

# Index

- $1 \times 1$  convolution, 407
- $L_2$  norm, 232
- $k$ -means, 161
- $k$ -means++, 162
  
- absorbing state, 307
- accuracy, 4
- additive, zero mean, independent
  - Gaussian noise, 103
- affine transformation, 80
- affinity, 165
- AIC, 248
- albedo, 144
- AlexNet, 428
- all-vs-all, 33
- anchor box, 442
- approximate nearest neighbor, 9
- autoencoder, 461
- average pooling, 408
- average precision, 447
  
- backpropagation, 383
- Backward stagewise regression, 251
- backward variable, 326
- bag, 42
- bag-of-words, 126
- bagging, 42
- baselines, 4
- batch, 26
- batch normalization layer, 432
- batch size, 26
- Bayes risk, 4
- bias, 246, 367
- biased random walk, 306
- BIC, 249
- bigram models, 314
- bigrams, 314
- binary terms, 334
- blocks, 404
- Boltzmann machine, 351
- boosting, 275
  
- bounding box regression, 439
- Box-Cox transformation, 216
  
- canonical correlations, 140
- canonical variables, 140
- chain graph, 334
- class conditional probability, 11
- class error rate, 6
- class-confusion matrix, 5
- classifier, 3
  - definition, 3, 21
  - nearest neighbors, 7
- clustering, 155
  - complete-link clustering, 156
  - dendrogram, 156, 157
  - group average clustering, 156
  - grouping and agglomeration, 155
  - partitioning and division, 155
  - single-link clustering, 156
  - using K-means, 159
- clusters, 79, 155
- coefficients, 97
- color constancy, 106
- comparing to chance, 4
- computation graph, 434
- condition number, 242
- conditional random field, 341
- configuration, 439
- convolution, 401
- convolutional layer, 405
- Cook's distance, 223
- cosine distance, 127
- cost to go function, 320
- cost-to-go function, 334
- covariance ellipses, 85
- cross-entropy loss, 370
- cross-validation, 7
  - folds, 250

- data augmentation, 427
- dead units, 376
- decay rate, 392
- decision boundary, 21
- decision forest, 35
- decision function, 35
- decision stump, 288
- decision tree, 34
- decoder, 461
- deconvolution layer, 462
- decorrelation, 157
- deep networks, 383
- dendrogram, *see* clustering
- denoising autoencoder, 465
- dependent variable, 209
- descent direction, 25
- descriptive statistics, 86
- deviance, 262
- discrete Markov random field, 353
- discriminative, 338
- discriminator, 471
- distributional semantics, 131
- document-term matrix, 128
- dropout, 387
- dynamic programming, 318
  
- E-step, 189
- edge points, 469
- edge terms, 334
- eigenvalue, 81
- eigenvector, 81
- elastic net, 267
- EM, 183, 188
- embedding, 456
- emission distribution, 316
- empirical distribution, 86
- empirical loss, 284
- encoder, 461
- energy, 352
- epoch, 27
- error, 4
- error cost, 263
- expectation maximization, 183, 188
- explanatory variables, 209
- exponential loss, 289
  
- false negative rate, 5
- false positive rate, 5
- Fast R-CNN, 441
- Faster R-CNN, 441
- feature maps, 404
- feature stack, 431
- feature vector, 4
- features, 406
- fold, 7
- folders, *see* cross-validation
- forward stagewise regression, 251
- forward variable, 326
- Frobenius norm, 100
- fully connected, 377
  
- GAN, 471
- Gaussian distribution, 83
- generalized linear model, 258
- generalizing badly, 6
- generative, 338
- generative adversarial networks, 471
- generator, 471
- GLM, 258
- gradient boost, 287
- gradient descent, 25
- graphical model, 336
- greedy stagewise linear regression, 277
- greedy stagewise regression, 281
- group normalization, 433
- growth function, 58
  
- Hamming distance, 347
- hat matrix, 222
- hidden layers, 377
- hidden Markov model, 316
- hinge loss, 23
- hourglass networks, 461
- Huber loss, 254
- Huber's proposal 2, 257
  
- idempotence, 243
- image classification, 399
- ImageNet, 426
- inception modules, 434
- inference, 317
- information gain, 38

- inlier, 254
- inpainting autoencoder, 465
- intensity, 261
- interpolated precision, 446
- IoU, 445
- irreducible, 311
- irreducible error, 246
- iteratively reweighted least squares, 256
  
