49
- missing data, 10
- missing feedback, 13
- MLIB, 6
- MOA, 10, 21, 187
 - adding classes to, 227
 - API, **221**
 - classification, 201, 218
 - clustering, 160
 - Command Line Interface (CLI), 29, **217**
 - compiling code for, 237
 - discretization, 190
 - distributed, *see* SAMOA
 - evaluation, 22–31, **203**, 218
 - extensions, 189
 - for Android, 190
 - for social media analysis, 189, 190, 192
 - for video processing, 193
 - generators, 160, **204**, 204, 212
 - good programming practices, 237
 - GUI, 22, 23, **201**
 - Hadoop, 196
 - installing, 21, 188
 - modifying the behavior of, 227
 - multi-target learning, 188
 - outlier detection, 188
 - platforms, 187, 188, 190
 - programming applications that use, 221
 - recent developments, 188
 - recommender systems, 189
 - regression, 148, 218
 - running tasks, 22, 123, 201, 217
 - SAMOA, 196
 - Spark, 195
 - tasks, 188, 203, 217
 - visualization, 212
- MOA-TweetReader, 189
- MOAReduction, 190
- Moment algorithm, 19, 174, 189
- moment computation, 56
- Morris's counter, **41**, 61, 63
- motif discovery, 10
- MTD, Mean Time to Detection, 75
- MTFA, Mean Time between False Alarms, 75
- multi-label classification, **115**
 - BR method, 115
 - in MOA, 193
 - LC method, 115
 - multi-label Hoeffding Tree, 116
 - PW method, 116
- multi-target learning, 188
- Multinomial Naive Bayes classifier, **98**, 208
- Naive Bayes
 - Multinomial, *see* Multinomial Naive Bayes classifier
- Naive Bayes classifier, 16, **95**, 105, 208
- neighborhood function (in graphs), 47
- No-change classifier, 15, **94**, 210
- normal approximation, **38**, 92, 172
- normal distribution, 38, 111
- numeric attributes, **109**, 143
 - in MOA, 190
- OCBoost, 209
- Online Bagging, 133
- Online Bagging algorithm, 209
- Online Boosting algorithm, 209
- Online Boosting algorithm, **135**, 209
- OpenML project, 194
- outliers, 70, 81, 109, 113, 188
- overfitting (in classifiers), 94
- OzaBag, 133, 209
- OzaBagADWIN, 133, 209
- OzaBagASHT, 138, 209
- OzaBoost, 135, 209
- PAC-learning, 37
- Page-Hinkley test, **76**, 82, 83, 146, 211
- pattern mining, 11, 18, 165, **167**
 - AdaGraphMiner, 179
 - Apriori, 168
 - association rules, 182
 - candidate pattern, 168
 - CHARM, 170, 178
 - closed pattern, 169, 182, 183
 - CloseGraph, 170, 179, 182
 - coresets, 172, 178, 182
 - Eclat, 169
 - FP-Growth, 168
 - FP-Stream, 175
 - generic algorithm on streams, **170**
 - graph, **166**, 178, 182
 - in MOA, 178, 182, 189
 - IncMine, 176, 183
 - itemset, 18, 165, 181
 - maximal pattern, 169
 - Moment, 174
 - other algorithms, 170, 181
 - pattern, 165, **166**

- pattern size, 167
- sequence, **165**, 181
- SPMF, 178, 182
- subpattern, 166
- superpattern, 166
- support, 166
- surveys, 181
- tree, **166**, 181
- WinGraphMiner, 179
- Perceptron, 132, 146, 210
 - for regression, **145**
 - stacking on Hoeffding Trees, 137, 210
- perceptron
 - for classification, **113**
- Poisson distribution, 133, 134
- prequential evaluation, 14, **88**, 90, 204
- Probabilistic counter, *see* Flajolet-Martin counter
- purity measure (clustering), 150
- Python language, 195
- quantiles, 54
 - FrugalStreaming sketch, 54
 - Greenwald and Khanna's sketch, 111, 190
 - in MOA, 190
- R language, 191, 195
- RAM-hour, **94**
- random forests, **136**
- randomized algorithm, 36
- RandomRBFGenerator, 207
- RandomRBFGeneratorDrift, 207
- RandomSEAGenerator, 207
- RandomTreeGenerator, 207
- range-sum queries, **53**, 64
- ranking / learning to rank, 10
- real-time analytics, *see* data streams
- recommender systems, 10, 189
- recurrent concepts, 10, 69, **139**
- regression, **143**
 - AMRules, 147
 - error measures, 144
 - FIMT-DD, 146
 - IBLStreams, 145
 - in MOA, 148, 189, 210
 - k -NN, 145
 - linear regression, 143
 - Perceptron, 145
 - Spegasos, 148
 - stochastic gradient descent, 148
- reservoir sampling, **40**
- rule learners, 94, 146
- SAMOA, **196**
- sampling, **39**, 63
 - for heavy hitters, 49
 - reservoir, *see* reservoir sampling
- Samza, 196
- semi-supervised learning, 13
- SGD, 210
- sigmoid, 25, 204
- silhouette coefficient, 150
- six degrees of separation, 48
- sketches, 35, **36**
 - ADWIN, **79**, 82, 108, 179
 - AMS (Alon-Matias-Szegedy), 57
 - Cohen's counter, 44
 - Count-Min, 51
 - CountSketch, 54
 - Exponential Histograms, 57, 73
 - Flajolet-Martin counter, 45
 - for linear algebra, 63
 - for massive graphs, 48
 - Frequent, 49
 - FrugalStreaming, 54
 - HyperLogLog counter, 46
 - Linear counting, 43
 - Lossy Counting, 49, 174
 - merging, 60
 - Misra-Gries, 49
 - Morris's counter, 41
 - other sketches, 63
 - quantiles, 54, 111
 - range-sum queries, 53
 - reservoir sampling, 40
 - Space Saving, **50**, 64, 82, 174, 183
 - Sticky Sampling, 49
 - Stream-Summary, 51
- skip counting, 41
- sliding windows, **58**, 73, 79, 83, 178
- Space Saving sketch, **50**, 61, 64, 82, 174, 183
- spam, 11, 85, 100
- Spark, 6, 195
- Spark Streaming, 6, 195
- Spegasos, 148, 210
- split criteria, 101
- split-validation, **89**
- SPMF framework, 178, 182
- SSQ measure (clustering), 150
- stacking, 132
 - Perceptron on Hoeffding Trees, 137, 210
- STAGGERGenerator, 208
- statistical significance, 92
 - McNemar's test, 92
- statistical tests, 76, 81
- Sticky Sampling sketch, 49
- stochastic averaging, 46
- stochastic gradient descent, 114, 148, 210
- Storm, 196

