Bibliography

REFERENCES

1. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Control 1974;AC-19(6):716–23.
2. Ali SM, Silvey SD. A general class of coefficients of divergence of one distribution from another. J Roy Statist Soc Ser B 1966;28(1):131–42.
3. Aloise D, Deshpande A, Hansen P, Popat P. NP-hardness of Euclidean sum-of-squares clustering. Mach Learn 2009;75(2):245–9.
4. Amari S. Theory of adaptive pattern classifiers. IEEE Trans Electron Comput 1967;EC-16(3):299–307.
5. Amari S, Nagaoka H. Methods of information geometry. Providence (RI, USA): Oxford University Press; 2000.
6. Amit Y, Fink M, Srebro N, Ullman S. Uncovering shared structures in multiclass classification. In: Ghahramani Z, editor. Proceedings of the 24th annual international conference on machine learning. Omnipress; 2007. p. 17–24.
7. Anderson TW. An introduction to multivariate statistical analysis. 2nd ed. New York (NY, USA): Wiley; 1984.
8. Argyriou A, Evgeniou T, Pontil M. Convex multi-task feature learning. Mach Learn 2008;73(3):243–72.
9. Aronszajn N. Theory of reproducing kernels. Trans Amer Math Soc 1950;68:337–404.
10. Auer P. Using confidence bounds for exploitation-exploration trade-offs. J Mach Learn Res 2002;3:397–422.
11. Bach FR, Lanckriet GRG, Jordan MI. Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the twenty-first international conference on machine learning. New York (NY, USA): ACM Press; 2004. p. 6–13.
12. Bartlett P, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ, editors. Practical Bayesian optimization of machine learning algorithms; 2012.
13. Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 2003;15(6):1373–96.
14. Bengio Y. Learning deep architectures for AI. Found Trends Mach Learn 2009;1(2):1–127.
15. Bishop CM. Pattern recognition and machine learning. New York (NY, USA): Springer; 2006.
16. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993–1022.
17. Borgwardt KM, Gretton A, Rasch MJ, Kriegel H-P, Schölkopf B, Smola AJ. Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 2006;22(14):e49–57.
18. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Haussler D, editor. Proceedings of the fifth annual ACM workshop on computational learning theory. ACM Press; 1992. p. 144–52.
19. Boucheron S, Lugosi G, Bousquet O. Concentration inequalities. In: Bousquet O, von Luxburg U, Rätsch G, editors. Advanced lectures on machine learning. Lecture notes in computer science, vol. 3176. Berlin (Heidelberg): Springer; 2004. p. 208–40.
20. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn 2011;3(1):1–122.
21. Breiman L. Bagging predictors. Mach Learn 1996;26(2):123–40.
22. Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