- Jacobian, 379
  
- k-means, *see* clustering
- kernel, 401
- kernel block, 404
- KL divergence, 359
- Kullback–Leibler divergence, 359
  
- L2 regularized error, 263
- label bias problem, 340
- lasso, 266
- latent semantic analysis, 128
- layers, 377
- learning curves, 27
- learning rate, 26, 372
- learning schedule, 27
- leave-one-out cross-validation, 7
- leverage, 222
- likelihood, 11
- Likert scales, 126
- line search, 25, 286
- link function, 258
- loadings, 97
- localize, 439
- log-loss, 370
- logistic regression, 259
- logit function, 259
- loss, 284
- loss augmented constraint violation, 347
- low rank approximation, 117
  
- M-estimator, *see* robustness
- M-step, 189
- MAD, 257
- MAP, 317
- Markov chain, 305
- Markov random field, 353
  
- mask, 401
- max pooling, 408
- max-cut, 353
- maximum a posteriori, 317
- maximum entropy Markov models, 339
- mean average precision, 447
- mean field method, 363
- mean-squared error, 212
- median absolute deviation, 257
- MEMM, 339
- minibatch training, 371
- mixing weights, 184
- mixture of normal distributions, 184
- mode collapse, 472
- multidimensional scaling, 123
- multivariate normal distribution, 83
  
- n-gram models, 314
- n-grams, 314
- neural network, 377
- neurons, 367
- non-maximum suppression, 439
- normalizing constant, 337
  
- object detection, 399
- one hot, 369
- one-hot vectors, 354
- one-vs-all, 33
- outliers, 219
- overcomplete, 461
- overfitting, 6
  
- padding, 403
- Pascal, 425
- PCA, 97
- perceptual loss, 466
- perplexity, 459
- phonemes, 317
- Places-2, 427
- pointwise loss, 50, 285
- pooling, 408
- posterior, 11
- precision, 446
- predictor, 49, 275, 284
- principal components, 97
- principal components analysis, 97

- principal coordinate analysis, 123
- prior, 11
  
- R-CNN, 441
- recall, 446
- receptive field, 407
- recurrent, 307
- regions, 440
- Regression, 205
- regression tree, 279
- regularization, 24
- regularization parameter, 24
- regularization path, 264
- regularizer, 24
- ReLU, 367
- residual, 208, 212
- residual connections, 436
- ResNets, 436
- ridge regression, 230
- robust regression, 254
- robustness
  - M-estimator, 254
    - influence function, 254
  - M-estimators
    - scale, 257
- ROI pooling layer, 441
  
- Sammon mapping, 456
- scale, 254
  - of an M-estimator, 257
- scene, 426
- scores, 97
- selection bias, 6
- selective search, 440
- sensitivity, 5
- shading, 144
- shattering number, 58
- sigmoid layer, 463
- singular value decomposition, 117
- singular values, 118
- smooth, 105
- smoothing, 315
- softmax function, 368
- sparse models, 262
- specificity, 5
- standardizing, 225
- stationary distribution, 311
  
- statistical significance, 252
- stem, 127
- steplength, 26
- steplength schedule, 27
- stepsize, 26, 372
- Stochastic gradient descent, 26
- stochastic matrices, 309
- stop words, 127
- stride, 403
- sum-products algorithm, 343
- SUN, 426
- support vector machine, 23
- SVD, 117
- SVM, 23
- symmetric, 81
  
- term frequency-inverse document
  - frequency, 134
- term-document matrix, 128
- test error, 6
- test examples, 209
- TF-IDF, 134
- topic, 185
- topic model, 186
- total error rate, 4
- training error, 6
- training examples, 209
- training loss, 49
- transition probabilities, 305
- transposed convolution layer,
  - 462
- trellis, 318
- trigram models, 314
- trigrams, 314
  
- unary terms, 334
- unbiased, 7
- unigram models, 314
- unigrams, 314
- union bound, 55
- unit, 367
- unpooling, 463
  
- validation set, 7
- variance, 246
- variational autoencoders, 470
- variational free energy, 360

- variational inference, 358
- VC dimension, 59
- vector quantization, 172
- vertex terms, 334
- VGG-19, 430
- Viterbi algorithm, 318
- VQ, 172
  
