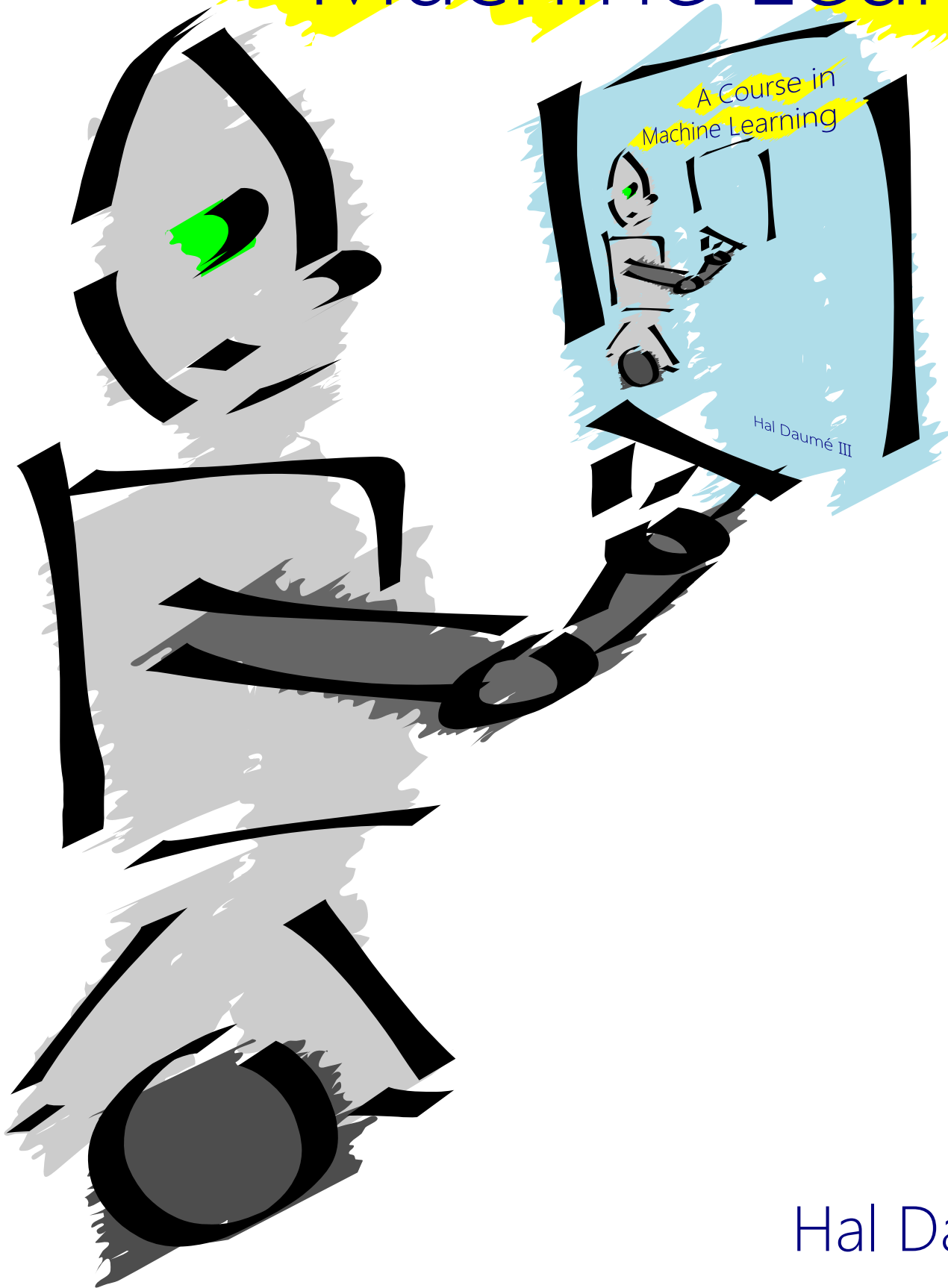


A Course in Machine Learning



Hal Daumé III

CODE AND DATASETS

Rating	Easy?	AI?	Sys?	Thy?	Morning?
+2	y	y	n	y	n
+2	y	y	n	y	n
+2	n	y	n	n	n
+2	n	n	n	y	n
+2	n	y	y	n	y
+1	y	y	n	n	n
+1	y	y	n	y	n
+1	n	y	n	y	n
0	n	n	n	n	y
0	y	n	n	y	y
0	n	y	n	y	n
0	y	y	y	y	y
-1	y	y	y	n	y
-1	n	n	y	y	n
-1	n	n	y	n	y
-1	y	n	y	n	y
-2	n	n	y	y	n
-2	n	y	y	n	y
-2	y	n	y	n	n
-2	y	n	y	n	y

