

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

Table 1: Course rating data set

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

Steffen Bickel, Michael Bruckner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2007.

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

Hal Daumé III. Frustratingly easy domain adaptation. In *Conference of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic, 2007.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

Matti Kääriäinen. Lower bounds for reductions. Talk at the Atomic Learning Workshop (TTI-C), March 2006.

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

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1): 81–106, 1986.

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).

Stéphane Ross, Geoff J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Workshop on Artificial Intelligence and Statistics (AISTats)*, 2011.

- K-nearest neighbors, 58
- d_A -distance, 82
- p -norms, 104
- 0/1 loss, 100
- 80% rule, 80

- absolute loss, 14
- activation function, 130
- activations, 41
- AdaBoost, 166
- adaptation, 74
- algorithm, 99
- all pairs, 92
- all versus all, 92
- approximation error, 71
- architecture selection, 139
- area under the curve, 64, 96
- argmax problem, 199
- AUC, 64, 95, 96
- AVA, 92
- averaged perceptron, 52

- back-propagation, 134, 137
- bag of words, 56
- bagging, 165
- base learner, 164
- batch, 173
- Bayes error rate, 20
- Bayes optimal classifier, 19
- Bayes optimal error rate, 20
- Bayes rule, 117
- Bernoulli distribution, 121
- bias, 42
- bias/variance trade-off, 72
- binary features, 30
- bipartite ranking problems, 95
- boosting, 155, 164
- bootstrap resampling, 165
- bootstrapping, 67, 69

- categorical features, 30
- chain rule, 117, 120
- chord, 102
- circuit complexity, 138
- clustering, 35, 178
- clustering quality, 178
- complexity, 34
- compounding error, 215
- concave, 102
- concavity, 193
- concept, 157
- confidence intervals, 68
- constrained optimization problem, 112
- contour, 104
- convergence rate, 107
- convex, 99, 101
- covariate shift, 74
- cross validation, 65, 68
- cubic feature map, 144
- curvature, 107

- data covariance matrix, 184
- data generating distribution, 15
- decision boundary, 34
- decision stump, 168
- decision tree, 8, 10
- decision trees, 57
- density estimation, 76
- development data, 26
- dimensionality reduction, 178
- discrepancy, 82
- discrete distribution, 121
- disparate impact, 80
- distance, 31
- domain adaptation, 74
- dominates, 63
- dot product, 45
- dual problem, 151

- dual variables, 151

- early stopping, 53, 132
- embedding, 178
- ensemble, 164
- error driven, 43
- error rate, 100
- estimation error, 71
- Euclidean distance, 31
- evidence, 127
- example normalization, 59, 60
- examples, 9
- expectation maximization, 186, 189
- expected loss, 16
- expert, 212
- exponential loss, 102, 169

- feasible region, 113
- feature augmentation, 78
- feature combinations, 54
- feature mapping, 54
- feature normalization, 59
- feature scale, 33
- feature space, 31
- feature values, 11, 29
- feature vector, 29, 31
- features, 11, 29
- forward-propagation, 137
- fractional assignments, 191
- furthest-first heuristic, 180

- Gaussian distribution, 121
- Gaussian kernel, 147
- Gaussian Mixture Models, 191
- generalize, 9, 17
- generative story, 123
- geometric view, 29
- global minimum, 106
- GMM, 191

- gradient, 105
- gradient ascent, 105
- gradient descent, 105

- Hamming loss, 202
- hard-margin SVM, 113
- hash kernel, 177
- held-out data, 26
- hidden units, 129
- hidden variables, 186
- hinge loss, 102, 203
- histogram, 12
- horizon, 213
- hyperbolic tangent, 130
- hypercube, 38
- hyperparameter, 26, 44, 101
- hyperplane, 41
- hyperspheres, 38
- hypothesis, 71, 157
- hypothesis class, 71, 160
- hypothesis testing, 67

- i.i.d. assumption, 117
- identically distributed, 24
- ILP, 195, 207
- imbalanced data, 85
- imitation learning, 212
- importance sampling, 75
- importance weight, 86
- independent, 24
- independently, 117
- independently and identically distributed, 117
- indicator function, 100
- induce, 16
- induced distribution, 88
- induction, 9
- inductive bias, 20, 31, 33, 103, 121
- integer linear program, 207
- integer linear programming, 195
- iteration, 36

- jack-knifing, 69
- Jensen's inequality, 193
- joint, 124

- K-nearest neighbors, 32
- Karush-Kuhn-Tucker conditions, 152
- kernel, 141, 145
- kernel trick, 146
- kernels, 54