- stream cross-validation, **90**
- Stream-Summary structure, **51**
- StreamDM-C++ project, 195
- streaming, *see* data streams
- StreamKM++ algorithm, **158**, 212
- Streams project, 196
- subpattern, *see* pattern mining
- summaries, *see* sketches
- superpattern, *see* pattern mining
- supervised learning, 11, 85
- support (of a pattern), 166
- support vector machines (SVM), *see* kernel methods
- TemporallyAugmentedClassifier, 95, 210
- TensorFlow, 6
- test-then-train evaluation, 14, **87**, 204
- time series, 68
- Twitter, 15, 85, 96, 99, 121, 189, 192
- UFFFT, **107**, 112
- unique items, *see* counting
- unsupervised learning, 11, 149, 165
- Vertical Hoeffding Tree, 200
- VFDT, **104**, 110
- VFDTc, **107**, 110
- VFML, **110**
- video processing, 193
- WaveformGenerator, 208
- WaveformGeneratorDrift, 208
- Weighted Majority algorithm, 130
- WEKA, 10, 22, 190, 193, 203
- WinGraphMiner algorithm, 179

Adaptive Computation and Machine Learning

Francis Bach, Editor

Christopher Bishop, David Heckerman, Michael Jordan, and Michael Kearns,
Associate Editors

Bioinformatics: The Machine Learning Approach, Pierre Baldi and Søren Brunak

Reinforcement Learning: An Introduction, Richard S. Sutton and Andrew G.
Barto

Graphical Models for Machine Learning and Digital Communication, Brendan J.
Frey

Learning in Graphical Models, Michael I. Jordan

Causation, Prediction, and Search, second edition, Peter Spirtes, Clark Glymour,
and Richard Scheines

Principles of Data Mining, David Hand, Heikki Mannila, and Padhraic Smyth

Bioinformatics: The Machine Learning Approach, second edition, Pierre Baldi
and Søren Brunak

Learning Kernel Classifiers: Theory and Algorithms, Ralf Herbrich

*Learning with Kernels: Support Vector Machines, Regularization, Optimization,
and Beyond*, Bernhard Schölkopf and Alexander J. Smola

Introduction to Machine Learning, Ethem Alpaydin

Gaussian Processes for Machine Learning, Carl Edward Rasmussen and
Christopher K.I. Williams

Semi-Supervised Learning, Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, Eds.

The Minimum Description Length Principle, Peter D. Grünwald

Introduction to Statistical Relational Learning, Lise Getoor and Ben Taskar, Eds.

Probabilistic Graphical Models: Principles and Techniques, Daphne Koller and Nir Friedman

Introduction to Machine Learning, second edition, Ethem Alpaydin

Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation, Masashi Sugiyama and Motoaki Kawanabe

Boosting: Foundations and Algorithms, Robert E. Schapire and Yoav Freund

Machine Learning: A Probabilistic Perspective, Kevin P. Murphy

Foundations of Machine Learning, Mehryar Mohri, Afshin Rostami, and Ameet Talwalker

Introduction to Machine Learning, third edition, Ethem Alpaydin

Deep Learning, Ian Goodfellow, Yoshua Bengio, and Aaron Courville

Toward Causal Learning, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf

Machine Learning for Data Streams with Practical Examples in MOA, Albert Bifet, Ricard Gavaldà, Geoff Holmes, and Bernhard Pfahringer

This is a section of [doi:10.7551/mitpress/10654.001.0001](https://doi.org/10.7551/mitpress/10654.001.0001)

Machine Learning for Data Streams

with Practical Examples in MOA

By: Albert Bifet, Ricard Gavaldà, Geoffrey Holmes, Bernhard Pfahringer

Citation:

Machine Learning for Data Streams: with Practical Examples in MOA

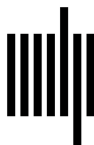
By: Albert Bifet, Ricard Gavaldà, Geoffrey Holmes, Bernhard Pfahringer

DOI: 10.7551/mitpress/10654.001.0001

ISBN (electronic): 9780262346047

Publisher: The MIT Press

Published: 2023



The MIT Press

© 2017 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form or by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

This book was set in Times Roman and Mathtime Pro 2 by the authors.

Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data is available

ISBN: 978-0-262-03779-2

10 9 8 7 6 5 4 3 2 1