- weak law of large numbers, 86
- weak learners, 275
- weights, 367
- whitening, 8, 157
  
- Wilks' lambda, 148
- witness function, 473
- word embedding, 133
- word probabilities, 185
- word vectors, 127
- WordNet, 426
  
- XGBoost, 294
  
- YOLO, 443
  
- Zipf's law, 215

# Index: Useful Facts

- Bernoulli Random Variable, 259
- Chebyshev's Inequality, 52
- Convolutional Layer, 405
- Covariance, 74
- Covariance Matrix, 75
- Definition: Bernoulli Random Variable, 259
- Definition: Chebyshev's Inequality, 52
- Definition: Convolutional Layer, 405
- Definition: Covariance, 74
- Definition: Covariance Matrix, 75
- Definition: Hoeffding's Inequality, 54
- Definition: Linear Regression, 209
- Definition: Poisson Distribution, 261
- Definition: Regression, 208
- Definition: The VC Dimension, 60
- Generalization Bound in Terms of VC Dimension, 62
- Held-Out Error Predicts Test Error, from Chebyshev, 52
- Hoeffding's Inequality, 54
- Linear Regression, 209
- Many Markov Chains Have Stationary Distributions, 312
- Markov Chains, 309
- Mean and Variance of an Expectation Estimated from Samples, 51
- Orthonormal Matrices Are Rotations, 82
- Parameters of a Multivariate Normal Distribution, 84
- Poisson Distribution, 261
- Properties of the Covariance Matrix, 75
- Regression, 208, 213
- The Growth Number of a Family of Finite VC Dimension, 61
- The Largest Variation of Sample Means Yields a Bound, 62
- The VC Dimension, 60
- Transition Probability Matrices, 311
- Weak Law of Large Numbers, 87
- You Can Transform Data to Zero Mean and Diagonal Covariance, 83

# Index: Procedures

- k*-Means Clustering, 161
- k*-Means with Soft Weights, 166
  
- Agglomerative Clustering, 156
  
- Building a Decision Forest, 42
- Building a Decision Forest Using Bagging, 42
- Building a Decision Tree: Overall, 40
- Building a Dictionary for VQ, 173
  
- Canonical Correlation Analysis, 141
- Choosing a  $\lambda$ , 28
- Classification with a Decision Forest, 42
- Computing Cook's Distance, 223
- Computing the Backward Variable for Fitting an HMM, 328
- Computing the Forward Variable for Fitting an HMM, 328
- Cross-Validation to Choose a Model, 14
  
- Diagonalizing a Symmetric Matrix, 81
- Divisive Clustering, 156
  
- EM, 194
- EM for Mixtures of Normals: E-step, 194
- EM for Mixtures of Normals: M-step, 194
- EM for Topic Models: E-step, 195
- EM for Topic Models: M-step, 195
  
- Fitting a Regression with Iteratively Reweighted Least Squares, 257
  
- Fitting Hidden Markov Models with EM, 328
  
- Gradient Boost, 287
- Greedy Stagewise Linear Regression, 278
- Greedy Stagewise Regression with Trees, 282
  
- Learning a Decision Stump, 289
- Linear Regression Using Least Squares, 218
  
- Obtaining Some Principal Components with NIPALS, 103
  
- Principal Components Analysis, 99
- Principal Coordinate Analysis, 124
  
- Representing a Signal Using VQ, 174
  
- Simple Image Whitening, 415
- Singular Value Decomposition, 118
- Splitting a Non-ordinal Feature, 41
- Splitting an Ordinal Feature, 40
  
- Training an SVM: Estimating the Accuracy, 29
- Training an SVM: Overall, 29
- Training an SVM: Stochastic Gradient Descent, 30
  
