Index

- active learning, 181
  - definition, 12
- Adaline, 35, 110
- approximation, 27
  - versus generalization, 62–68, 106
- artificial intelligence, 5
- augmented error, 132, 157
- axiom of non-falsifiability, 178

- $B(N,k)$
  - definition, 46
  - lower bound, 69
  - upper bound, 48
- backgammon, 12
- Bayes optimal decision theory, 10
- Bayes theorem, 33
- Bayesian learning, 181
- bias-variance, 62–66
  - average function, 63
  - dependence on N, d, 158
  - example, 65
  - impact of noise, 125
  - linear models, 158–159
  - linear regression, 114
  - noisy target, 74
- bin model, 18
  - multiple bins, 22
  - relationship to learning, 20
- binomial distribution, 36
- boosting, 181
- break point
  - definition, 45

- Chebyshev inequality, 36
- Chernoff bound, 37
- classification
  - for regression, 113
  - linear programming algorithm, 110
- classification error
  - bound by cross-entropy error, 97
  - bound by squared error, 97
- clustering, 13
- coin classification, 9, 13
- combinatorial optimization, 80
- complexity
  - of $\mathcal{H}$, 26
  - of f, 27
- computational complexity, 181
- computational finance, 181
- computer vision, 1
- convex function, 93
- convex set, 44
- cost, 28
- cost matrix, 29, 115
- credit approval, 3, 82, 96
- cross validation, 145–150
  - V-fold, 150
  - choosing λ, 149
  - digits data, 151
  - effective number of examples, 163
  - exact computation, 149
  - leave-one-out, 146
  - linear model, 149
  - linear model, analytic, 164
  - model selection, 148
  - regularized, 165
  - summary, 147
  - unbiased, 147
  - variance, 162
- cross-entropy, 92

- data contamination, 145, 151, 176
- data mining, 15
- data point, **3**
- data set, **3**
  - ghost, 188
  - space of, 54
- data snooping, 173–177, 181

financial trading, 174
nonlinear transform, 103
normalization bias, 174
versus sampling bias, 177
decision stump, 106
design
versus learning, 9
deterministic noise, **124**, 128
effect on learning, 151
regularization, 136
similarity to stochastic noise, 136
Dewey, 171
dichotomy, **42**
maximum number, 46
perceptron, 43
table, 47
differentiable, 85
twice-, 93, 95

effective number of hypotheses, **41**, 53
effective number of parameters, 52, **137**, 159
Einstein, 167
ensemble learning, 181
entropy, 168
error measure, 28–30
L_1 versus L_2, 38
classification, 28
cross-entropy, 92
fingerprint example, 28
logistic regression, 91
example, **3**

false accept, 29, 115
false reject, 29, 115
falsifiability, 178
feasibility of learning
Boolean example, 16
probabilistic, 18
two main questions, 26
visual example, 15
feature selection, 151
feature space, 100
features, 81
nonlinear transform, 99
feature transform, **100**, 111, 116–117
final exam, 39
financial forecasting, 1
fingerprint verification, 28, 115
football scam, 170

Gaussian processes, 181
generalization, **39–59**
VC bound, 50–59
VC dimension, 50
generalization bound
definition, 40
Devroye, 73
Parrondo and Van den Broek, 73
Rademacher penalty, 73
relative error, 74
VC, *see* VC generalization bound
generalization error
definition, 40
global minimum, 93
gradient descent, 92–99
algorithm, 95
batch, 97
initialization and termination, 95
stochastic, 97
growth function, 41–50
2-dimensional perceptron, 43
bound, 46–49
convex set, 44
definition, 42
in VC proof, 190
polynomial bound, 50
positive interval, 44
positive ray, 43
two-dimensional perceptron, 43

handwritten digit recognition, 4, 11, 81–82, 106–107, 151
hat matrix, **87**, 112
Hessian matrix, 116
Hoeffding bound, *see* Hoeffding Inequality
Hoeffding Inequality, **19**, 19–27
and binomial distribution, 36
uniform version, 24
without replacement, 192
hypothesis set, **3**
composition, 72
concentric spheres, 69
convex set, 44
monotonic, 71
polynomial, 120
positive interval, 44

positive ray, 43
positive rectangles, 69
positive-negative interval, 69
positive-negative ray, 69
restricted to inputs, 42

in-sample error, **21**
input space, **3**
iterative learning, 7

kernel methods, 181

Lagrange multiplier, 131, 157
lasso, 161
law of large numbers, 36, 37
learning
criteria, 26, 78
feasibility, 15–18, 24–26
learning algorithm, **3**
learning curve, 66–68, 140, 147
linear regression, 88
learning model
definition, 5
learning problem
summary figure, 30
learning rate, 94, 95
leave-one-out, 146
Legendre polynomials, 123, 128–129, 154, 155
likelihood, 91
linear classification, 77
linear model, **77**
bias-variance, 158–159
building block, 181
cross validation, analytic, 164
optimal weight decay, 161
overlooked resource, 107
summary, 96
linear programming, 110, 111
linear regression, 82–88, 111
algorithm, 86
bias and variance, 114
for classification, 96–97, 109–110
learning curve, 88
optimal hypothesis, 111
out of sample, 87–88
out-of-sample error, 112
projection matrix, 86, 113
rank deficient, 114
using classification algorithm, 113
linearly separable, 6, 78
example, 6
local minimum, 93
logistic function, 89
logistic regression, 88–99
algorithm, 95
cross-entropy error, 92
error measure, 91–92
for classification, 96–97, 115
hard threshold, 115
initialization, 95
optimal decision theory, 115
termination, 96
loss matrix, 38

