# Statistical Machine Learning from Data
# - References -

Samy Bengio
IDIAP Research Institute
CP 592, rue du Simplon 4, 1920 Martigny, Switzerland
bengio@idiap.ch, http://www.idiap.ch/~bengio

January 17, 2006

## Statistical Learning Theory

[1] Bernhard Schölkopf and Alex J. Smola. *Learning with Kernels*. MIT Press, 2002. NOTE: A good introduction to various kernel machines.

[2] Trevor Hastie, Rob Tibshirani, and Jerome Friedman. *The elements of Statistical Learning*. Springer, 2001. NOTE: A good introduction to various machine learning models.

[3] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998. NOTE: The theory is explained here with all the equations.

[4] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer, second edition, 1995. NOTE: A good introduction to the theory, not much equations.

## Classical Methods

[5] Christopher Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, London, UK, 1995. NOTE: A good general book on machine learning.

[6] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973. NOTE: A good general book on pattern classification.

## Artificial Neural Networks

[7] Yann LeCun, Léon Bottou, G. Orr, and Klaus Muller. Efficient backprop. In G. Orr and Muller K., editors, *Neural Networks: Tricks of the trade*. Springer, 1998. NOTE: Very good paper proposing a series of tricks to make neural networks really working.

[8] Brian D. Ripley. *Pattern recognition and Neural networks.* Cambridge University Press, Cambridge, UK, 1996. NOTE: A good general book on machine learning and neural networks. Orientation: statistics.

[9] Christopher Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, London, UK, 1995. NOTE: A good general book on machine learning and neural networks. Orientation: physics.

[10] T.G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995. NOTE: The paper introducing ECOC in the machine learning literature.

[11] Simon Haykin. *Neural Networks. A Comprehensive Foundation, 2nd edition.* Macmillan College Publishing, New York, 1994. NOTE: A good general book on machine learning and neural networks. Orientation: signal processing.

[12] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994. NOTE: The extension of mixtures of experts to EM and hierarchical mixtures.

[13] Léon Bottou. *Une Approche théorique de l'Apprentissage Connexionniste: Applications à la Reconnaissance de la Parole.* PhD thesis, Université de Paris XI, Orsay, France, 1991. NOTE: Very good thesis on stochastic gradient for neural networks and speech recognition.

[14] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991. NOTE: The original paper introducing the concept of mixtures of experts.

[15] Yann LeCun. A theoretical framework for back-propagation. In D. Touretzky, G. Hinton, and T. Sejnowski, editors, *Proceedings of the 1988 Connectionist Models Summer School*, pages 21–28, CMU, Pittsburgh, Pa, 1988. Morgan Kaufmann. NOTE: A very good Lagrangian technique to derive gradients.

## Gaussian Mixture Models and EM

[16] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1–3), 2000. NOTE: How GMMs are applied to text-independent speaker verification.

[17] Jeff Bilmes. A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical Report ICSI-TR 97-021, International Computer Science Institute, 1997.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society B*, 39:1–38, 1977. NOTE: A theoretical paper introducing the EM algorithm.

## Hidden Markov Models

[19] Laurence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. NOTE: A good introduction to HMMs and speech recognition.

[20] Laurence R. Rabiner and B. H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 1986. NOTE: A very good introduction to HMMs.

## Ensemble Models

[21] Ron Meir and Gunnar Ratsch. An introduction to boosting and leveraging. In *Advanced Lectures on Machine Learning, LNCS*, pages 119–184. Springer Verlag, 2003. NOTE: Very good theoretical and practical introduction to boosting and similar algorithms.

[22] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, 1995. NOTE: A paper on boosting and AdaBoost.

[23] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1994. NOTE: The Bagging algorithm explained.

## Support Vector Machines

[24] Ronan Collobert and Samy Bengio. SVMTorch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, 1:143–160, 2001. NOTE: How to implement efficiently Support Vector Machines.

[25] Chris Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998. NOTE: A good tutorial on SVMs.

## Feature Selection

[26] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 1-2:245–272, 1997. NOTE: A broad review of various feature selection algorithms.

## Parameter Sharing

[27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. NOTE: How convolutional networks such as LeNet works.