- KKT conditions, 152
- label, 11
- Lagrange multipliers, 119
- Lagrange variable, 119
- Lagrangian, 119
- lattice, 200
- layer-wise, 139
- learning by demonstration, 212
- leave-one-out cross validation, 65
- level-set, 104
- license, 2
- likelihood, 127
- linear classifier, 169
- linear classifiers, 169
- linear decision boundary, 41
- linear regression, 110
- linearly separable, 48
- link function, 130
- log likelihood, 118
- log posterior, 127
- log probability, 118
- log-likelihood ratio, 122
- logarithmic transformation, 61
- logistic loss, 102
- logistic regression, 126
- LOO cross validation, 65
- loss function, 14
- loss-augmented inference, 205
- loss-augmented search, 205

- margin, 49, 112
- margin of a data set, 49
- marginal likelihood, 127
- marginalization, 117
- Markov features, 198
- maximum a posteriori, 127
- maximum depth, 26
- maximum likelihood estimation, 118
- mean, 59
- Mercer's condition, 146
- model, 99
- modeling, 25
- multi-layer network, 129

- naive Bayes assumption, 120
- nearest neighbor, 29, 31
- neural network, 169
- neural networks, 54, 129
- neurons, 41
- noise, 21

- non-convex, 135
- non-linear, 129
- Normal distribution, 121
- normalize, 46, 59
- null hypothesis, 67

- objective function, 100
- one versus all, 90
- one versus rest, 90
- online, 42
- optimization problem, 100
- oracle, 212, 220
- oracle experiment, 28
- output unit, 129
- OVA, 90
- overfitting, 23
- oversample, 88

- p-value, 67
- PAC, 156, 166
- paired t-test, 67
- parametric test, 67
- parity, 21
- parity function, 138
- patch representation, 56
- PCA, 184
- perceptron, 41, 42, 58
- perpendicular, 45
- pixel representation, 55
- policy, 212
- polynomial kernels, 146
- positive semi-definite, 146
- posterior, 127
- precision, 62
- precision/recall curves, 63
- predict, 9
- preference function, 94
- primal variables, 151
- principle components analysis, 184
- prior, 127
- probabilistic modeling, 116
- Probably Approximately Correct, 156
- programming by example, 212
- projected gradient, 151
- projection, 46
- psd, 146

- radial basis function, 139
- random forests, 169
- random variable, 117
- RBF kernel, 147

- RBF network, 139
- recall, 62
- receiver operating characteristic, 64
- reconstruction error, 184
- reductions, 88
- redundant features, 56
- regularized objective, 101
- regularizer, 100, 103
- reinforcement learning, 212
- representer theorem, 143, 145
- ROC curve, 64

- sample complexity, 157, 158, 160
- sample mean, 59
- sample selection bias, 74
- sample variance, 59
- semi-supervised adaptation, 75
- sensitivity, 64
- separating hyperplane, 99
- sequential decision making, 212
- SGD, 173
- shallow decision tree, 21, 168
- shape representation, 56
- sigmoid, 130
- sigmoid function, 126
- sigmoid network, 139
- sign, 130
- single-layer network, 129
- singular, 110
- slack, 148
- slack parameters, 113
- smoothed analysis, 180
- soft assignments, 190
- soft-margin SVM, 113

- span, 143
- sparse, 104
- specificity, 64
- squared loss, 14, 102
- statistical inference, 116
- statistically significant, 67
- steepest ascent, 105
- stochastic gradient descent, 173
- stochastic optimization, 172
- strong law of large numbers, 24
- strong learner, 166
- strong learning algorithm, 166
- strongly convex, 107
- structural risk minimization, 99
- structured hinge loss, 203
- structured prediction, 195
- sub-sampling, 87
- subderivative, 108
- subgradient, 108
- subgradient descent, 109
- sum-to-one, 117
- support vector machine, 112
- support vectors, 153
- surrogate loss, 102
- symmetric modes, 135

- t-test, 67
- test data, 25
- test error, 25
- test set, 9
- text categorization, 56
- the curse of dimensionality, 37
- threshold, 42
- Tikhonov regularization, 99

- time horizon, 213
- total variation distance, 82
- train/test mismatch, 74
- training data, 9, 16, 24
- training error, 16
- trajectory, 213
- trellis, 200
- truncated gradients, 175
- two-layer network, 129

- unary features, 198
- unbiased, 47
- underfitting, 23
- unit hypercube, 39
- unit vector, 46
- unsupervised adaptation, 75
- unsupervised learning, 35

- validation data, 26
- Vapnik-Chernovenkis dimension, 162
- variance, 59, 165
- variational distance, 82
- VC dimension, 162
- vector, 31
- visualize, 178
- vote, 32
- voted perceptron, 52
- voting, 52

- weak learner, 166
- weak learning algorithm, 166
- weights, 41

- zero/one loss, 14