- Updating Parameters for Fitting an HMM, 329

# Index: Worked Examples

- $s(\mathcal{B}, 3)$  for a Simple Linear Classifier on the Line, 58
- $s(\mathcal{B}, 4)$  for a Simple Linear Classifier on the Plane, 59
- A Simple Discrete MRF for Image Denoising, 355
- Agglomerative Clustering, 159
- AIC and BIC, 250
- Anxiety and Wildness in Mice, 141
- Building an L1 Regularized Regression, 265
- Classifying Breast Tissue Samples, 13
- Classifying Heart Disease Data, 43
- Classifying Using Nearest Neighbors, 9
- Cross-Validation, 251
- Denoising MRF—II, 356
- Denoising MRF—III, 356
- Greedy Stagewise Regression for Prawns, 281
- L1 Regularized Regression for a “Wide” Dataset, 269
- MAP Inference for MRFs Is a Linear Program, 358
- Modelling Short Words, 314
- Modelling Text with n-Grams of Words, 315
- Multiclass Logistic Regression with an L1 Regularizer, 270
- Multiple Coin Flips, 308
- Opioid Prescribers with XGBoost, 297
- Predicting the Quality of Education of a University, 296
- Regressing Prawn Scores Against Location, 279
- Sammon Mapping MNIST Data, 457
- T-SNE on MNIST Data, 459
- The Gambler’s Ruin, 307
- The VC Dimension of the Binary Functions Produced by a Linear Classifier on the Line, 60
- The VC Dimension of the Binary Functions Produced by a Linear Classifier on the Plane, 60
- Umbrellas, 306
- Umbrellas, but Without a Stationary Distribution, 312
- Useful Facts About MRFs, 357
- Viruses, 310, 311



# Index: Remember This

- Boosting: classification and regression differ by training loss, 285
- Boosting: gradient boosting builds a predictor greedily, 288
- Boosting: gradient boosting decision stumps is a go-to, 289
- Boosting: predicting the weights in gradient boost is easier than it looks, 291
- Boosting: the lasso can prune boosted models, 294
- Boosting: use XGBoost for big gradient boosting problems, 296
  
- CCA can mislead you, 150
- Choosing Models: AIC and BIC, 250
- Choosing models: an  $L_1$  regularization penalty encourages zeros in models, 264
- Choosing Models: stagewise regressions are greedy searches, 252
- Choosing models: Use the lasso, 268
- Classifier: an SVM is a linear classifier trained with the hinge loss, 24
- Classifier: any SVM package should build a multiclass classifier for you, 34
- Classifier: definition, 3
- Classifier: do not evaluate a classifier on training data, 7
- Classifier: enough examples make a bad predictor unlikely, 56
- Classifier: good and bad properties of nearest neighbors, 10
- Classifier: held-out error predicts test error (Chebyshev), 53
- Classifier: held-out error predicts test error (Hoeffding), 55
- Classifier: linear SVM's are a go-to classifier, 33
- Classifier: look at false positive rate and false negative rate together, 6
- Classifier: naive Bayes is simple, and good for high dimensional data, 16
- Classifier: nearest neighbors has theoretical guarantees on the error rate, 8
- Classifier: performance summarized by accuracy or error rate, 5
- Classifier: random forests are good and easy, 44
- Classifier: regularization discourages large errors on future data, 25
- Classifier: test error bounds from training error, finite set of predictors, 56
- Classifier: train linear SVM's with stochastic gradient descent, 27
- Classifier: VC dimension of linear classifiers, 61
- Clustering: agglomerative and divisive clustering, 159
- Clustering: K-means is the "go-to" clustering recipe, 171
- Clustering: tips for using EM to cluster, 196
- Covariance: correlation from covariance, 74
- Covariance: mean and covariance of affine transformed dataset, 81
- Covariance: mean, variance and covariance can be used in

- two senses, 87
- CRFs: a CRF has edge weights that are not joint probabilities, 341
- CRFs: dynamic programming works for forest models, 336
- CRFs: HMM's are generative, which is inconvenient, 339
- CRFs: label bias means you should not use MEMM's, 341
- CRFs: learn CRF's discriminatively, 347
- CRFs: learning a CRF takes care, 343
  
- EM: EM is a quite general algorithm, 197
  
- Graphical models: a natural denoiser for binary images is intractable, 353
- Graphical models: A natural denoiser for images is also intractable, 358
- Graphical models: KL divergence measures the similarity of two probability distributions, 359
- Graphical models: maximum likelihood estimation uses KL divergence, 360
- Graphical models: mean field inference works well for denoising binary images, 364
- Graphical models: the variational free energy bounds KL divergence, and is tractable, 361
- Graphical models: variational inference uses an easy distribution close to an intractable model, 359
  
