Machine Learning Bibliography (source):
http://web.eecs.umich.edu/~cscott/past_courses/eecs545f09/bib.html

Books

- Hastie, Friedman, and Tibshirani, *The Elements of Statistical Learning*, 2001
- Bishop, *Pattern Recognition and Machine Learning*, 2006
- Ripley, *Pattern Recognition and Neural Networks*, 1996
- Duda, Hart, and Stork, *Pattern Classification*, 2nd Ed., 2002
- Tan, Steinbach, and http://web.eecs.umich.edu/~cscott/past_courses/eecs545f09/bib.html, Introduction to Data Mining, Addison-Wesley, 2005.
- Scholkopf and Smola, *Learning with Kernels*, 2002
- Mardia, Kent, and Bibby, *Multivariate Analysis*, 1979
- Computational Statistics (online book)
- Sutton and Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.
- Bertsekas and Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, 1996.

Other machine learning courses

- Andrew Ng
- Max Welling

Data repositories

- UCI Machine Learning Repository
- Physionet
- MNIST Handwritten Digits
- Handwritten digits, faces, text in Matlab format
- Medical imaging
- Image registration

Background

- The Matrix Cookbook by Kaare Brandt Petersen and Michael Syskind Pedersen.
- Convex Optimization by Boyd and Vandenberghe

Matlab Software

- CVX convex program solver by Stephen Boyd
- YALMIP, a high-level Matlab interface to a variety of convex program solvers, such as SeDuMi
- SeDuMi, for solving second order cone programs. Most if not all tractable convex programs can be cast as such.
- LIBSVM, for support vector classification (including multiclass), regression, and one-class classification (novelty detection).

Conferences/Publications

- NIPS

- ICML
- AISTATS
- JMLR
- MLJ
- TPAMI

Nearest Neighbors

- The primary research area relating to nearest neighbor methods is the problem of storage, data reduction, and rapid calculation of nearest neighbors. A search on "nearest neighbor search" or "condensed nearest neighbors" or "editted nearest neighbors" will return a number of references.
- Theory: Devroye, Gyorfi and Lugosi, *A Probabilistic Theory of Pattern Recognition*, 1996

Density Estimation

- David Scott, *Multivariate Density Estimation*, 1992
- Theory: Devroye and Lugosi, *Combinatorial Methods in Density Estimation*, 2001

Linear methods for classification

- Hastie et al, Bishop, and Duda et al. all have chapters on LDA, logistic regression, and other linear classifiers.

Decision Trees

- The first comprehensive treatment and still a standard reference: Brieman, Friedman, Olshen and Stone, *Classification and Regression Trees*, 1984
- The other standard reference is Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
- A somewhat recent survey of research on decision trees: Sreerama K. Murthy: Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. Data Min. Knowl. Discov. 2(4): 345-389 (1998)
- Ripley has a nice chapter on decision trees -- probably the best place to start.

Error estimation

- On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, Gavin C. Cawley, Nicola L. C. Talbot; JMLR 11(Jul):2079-2107, 2010. Discussion of how certain model selection strategies are more biased than others; essential reading if you are doing comparative studies of different machine learning methods.
- Descent Methods for Tuning Parameter Refinement, Alexander Lorbert, Peter Ramadge ; AISTATS 2010. A natural idea.
- An entry level discussion of the bootstrap, cross-validation, and other error estimates is given in Efron and Tibshirani, *An Introduction to the Bootstrap*, 1993.

Boosting

- Adaboost was first developed in Freund and Schapire, A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119-139, 1997.
- A simpler proof the Adaboost's weak learning property is given in Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297-336, 1999.
- The view of Adaboost as performing functional gradient descent was observed by a number of researchers in the late 90's and early 00's. A representative work is L. Mason, J. Baxter, P. L. Bartlett, and M. Frean. Functional gradient techniques for combining hypotheses. In A. J. Smola, P. L. Bartlett, B. Sch?kopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 221-246. MIT Press, 2000.
- Logitboost was introduced in Friedman, J.H., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Annals of Statistics **28**, 337-407 (with discussion) (2000).
- Empirical Bernstein Boosting, Pannagadatta Shivaswamy, Tony Jebara; AISTATS 2010.
- Many other references to boosting can be found on Robert Schapire's web page.

Support Vector Machines

- The original paper: Corinna Cortes and V. Vapnik, "Support-Vector Networks", *Machine Learning*, **20**, 1995
- The standard reference: Scholkopf and Smola, *Learning with Kernels*, 2002
- Algorithms for solving the SVM are discussed E. Osuna, R. Freund, and F. Girosi. "Improved training algorithm for support vector machines." NNSP'97, 1997. http://citeseer.ist.psu.edu/osuna97improved.html, and in J. Platt, *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, in Advances in Kernel Methods - Support Vector Learning, B. Sch?kopf, C. Burges, and A. Smola, eds., MIT Press, 1999.

Clustering

- K-means, EM for Gaussian mixture models, and hierarchical clustering: see the recommended texts, especially Hastie et al., Duda et al., and Bishop (although Bishop doesn't discuss hierarchical clustering). K-means is also known as the Lloyd-Max algorithm in the context of vector quantization.
- EM was originally introduced by Dempster, A. P., Laird, N. M. Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B, 39, 1-38.
- Spectral clustering: an excellent introduction to spectral clustering is the following: U. Luxburg, ``A Tutorial on Spectral Clustering," Statistics and Computing 17(4), 395-416 (2007).

Dimensionality reduction

- Principal components analysis: The book by Mardia, Kent and Bibby derives PCA for the ``population" case (the sample case being analagous) for both the maximum

orthogonal variance perspective and the least squares linear approximation perspective. Note that PCA is also known as the Karhunen-Loeve transform (KLT).
- Multidimensional scaling: The book by Mardia, Kent and Bibby has a clean and rigorous derivation of classical MDS, associated optimality properties, and connections to PCA. It also discusses nonmetric MDS methods.
- The ``majorization" approach to metric MDS via stress minimization is reviewed and analyzed by Jan de Leeuw, "Convergence of the Majorization Method for Multidimensional Scaling," *Joumal of Classification* 5:163-180 (1988)
- Isomap
- Local linear embedding (LLE)
- Laplacian eigenmaps
- Kernel PCA is covered in the book by Scholkopf and Smola, or see the original paper referenced therein.
- Manifold learning resource page
- Self-organizing maps, principal curves, and independent component analysis (ICA) may be reviewed in Hastie et al.
- Factor analysis is treated in Mardia et al.
- An Introduction to Variable and Feature Selection, an excellent survey and introduction to methods of variable section that appeared in *Journal of Machine Learning Research* 3 (2003) 1157-1182.
- The following article describes extensive simulations for various learning algorithms combined with different feature selection methods, and offers some good intuition: Hua, J., Xiong, Z., Lowey, J., Suh, E., and E. R. Dougherty, Optimal Number of Features as a Function of Sample Size for Various Classification Rules, Bioinformatics, 21, No. 8, 1509-1515, 2005.

Nonlinear regression and Gaussian Processes

- Introduction to Gaussian Processes by David MacKay.
- Kernel ridge regression by Max Welling
- Support vector regression by Max Welling
- Approximations for Binary Gaussian Process Classification by Hannes Nickisch and Carl Edward Rasmussen.