23. Breunig MM, Kriegel H-P, Ng RT, Sander J. LOF: identifying density-based local outliers. In: Chen W, Naughton JF, Bernstein PA, editors. Proceedings of the ACM SIGMOD international conference on management of data; 2000. p. 93–104.
24. Caruana R. Multitask learning. Mach Learn 1997;28:41–75.
25. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. Technical report, Department of Computer Science, National Taiwan University, http://www.csie.ntu.edu.tw/čjlin/libsvm/; 2001.
26. Chang C-C, Tsai H-C, Lee Y-J. A minimum enclosing balls labeling method for support vector clustering. Technical report, National Taiwan University of Science and Technology, 2007.
27. Chang W-C, Lee C-P, Lin C-J. A revisit to support vector data description. Technical report, National Taiwan University, 2013.
28. Chapelle O, Schölkopf B, Zien A, editors. Semi-supervised learning. Cambridge (MA, USA): MIT Press; 2006.
29. Chung FRK. Spectral graph theory. Providence (RI, USA): American Mathematical Society; 1997.
30. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.
31. Cramér H. Mathematical methods of statistics. Princeton (NJ, USA): Princeton University Press; 1946.
32. Crammer K, Dekel O, Keshet J, Shalev-Shwartz S, Singer Y. Online passive-aggressive algorithms. J Mach Learn Res 2006;7(March):551–85.
33. Crammer K, Kulesza A, Dredze M. Adaptive regularization of weight vectors. In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A, editors. Advances in neural information processing systems, vol. 22. 2009. p. 414–22.
34. Crammer K, Singer Y. On the algorithmic implementation of multiclass kernel-based vector machines. J Mach Learn Res 2001;2:265–92.
35. Csiszár I. Information-type measures of difference of probability distributions and indirect observation. Studia Sci Math Hungar 1967;2:229–318.
36. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc Ser B 1977;39(1):1–38.
37. Domingo C, Watanabe O. MadaBoost: a modification of AdaBoost. In: Proceedings of the thirteenth annual conference on computational learning theory; 2000. p. 180–9.
38. du Plessis MC, Sugiyama M. Semi-supervised learning of class balance under class-prior change by distribution matching. Neural Netw 2014;50:110–9.
39. Duchi J, Shalev-Shwartz S, Singer Y, Chandra T. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In: McCallum A, Roweis S, editors. Proceedings of the 25th annual international conference on machine learning. Omnipress; 2008. p. 272–9.
40. Evgeniou T, Pontil M. Regularized multi-task learning. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2004. p. 109–17.
41. Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugenics 1936;7(2):179–88.
42. Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugenics 1936;7:179–88.
43. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting. Ann Statist 2000;28(2):337–407.
44. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 2008;9(3):432–41.
45. Fukumizu K, Sriperumbudur BK, Gretton A, Schölkopf B. Characteristic kernels on groups and semigroups. In: Koller D, Schuurmans D, Bengio Y, Bottou L, editors. Advances in neural information processing systems, vol. 21. 2009. p. 473–80.
46. Gärtner T. Kernels for structured data. Singapore: World Scientific; 2008.

47. Geman S, Geman D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 1984;6:721–41.
48. Girolami M. Mercer kernel-based clustering in feature space. IEEE Trans Neural Netw 2002;13(3):780–4.
49. Gretton A, Borgwardt KM, Rasch M, Schölkopf B, Smola AJ. A kernel method for the two-sample-problem. In: Schölkopf B, Platt J, Hoffman T, editors. Advances in neural information processing systems, vol. 19. Cambridge (MA, USA): MIT Press; 2007. p. 513–20.
50. Gretton A, Sriperumbudur B, Sejdinovic D, Strathmann H, Balakrishnan S, Pontil M, et al. Optimal kernel choice for large-scale two-sample tests. In: Bartlett P, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in neural information processing systems, vol. 25. 2012. p. 1214–22.
51. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci USA 2004;101:5228–35.
52. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 1970;57(1):97–109.
53. He X, Niyogi P. Locality preserving projections. In: Thrun S, Saul L, Schölkopf B, editors. Advances in neural information processing systems, vol. 16. Cambridge (MA, USA): MIT Press; 2004. p. 153–60.
54. Hinton GE. Training products of experts by minimizing contrastive divergence. Neural Comput 2002;14(8):1771–800.
55. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313(5786):504–7.
56. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970;12(3):55–67.
57. Holland PW, Welsch RE. Robust regression using iteratively reweighted least-squares. Comm Statist Theory Methods 1978;6(9):813–27.
58. Huber PJ. Robust statistics. New York (NY, USA): Wiley; 1981.
59. Jolliffe IT. Principal component analysis. New York (NY, USA): Springer-Verlag; 1986.
60. Kawahara Y, Sugiyama M. Sequential change-point detection based on direct density-ratio estimation. Stat Anal Data Min 2012;5(2):114–27.
61. Kawakubo H, du Plessis M, Sugiyama MC. Coping with class balance change in classification: class-prior estimation with energy distance. Technical report IBISML2014-71, IEICE, 2014.
62. Knuth DE. Seminumerical algorithms. The art of computer programming. vol. 2 Reading (MA, USA): Addison-Wesley; 1998.
63. Konishi S, Kitagawa G. Generalized information criteria in model selection. Biometrika 1996;83(4):875–90.
64. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat 1951;22:79–86.
65. Lafferty J, Mccallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th international conference on machine learning; 2001. p. 282–9.
66. Langford J, Li L, Zhang T. Sparse online learning via truncated gradient. J Mach Learn Res 2009;10:777–801.
67. Li K. Sliced inverse regression for dimension reduction. J Amer Statist Assoc 1991;86(414):316–42.
68. Liu JS. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. J Amer Statist Assoc 1994;89(427):958–66.
69. Liu S, Quinn J, Gutmann MU, Sugiyama M. Direct learning of sparse changes in Markov networks by density ratio estimation. Neural Comput 2014;26(6):1169–97.
70. Loftsgaarden DO, QuesenBerry CP. A nonparametric estimate of a multivariate density function. Ann Math Stat 1965;36(3):1049–51.