- High dimensions: high dimensional data displays odd behavior, 79
  
- High dimensions: the multivariate normal distribution, 86
  
- Image classification:  $1 \times 1$  convolution=linear map, 407
- Image classification: Alexnet was a spectacular success at image classification, 430
- Image classification: batch or group normalization can help training, 433
- Image classification: CIFAR-100 is a small hard dataset, 425
- Image classification: convolutional layer + ReLU=Pattern detector, 406
- Image classification: ImageNet is the a standard large scale image classification dataset, 426
- Image classification: Inception networks handle features at multiple spatial scales, 436
- Image classification: MNIST and CIFAR-10 are warmup datasets, 425
- Image classification: PASCAL VOC 2007 remains a standard image classification dataset, 426
- Image classification: pattern detector responses are usually sparse, 406
- Image classification: Places-2 is a large-scale scene classification dataset, 427
- Image classification: ResNets can be very deep and very accurate, 438
- Image classification: SUN is a large-scale scene classification dataset, 427
- Image classification: there are two common meanings of "convolutional layer", 406
- Image classification: VGGNet was a spectacular success at

- image classification, 432
- Image generation: adversarial losses improve generators, 473
- Image generation: generators can be trained by matching expectations, 474
- Image generation: generators can be trained by matching one dimensional distributions, 475
- Image generation: variational autoencoders can generate images, 471
  
- Mapping and Autoencoding: perceptual loss results in less blurry autoencoded images, 469
- Mapping and Auto-encoding: T-SNE is the first choice to embed data in a low dimensional space., 460
- Mapping and Autoencoding: autoencoders learn to make small codes which allow reconstruction., 462
- Mapping and Autoencoding: Sammon mapping embeds data in a low dimensional space, 457
- Mapping and autoencoding: training an autoencoder to denoise can be helpful, 466
- Mapping and Autoencoding: transposed convolution (deconvolution) makes small blocks larger, 464
- Modelling: three kinds of error: irreducible, bias and variance, 248
  
- Neural: backpropagation yields gradients, 383
- Neural: basic ideas for multilayer networks, 380
- Neural: dropout can be useful, 387
- Neural: gradient tricks can help, 393
  
- Neural: making fully connected layers with convolutional layers, 405
- Neural: Multilayer networks work well, 390
- Neural: softmax yields class posterior, 369
- Neural: train networks by minimizing loss, 370
- Neural: training can be hard, 385
- Neural: use a software environment, 386
  
- Object detection: evaluating object detectors is fiddly, 447
- Object detection: how object detectors work, 440
- Object detection: how R-CNN, Fast R-CNN and Faster R-CNN work, 443
- Object detection: selective search finds boxes likely to contain objects, 441
- Object detection: YOLO trades off speed with accuracy, 445
  
- PCA: a few principal components can represent a high-D dataset, 98
- PCA: PCA can significantly reduce noise, 105
  
- Regression: appending functions of a measurement to  $\mathbf{x}$  is useful, 227
- Regression: be suspicious of points with high Cook's distance, 224
- Regression: be suspicious of points with high leverage, 223
- Regression: Estimating  $\beta$ , 211
- Regression: evaluate a GLM with the model's deviance, 262
- Regression: Evaluates the quality of predictions made by a regression with  $R^2$ , 214
- Regression: generalize linear regression with a GLM, 258

- Regression: greedy stagewise linear regression is an important core recipe, 278
- Regression: greedy stagewise regression can fit using many regression trees, 284
- Regression: interpreting regression coefficients is harder than you think, 253
- Regression: linear regressions can fail, 217
- Regression: logistic regression is one useful GLM, 260
- Regression: multiclass logistic regression is another useful GLM, 260
- Regression: outliers can affect linear regressions significantly, 222
- Regression: predict count data with a GLM, 261
- Regression: regression trees are like classification trees, 279
- Regression: samples of a standard normal random variable, 225
- Regression: the Box-Cox transformation, 217
- Regression: the hat matrix mixes training  $y$ -values to produce predictions, 222
- Regression: transforming variables is useful, 215
- Regression: you can regularize a regression, 232
- SVD: represent documents with smoothed word counts, 129
- SVD: reweight word counts with TF-IDF, 135
- SVD: smoothed document counts are a clue to word meanings, 133
- SVD: strong word embeddings require finer measures of word similarity, 134
- SVD: the SVD decomposes a matrix in a useful way, 118
- SVD: the SVD smoothes a data matrix, 121
- SVD: the SVD yields principal components, 120