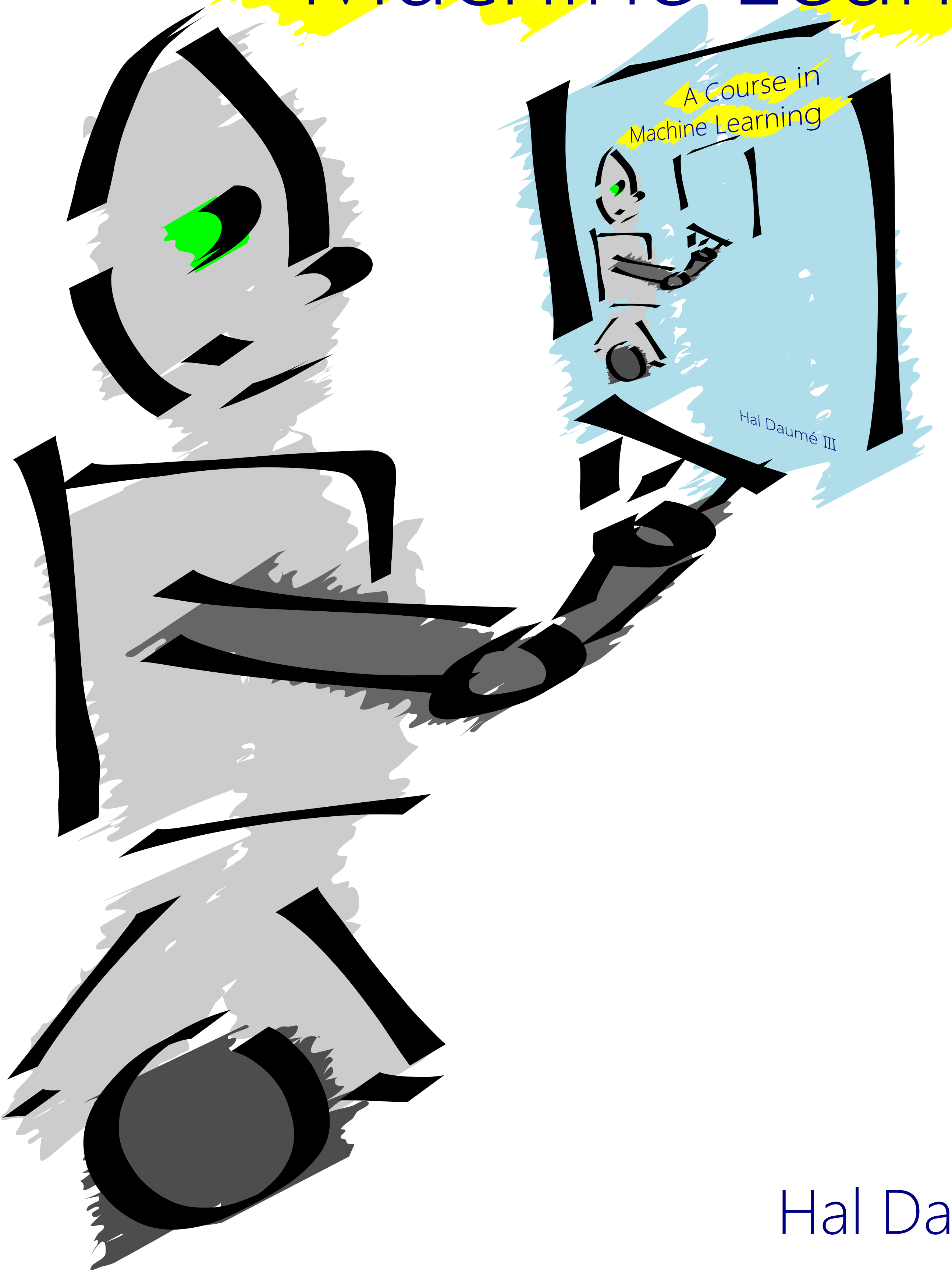
A Course in Machine Learning



Hal Daumé III

Draft: Do Not Distribute

Copyright © 2012 Hal Daumé III

http://ciml.info

This book is for the use of anyone anywhere at no cost and with almost no restrictions whatsoever. You may copy it or re-use it under the terms of the CIML License online at ciml.info/LICENSE. You may not redistribute it yourself, but are encouraged to provide a link to the CIML web page for others to download for free. You may not charge a fee for printed versions, though you can print it for your own use.

version 0.8 , August 2012

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

Draft: Do Not Distribute

Notation

Draft: Do Not Distribute

Bibliography

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).

Draft: Do Not Distribute

Index

K-nearest neighbors, 54
ϵ-ball, 35
p-norms, 89
0/1 loss, 85

absolute loss, 14
activation function, 114
activations, 37
active learning, 177
AdaBoost, 151
algorithm, 84
all pairs, 74
all versus all, 74
architecture selection, 123
area under the curve, 60, 79
AUC, 60, 77, 79
AVA, 74
averaged perceptron, 47

back-propagation, 118, 121
bag of words, 52
bagging, 150
base learner, 149
batch, 158
batch learning, 180
Bayes error rate, 102, 147
Bayes optimal classifier, 101, 147
Bayes optimal error rate, 102
Bernouilli distribution, 106
bias, 38
binary features, 25
bipartite ranking problems, 77
boosting, 139, 149
bootstrap resampling, 150
bootstrapping, 63, 65

categorical features, 25
chain rule, 105
chord, 87
circuit complexity, 122
clustering, 30, 163
clustering quality, 163
collective classification, 81
complexity, 29
concave, 86
concavity, 175
concept, 141
confidence intervals, 64
constrained optimization problem, 96
contour, 89
convergence rate, 92
convex, 84, 86
cross validation, 60, 64
cubic feature map, 128
curvature, 92

data covariance matrix, 169
data generating distribution, 15
decision boundary, 29
decision stump, 153
decision tree, 8, 10
decision trees, 53
development data, 22
dimensionality reduction, 163
discrete distribution, 106
distance, 26
dominates, 59
dot product, 41
dual problem, 135
dual variables, 135

early stopping, 49, 117
embedding, 163
ensemble, 149
error driven, 39
error rate, 85
Euclidean distance, 26
evidence, 112
example normalization, 55, 56
examples, 9
expectation maximization, 171
expected loss, 15
exponential loss, 87, 154

feasible region, 97
feature combinations, 49
feature mapping, 49
feature normalization, 55
feature scale, 28
feature space, 25
feature values, 11, 24
feature vector, 24, 26
features, 11, 24
forward-propagation, 121
fractional assignments, 172
furthest-first heuristic, 165