71. Mackay DJC. Information theory, inference, and learning algorithms. Cambridge (UK): Cambridge University Press; 2003.
72. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. J Chem Phys 1953;21(6):1087–92.
73. Mosteller F, Tukey JW, editors. Data analysis and regression. Reading (MA, USA): Addison-Wesley; 1977.
74. Murphy KP. Machine learning: a probabilistic perspective. Cambridge (Massachusetts, USA): MIT Press; 2012.
75. Nguyen X, Wainwright MJ, Jordan MI. On surrogate loss functions and f-divergences. Ann Statist 2009;37(2):876–904.
76. Nguyen X, Wainwright MJ, Jordan MI. Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Trans Inform Theory 2010;56(11):5847–61.
77. Orr MJL. Introduction to radial basis function networks. Technical report, Center for Cognitive Science, University of Edinburgh, 1996.
78. Parikh N, Boyd S. Proximal algorithms. Found Trends Optim 2013;1(3):123–231.
79. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil Mag 5 1900;50(302):157–75.
80. Petersen KB, Pedersen MS. The matrix cookbook. Technical report, Technical University of Denmark, 2012.
81. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence N, editors. Dataset shift in machine learning. Cambridge (Massachusetts, USA): MIT Press; 2009.
82. Rao C. Information and the accuracy attainable in the estimation of statistical parameters. Bull Calcutta Math Soc 1945;37:81–9.
83. Ricci F, Rokach L, Shapira B, Kantor PB, editors. Recommender systems handbook. New York (NY, USA): Springer; 2010.
84. Rissanen J. Modeling by shortest data description. Automatica 1978;14(5):465–71.
85. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323:533–6.
86. Salakhutdinov RR, Hinton GE. Deep Boltzmann machines. In: van Dyk D, Welling M, editors. Proceedings of twelfth international conference on artificial intelligence and statistics. JMLR workshop and conference proceedings, vol. 5. Beach (FL, USA): Clearwater; 2009. p. 448–55.
87. Schapire RE, Freund Y. Boosting: foundations and algorithms. Cambridge (Massachusetts, USA): MIT Press; 2012.
88. Schölkopf B, Smola A, Müller K-R. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 1998;10(5):1299–319.
89. Schölkopf B, Smola AJ. Learning with kernels. Cambridge (MA, USA): MIT Press; 2002.
90. Scott DW. Multivariate density estimation: theory, practice and visualization. New York (NY, USA): Wiley; 1992.
91. Sejdinovic D, Sriperumbudur B, Gretton A, Fukumizu K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. Ann Statist 2013;41(5):2263–91.
92. Shannon C. A mathematical theory of communication. Bell Syst Tech J 1948;27:379–423.
93. Silverman BW. Density estimation for statistics and data analysis. London (UK): Chapman and Hall; 1986.
94. Sima V. Algorithms for linear-quadratic optimization. New York (NY, USA): Marcel Dekker; 1996.
95. Smith AFM, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J Roy Statist Soc Ser B 1993;55:3–24.