machine learning, vii, 14
maximum likelihood, 91
medical diagnosis, 1
minimum description length, 168
model selection, 141–145
choosing λ, 134, 149
cross validation, 148
experiment, 144
summary, 143
monotonic functions, 71
VC dimension, 71
movie rating, 1–3
multiclass, 81

Netflix, 1
neural network, 181
Newton's method, 116
noise
deterministic, 124
stochastic, 124
non-falsifiability, 178
axiom, 170
picking financial traders, 170
non-separable data, 79–81
nonlinear regression, 104
nonlinear transformation, 99
normalization, 175
NP-hard, 80

objective, 28
Occam's razor, 167–171, 181
off training set error, 37
Ω, 58

online learning, 98, 181
    definition, 12
ordinary least squares, 86
out-of-sample error, **21**
outliers, 79
output space, **3**
overfitting, **119–165**, 171
    definition, 119
    experiment, 123, 155
    learning curves, 122

pattern recognition, 9
penalty
    hypothesis complexity, 126, 133
    model complexity, 58
perceptron, 5–8, 78–82
    definition, 5
    experiment, 34
    learning algorithm (PLA), 7
    $m_{\mathcal{H}}(N)$, 70
    PLA convergence, 33
    pocket algorithm, 80
perceptron learning algorithm, 7, 77, 78, 98, 109–110
    and SGD, 98
    convergence, 33
    figure, 7, 83
PLA, *see* perceptron learning algorithm
pocket algorithm, **80**, 97, 109
    figure, 83
poll, 19
    Truman versus Dewey, 171
polynomial transform, 104
polynomials, 120
positive interval, 44
positive ray, 43
postal scam, 170
prediction of heart attacks, 89
probability
    logistic regression, 89
    union bound, 24, 41
projection matrix, 113
pseudo-inverse, 85
    numerical stability, 86
publication bias, 173

quadratic programming, 181

random sample, 19
recommender systems, 1, 15, 181
regression, 77, **82**
    logistic, **89**
regularization, **126–137**, 181
    E_{in} versus λ, 156
    augmented error, 132
    choosing λ, 134, 149
    input noise, 160
    lasso, 161
    linear model, 133
    ridge regression, 132
    soft order constraint, 128
    Tikhonov, 131, 160
    VC dimension, 137
    weight decay, 132
regularization parameter, λ, 133
reinforcement learning, 12, 181
ridge regression, 132
risk, 28
risk matrix, 38, *see also* cost matrix

sample complexity, 56–57
sampling bias, 171–173, 181
    versus data snooping, 177
Sauer's Lemma, 48
search engines, 1
selection bias, 173
SGD, *see* stochastic gradient descent
shatter, 42
sigmoid, 90
singular value decomposition, 114
soft order constraint, 157
soft threshold, 90
spam, 4, 6
squared error, 61, 66, 84, 140
SRM, *see* structural risk minimization
statistics, 14
stochastic gradient descent, 97–99, 110
stochastic noise, 124
streaming data, 12
structural risk minimization, 178
superstition, 119
supervised learning
    definition, 11
support vector machines, 181
supremum, 187
SVD, *see* singular value decomposition

tanh, **90**

- target distribution, 31
- target function, **3**
  - noisy, 30–32, 83, 87
- test set, 59
- Tikhonov regularizer, 131
- Tikhonov smoothness penalty, 162
- training examples, 4
- Truman, 171

- underfitting, 135
- union bound, 24, 41
- unlabeled data, 13, 181
- unsupervised learning, 13, 181
  - learning a language, 13

- validation, **137–141**
  - cross validation, 145
  - model selection, 141
  - summary, 141
  - validation set, 138
- validation error, 138
  - expectation, 138
  - optimistic bias, 142
  - variance, 139
- validation set
  - VC bound, 139, 163
- Vapnik-Chervonenkis, *see* VC
- VC dimension, **50**
  - d-dimensional perceptron, 52
  - and number of parameters, 72
  - definition, 50
  - effective, 137
  - intersection of hypothesis sets, 71
  - monotonic functions, 71
  - of composition, 72
  - union of hypothesis sets, 71
- VC generalization bound, 53, 78, 87, 102
  - definition, 53
  - proof, 187
  - sketch of proof, 53
- VC Inequality, 187
- vending machines, 9
- virtual examples, 157

- weight decay, 132
  - cross validation error, 149
  - example, 126
  - gradient descent, 156
  - invariance under linear transform, 162
  - linear model, 133
  - negative λ, 156
  - optimal λ, 161
  - virtual examples, 157

- $\mathcal{Z}$ space, 99–102