
Index

- F_1 score, 97
- ℓ_1, ℓ_2, ℓ_p norms, 29
- active path, 51
- algorithm
 - ν -SV classification, 30
 - conditional graphical model, 276
 - cost-sensitive SVM, 95
 - Eulerian circuit, 156
 - generic search, 171
 - joint kernel maps, 73
 - junction tree, 54
 - kernel dependency estimation, 22, 146
 - kernel PCA, 13
 - LaSO, 171
 - max-margin conditional random field, 120
 - max-margin Markov networks, 233, 272
 - multiclass perceptron, 28
 - norma, 27
 - parzen, 5
 - perceptron, 26
 - self-bounded PST-perceptron, 140
 - sparse greedy subspace approximation, 292
 - SV support estimation, 31
 - SVM – dual, 29
 - SVM – primal, 29
 - unbounded-depth PST-perceptron, 134
- approximate KL minimization, 284
- asymptotic consistency, 248
- Bayes rule, 267
- Bayesian conditional random fields, 270
- Bayesian networks, 47
- beam size, 181
- belief propagation, 299
- boosting loss, 40
- bottleneck approximation, 300
- bounds for parsing, 249
- Chomsky normal form, 57
- co-kriging, 304
- collapse of EBMs, 215
- compatibility, 87
- compatibility function, 61
- compatibility measure, 193
- concentration of measure, 37
- conditional density estimation, 196
- conditional gradient ascent, 119
- conditional graphical models, 274
- conditional independence, 44
- conditional models, 60
- conditional random fields, 233, 268
- conditional subspace gradient ascent, 121
- context function, 131
- context tree weighting, 129
- context-free grammars, 58
- contrastive divergence algorithm, 244
- correlation, 68
- cross-covariance, 305
- cummulant-generating, 56
- curse-of-dependency problem, 300
- cutting-plane algorithm, 94
- d-separation, 51
- De Bruijn graph, 155
- decoder, 247

- decoding, 22, 170
- decoding as energy minimization, 194
- dense densities, 285
- detection, 196
- directed acyclic graph, 47
- DTW, *see* dynamic time warping
- dynamic time warping, 235

- EBMs, 191
- EDT, 182
- energy function, 193
- energy loss, 201
- energy-based models, 191
- entity detection and tracking, 182
- Euler circuit, 155
- exponential families, 56, 285

- face detection, 212
- face detection and pose estimation, 194
- factor analysis, 305
- factor graph, 227
- feature sharing, 88
- feature space, 4
- Fisher information matrix, 21
- Fisher scores, 21

- Gauss-Seidel, 292
- Gauss-Southwell, 293
- Gaussian process belief propagation, 306
- Gaussian process classification loss, 33
- gene function prediction, 19
- generalization across classes, 88
- generalized margin loss, 201
- generalized perceptron loss, 201
- generalizing support vector machines, 69
- Gibbs distribution, 197
- Gram matrix, 6
- graph transformer network, 236
- GTN, *see* graph transformer network
- Hammersley-Clifford theorem, 47
- Hamming distance, 248
- handwriting recognition, 194
- handwriting segmentation and recognition, 239
- hierarchical classification, 88, 99
- hinge loss, 32, 202
- hyperedge potentials, 108

- identity map, 79
- image segmentation, restoration, 194
- implicit regression, 209
- incomplete Cholesky decomposition, 159
- independence graph, 47
- independence map, 44, 48
- input-output constraints, 148
- interactions between output labels, 89

- joint feature map, 62
- joint input-output space, 87
- joint kernel maps, 71
- joint labeling problems, 35
- joint learning, 187
- joint tagging and chunking, 179
- junction tree, 54
- junction tree algorithm, 47

- KDE, *see* algorithms, kernel dependency estimation
- Kernel
 - all-pairs shortest-path, 21
 - centering, 12
 - concatenation, 87
 - construction of kernels, 8
 - convolution kernels, 18
 - diffusion, 19
 - from generative models, 21
 - Gaussian, 9
 - graph kernel, 20
 - joint, 76
 - MRF decomposition, 288
 - n-gram kernel, 17
 - PCA, 13
 - positive definite, 6
 - rational kernel, 15

- reproducer property, 8
- set kernel, 14
- spline kernel, 10
- sufficient statistics, 287
- tensor product, 77
- tree kernel, 20
- vector-valued, 75
- kernel dependency estimation, 23
- kernel PCA map, 14
- label interdependency, 89
- label sequence learning, 89, 100
- label-bias problem, 229
- Lagrange function, 29
- large-margin separating hyperplane, 28
- late normalization, 229
- latent variables, 210
- learning as search optimization, *see* algorithm, LaSO
- linear regression, 68
- local loss function, 289
- log loss, 202
- log-linear model, 108
- log-partition, 56
- logistic regression, 248
- loopy belief propagation, 122
- LVQ2 loss, 202
- M³Ns, *see* algorithms, max-margin Markov networks
- Markov blanket, 45
- Markov networks, 44
- Markov property of graphs, 45
- Markov random fields, 44
- mass spectrometry, 80
- maximal cliques, 47
- maximum a posteriori, 267
- maximum entropy, 283
- maximum entropy distribution, 284
- mediator variable, 300
- microlabel, 107
- minimum classification error loss, 203
- missing probability mass problem, 229
- missing variables, 210
- MRF, 44
- multiclass logistic regression, 33
- multiclass loss, 32
- multilabel, 107
- multilabel estimation problems, 36
- named entity recognition, 296
- negative log-likelihood loss, 204
- neighbor graph, 19
- neighbor of a word, 19
- novelty detection loss function, 33
- offending answers, 202
- one-class SVM, *see* algorithms, SV support estimation
- online learning, 26
- online update, 27
- ordinal regression, 36
- PAC-Bayesian generalization bound, 249
- parsing, 90, 101
- partition function, 108, 197
- path cost, 171
- pitch accent prediction, 294
- potential functions, 47, 288
- pre-image, 170
- pre-image problem, 22
- pre-images for n-gram kernels, 154
- prediction suffix trees, 129
- probabilistic context-free grammars, 57
- pseudoexample, 108
- quadratic program, 29
- Rademacher averages, 37, 38
- ranking, 196
- relative mistake bound for unbounded PSTs, 134
- representer theorem, 10
- RKHS, 8
- search, 171
 - operators, 171

- problem, 171
- states, 171
- search margin, 173
- semi-infinite program, 73
- semi-parametric Markov random field,
294
- semisupervised learning, 19
- sequence labeling, 194, 293
- Siamese architectures, 210
- smiling face, 81
- soft-margin loss, 40
- softmax, 267
- sparse approximation schemes, 29
- sparse greedy subspace approxima-
tion, 291
- square-exponential loss, 203
- square-square loss, 203
- statistical independence, 43
- stochastic approximation of risk, 27
- stochastic gradient descent, 208
- string-to-string, 144
- string-to-string mapping, 143
- structured output, 71
- sufficient statistics, 56, 301
- support vectors, 29
- symmetrization, 38
- syntactic chunking, 177

- taxonomies in categorization, 33, 88
- TDNN, *see* time-delay neural net-
work
- time-delay neural network, 235
- transfer learning, 304
- tree reparametrization algorithm, 122
- Tsybakov's noise condition, 41

- variable-length Markov models, 129