NOTATION

BIBLIOGRAPHY

Sergey Brin. Near neighbor search in large metric spaces. In *Conference on Very Large Databases (VLDB)*, 1995.

Tom M. Mitchell. *Machine Learning*. McGraw Hill, 1997.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958. Reprinted in *Neurocomputing* (MIT Press, 1998).

- K-nearest neighbors, 56
- ϵ -ball, 37
- p -norms, 91
- 0/1 loss, 87

- absolute loss, 14
- activation function, 117
- activations, 39
- active learning, 181
- AdaBoost, 154
- algorithm, 86
- all pairs, 76
- all versus all, 76
- architecture selection, 126
- area under the curve, 62, 81
- AUC, 62, 80, 81
- AVA, 76
- averaged perceptron, 49

- back-propagation, 121, 124
- bag of words, 54
- bagging, 153
- base learner, 152
- batch, 161
- batch learning, 184
- Bayes error rate, 104, 150
- Bayes optimal classifier, 103, 150
- Bayes optimal error rate, 104
- Bernoulli distribution, 108
- bias, 40
- binary features, 27
- bipartite ranking problems, 79
- boosting, 142, 152
- bootstrap resampling, 153
- bootstrapping, 65, 67

- categorical features, 27
- chain rule, 107
- chord, 89

- circuit complexity, 125
- clustering, 32, 166
- clustering quality, 166
- collective classification, 83
- complexity, 31
- concave, 89
- concavity, 179
- concept, 144
- confidence intervals, 66
- constrained optimization problem, 98
- contour, 91
- convergence rate, 94
- convex, 86, 88
- cross validation, 62, 66
- cubic feature map, 131
- curvature, 94

- data covariance matrix, 172
- data generating distribution, 15
- decision boundary, 31
- decision stump, 156
- decision tree, 8, 10
- decision trees, 55
- development data, 24
- dimensionality reduction, 166
- discrete distribution, 109
- distance, 28
- dominates, 61
- dot product, 43
- dual problem, 138
- dual variables, 138

- early stopping, 51, 120
- embedding, 166
- ensemble, 152
- error driven, 41
- error rate, 87
- Euclidean distance, 28
- evidence, 114

- example normalization, 57, 58
- examples, 9
- expectation maximization, 175
- expected loss, 15, 16
- exponential loss, 89, 157

- feasible region, 99
- feature combinations, 51
- feature mapping, 51
- feature normalization, 57
- feature scale, 30
- feature space, 28
- feature values, 11, 26
- feature vector, 26, 28
- features, 11, 26
- forward-propagation, 124
- fractional assignments, 176
- furthest-first heuristic, 168

- Gaussian distribution, 109
- Gaussian kernel, 134
- Gaussian Mixture Models, 177
- generalize, 9, 16
- generative story, 110
- geometric view, 26
- global minimum, 92
- GMM, 177
- gradient, 92
- gradient ascent, 92
- gradient descent, 92
- graph, 83

- hard-margin SVM, 99
- hash kernel, 165
- held-out data, 24
- hidden units, 116
- hidden variables, 175
- hinge loss, 89
- histogram, 12

- hyperbolic tangent, 117
- hypercube, 35
- hyperparameter, 23, 42, 88
- hyperplane, 39
- hyperspheres, 35
- hypothesis, 144
- hypothesis class, 147
- hypothesis testing, 65

- i.i.d. assumption, 105
- imbalanced data, 70
- importance weight, 71
- independently, 104
- independently and identically distributed, 105
- indicator function, 87
- induce, 15
- induced distribution, 72
- induction, 9
- inductive bias, 18, 28, 30, 90, 109
- iteration, 32