Gaussian distribution, 106
Gaussian kernel, 131
Gaussian Mixture Models, 173
generalize, 9, 16
generative story, 108
geometric view, 24
global minimum, 90
GMM, 173
gradient, 90
gradient ascent, 90
gradient descent, 90
graph, 81

hard-margin SVM, 97
hash kernel, 162
held-out data, 22
hidden units, 113
hidden variables, 171
hinge loss, 87
histogram, 12
hyperbolic tangent, 114
hypercube, 33
hyperparameter, 21, 40, 86
hyperplane, 37
hyperspheres, 33
hypothesis, 141

Draft: Do Not Distribute

hypothesis class, 144
hypothesis testing, 63

i.i.d. assumption, 103
imbalanced data, 68
importance weight, 69
independently, 102
independently and identically distributed, 103
indicator function, 85
induce, 15
induced distribution, 70
induction, 9
inductive bias, 17, 26, 28, 88, 106
iteration, 30

jack-knifing, 65
Jensen's inequality, 175
joint, 109

K-nearest neighbors, 27
Karush-Kuhn-Tucker conditions, 136
kernel, 125, 129
kernel trick, 130
kernels, 50
KKT conditions, 136

label, 11
Lagrange multipliers, 104
Lagrange variable, 104
Lagrangian, 104
layer-wise, 123
leave-one-out cross validation, 61
level-set, 89
license, 2
likelihood, 112
linear classifier, 154
linear classifiers, 154
linear decision boundary, 37
linear regression, 94
linearly separable, 43
link function, 114
log likelihood, 103
log posterior, 112
log probability, 103
log-likelihood ratio, 107
logarithmic transformation, 57
logistic loss, 87
logistic regression, 111
LOO cross validation, 61
loss function, 14

margin, 44, 96
margin of a data set, 44
marginal likelihood, 112
maximum a posteriori, 111
maximum depth, 21
maximum likelihood estimation, 103
Mercer's condition, 130
model, 84
modeling, 21
multi-layer network, 113

naive Bayes assumption, 105
nearest neighbor, 24, 26
neural network, 154
neural networks, 50, 113
neurons, 37
noise, 17
non-convex, 119
non-linear, 113
Normal distribution, 106
normalize, 42, 55
null hypothesis, 63

objective function, 85
one versus all, 72
one versus rest, 72
online, 38
online learning, 180
optimization problem, 85
output unit, 113
OVA, 72
overfitting, 20
oversample, 71

p-value, 63
PAC, 140, 151
paired t-test, 63
parametric test, 63
parity function, 122
patch representation, 52
PCA, 169
perceptron, 37, 38, 54
perpendicular, 41
pixel representation, 51
polynomial kernels, 130
positive semi-definite, 130
posterior, 112
precision, 58
precision/recall curves, 58
predict, 9
preference function, 76
primal variables, 135
principle components analysis, 169
prior, 112
probabilistic modeling, 101
Probably Approximately Correct, 140
projected gradient, 135
psd, 130

radial basis function, 123
random forests, 154
RBF kernel, 131
RBF network, 123
recall, 58
receiver operating characteristic, 60
reconstruction error, 169
reductions, 70
redundant features, 52
regularized objective, 85
regularizer, 85, 88
representer theorem, 127, 129
ROC curve, 60

sample complexity, 141–143
semi-supervised learning, 177
sensitivity, 60
separating hyperplane, 84
SGD, 158
shallow decision tree, 17, 153
shape representation, 52
sigmoid, 114
sigmoid function, 110
sigmoid network, 123
sign, 114
single-layer network, 113
slack, 132
slack parameters, 97
smoothed analysis, 165
soft assignments, 172
soft-margin SVM, 97
span, 127
sparse, 89
specificity, 60
squared loss, 14, 87
stacking, 82
StackTest, 82
statistical inference, 101
statistically significant, 63
stochastic gradient descent, 158
stochastic optimization, 157
strong learner, 151
strong learning algorithm, 151

Draft: Do Not Distribute

strongly convex, 92
structural risk minimization, 84
sub-sampling, 70
subderivative, 93
subgradient, 93
subgradient descent, 94
support vector machine, 96
support vectors, 137
surrogate loss, 87
symmetric modes, 119

t-test, 63
test data, 20
test error, 20
test set, 9
text categorization, 52
the curse of dimensionality, 32
threshold, 38
Tikhonov regularization, 84
training data, 9, 15, 20
training error, 16
trucated gradients, 160
two-layer network, 113

unbiased, 43
underfitting, 20
unit hypercube, 34
unsupervised learning, 30

validation data, 22
Vapnik-Chernovenkis dimension, 146
variance, 150
VC dimension, 146
vector, 25
visualize, 163
vote, 27
voted perceptron, 47
voting, 47

weak learner, 151
weak learning algorithm, 151
weighted nearest neighbors, 35
weights, 37

zero/one loss, 14

Draft: Do Not Distribute