96. Smolensky P. Information processing in dynamical systems: foundations of harmony theory. In: Rumelhart DE, McClelland JL, editors. Parallel distributed processing: explorations in the microstructure of cognition, vol. 1. Cambridge (MA, USA): MIT Press; 1986. p. 194–281.
97. Stone M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J Roy Statist Soc Ser B 1977;39:44–7.
98. Sugiyama M. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. J Mach Learn Res 2007;8(May):1027–61.
99. Sugiyama M. Statistical reinforcement learning: modern machine learning approaches. Boca Raton (Florida, USA): Chapman and Hall, CRC; 2015.
100. Sugiyama M, Idé T, Nakajima S, Sese J. Semi-supervised local Fisher discriminant analysis for dimensionality reduction. Mach Learn 2010;78(1–2):35–61.
101. Sugiyama M, Kawanabe M. Machine learning in non-stationary environments: introduction to covariate shift adaptation. Cambridge (Massachusetts, USA): MIT Press; 2012.
102. Sugiyama M, Krauledat M, Müller K-R. Covariate shift adaptation by importance weighted cross validation. J Mach Learn Res 2007;8(May):985–1005.
103. Sugiyama M, Suzuki T, Kanamori T, du Plessis MC, Liu S, Takeuchi I. Density-difference estimation. Neural Comput 2013;25(10):2734–75.
104. Sugiyama M, Suzuki T, Nakajima S, Kashima H, von Bünau P, Kawanabe M. Direct importance estimation for covariate shift adaptation. Ann Inst Statist Math 2008;60(4):699–746.
105. Sutton RS, Barto GA. Reinforcement learning: an introduction. Cambridge (MA, USA): MIT Press; 1998.
106. Székely GJ, Rizzo ML. Energy statistics: a class of statistics based on distances. J Statist Plann Inference 2013;143(8):1249–72.
107. Takeuchi K. Distribution of information statistics and validity criteria of models. Math Sci 1976;153:12–8 [in Japanese].
108. Tangkaratt V, Sasaki H, Sugiyama M. Direct estimation of the derivative of quadratic mutual information with application in supervised dimension reduction. Technical report 1508.01019, arXiv; 2015.
109. Tax DMJ, Duin RPW. Support vector data description. Mach Learn 2004;54(1):45–66.
110. Tibshirani R. Regression shrinkage and subset selection with the lasso. J Roy Statist Soc Ser B 1996;58(1):267–88.
111. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. J Roy Statist Soc Ser B 2005;67:91–108.
112. Tomioka R, Aihara K. Classifying matrices with a spectral regularization. In: Ghahramani Z, editor. Proceedings of the 24th annual international conference on machine learning. Omnipress; 2007. p. 895–902.
113. Torkkola K. Feature extraction by non-parametric mutual information maximization. J Mach Learn Res 2003;3:1415–38.
114. Tsochantaridis I, Joachims T, Hofmann T, Altun Y. Large margin methods for structured and interdependent output variables. J Mach Learn Res 2005;6:1453–84.
115. Vapnik VN. Statistical learning theory. New York (NY, USA): Wiley; 1998.
116. Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. In: Proceedings of 25th annual international conference on machine learning; 2008. p. 1096–103.
117. von Neumann J. Various techniques used in connection with random digits. In: Householder AS, Forsythe GE, Germond HH, editors. Monte Carlo methods; National bureau of standards applied mathematics series, vol. 12. 1951. p. 36–8.

118. Wahba G. Spline models for observational data. Philadelphia (PA, USA): Society for Industrial and Applied Mathematics; 1990.
119. Watanabe S. Algebraic geometry and statistical learning theory. Cambridge (UK): Cambridge University Press; 2009.
120. Wu CFJ. On the convergence properties of the EM algorithm. Ann Statist 1983;11:95–103.
121. Yamada M, Suzuki T, Kanamori T, Hachiya H, Sugiyama M. Relative density-ratio estimation for robust distribution comparison. Neural Comput 2013;25(5):1324–70.
122. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J Roy Statist Soc Ser B 2006;68(1):49–67.
123. Zelnik-Manor L, Perona P. Self-tuning spectral clustering. In: Saul LK, Weiss Y, Bottou L, editors. Advances in neural information processing systems, vol. 17. Cambridge (MA, USA): MIT Press; 2005. p. 1601–8.
124. Zou H, Hastie T. Regularization and variable selection via the elastic net. J Roy Statist Soc Ser B 2005;67(2):301–20.