- jack-knifing, 67
- Jensen's inequality, 179
- joint, 111

- K-nearest neighbors, 29
- Karush-Kuhn-Tucker conditions, 139
- kernel, 128, 132
- kernel trick, 133
- kernels, 52
- KKT conditions, 139

- label, 11
- Lagrange multipliers, 106
- Lagrange variable, 106
- Lagrangian, 106
- layer-wise, 126
- leave-one-out cross validation, 63
- level-set, 91
- license, 2
- likelihood, 114
- linear classifier, 157
- linear classifiers, 157
- linear decision boundary, 39
- linear regression, 96
- linearly separable, 45
- link function, 117
- log likelihood, 106
- log posterior, 114
- log probability, 106
- log-likelihood ratio, 109
- logarithmic transformation, 59
- logistic loss, 89
- logistic regression, 113
- LOO cross validation, 63
- loss function, 14

- margin, 46, 98
- margin of a data set, 46
- marginal likelihood, 114
- maximum a posteriori, 114
- maximum depth, 23
- maximum likelihood estimation, 105
- Mercer's condition, 133
- model, 86
- modeling, 22
- multi-layer network, 116

- naive Bayes assumption, 107
- nearest neighbor, 26, 28
- neural network, 157
- neural networks, 52, 116
- neurons, 39
- noise, 19
- non-convex, 122
- non-linear, 116
- Normal distribution, 109
- normalize, 44, 57
- null hypothesis, 65

- objective function, 87
- one versus all, 75
- one versus rest, 75
- online, 41
- online learning, 184
- optimization problem, 87
- output unit, 116
- OVA, 75
- overfitting, 21
- oversample, 73

- p-value, 65
- PAC, 143, 154
- paired t-test, 65
- parametric test, 65
- parity function, 125
- patch representation, 54
- PCA, 172
- perceptron, 39, 40, 56
- perpendicular, 43
- pixel representation, 54

- polynomial kernels, 133
- positive semi-definite, 133
- posterior, 114
- precision, 60
- precision/recall curves, 60
- predict, 9
- preference function, 78
- primal variables, 138
- principle components analysis, 172
- prior, 114
- probabilistic modeling, 103
- Probably Approximately Correct, 143
- projected gradient, 138
- psd, 133

- radial basis function, 126
- random forests, 157
- RBF kernel, 134
- RBF network, 126
- recall, 60
- receiver operating characteristic, 62
- reconstruction error, 172
- reductions, 72
- redundant features, 54
- regularized objective, 88
- regularizer, 87, 90
- representer theorem, 130, 132
- ROC curve, 62

- sample complexity, 144, 145, 147
- semi-supervised learning, 181
- sensitivity, 62
- separating hyperplane, 86
- SGD, 161
- shallow decision tree, 18, 156
- shape representation, 54
- sigmoid, 117
- sigmoid function, 113
- sigmoid network, 126
- sign, 117
- single-layer network, 116
- slack, 135
- slack parameters, 99
- smoothed analysis, 168
- soft assignments, 176
- soft-margin SVM, 99
- span, 130
- sparse, 91
- specificity, 62
- squared loss, 14, 89
- stacking, 84

- StackTest, 84
- statistical inference, 103
- statistically significant, 65
- stochastic gradient descent, 161
- stochastic optimization, 160
- strong learner, 154
- strong learning algorithm, 154
- strongly convex, 94
- structural risk minimization, 86
- sub-sampling, 72
- subderivative, 95
- subgradient, 95
- subgradient descent, 96
- support vector machine, 98
- support vectors, 140
- surrogate loss, 89
- symmetric modes, 122
- t-test, 65
- test data, 22
- test error, 22
- test set, 9
- text categorization, 54
- the curse of dimensionality, 34
- threshold, 40
- Tikhonov regularization, 86
- training data, 9, 15, 22
- training error, 16
- truncated gradients, 163
- two-layer network, 116
- unbiased, 45
- underfitting, 21
- unit hypercube, 36
- unsupervised learning, 32
- validation data, 24
- Vapnik-Chernovenkis dimension, 149
- variance, 153
- VC dimension, 149
- vector, 28
- visualize, 166
- vote, 29
- voted perceptron, 49
- voting, 49
- weak learner, 154
- weak learning algorithm, 154
- weighted nearest neighbors, 37
- weights, 39
- zero/one loss, 14