Francis Bach (INRIA-ENS)

Sparse methods for machine learning: Theory and algorithms

Regularization by the L1-norm has attracted a lot of interest in recent years in statistics, machine learning and signal processing. In the context of least-square linear regression, the problem is usually referred to as the Lasso [1] or basis pursuit [2]. Much of the early effort has been dedicated to algorithms to solve the optimization problem efficiently, either through first-order methods [3, 4], or through homotopy methods that lead to the entire regularization path (i.e., the set of solutions for all values of the regularization parameters) at the cost of a single matrix inversion [5, 6]. A well-known property of the regularization by the L1-norm is the sparsity of the solutions, i.e., it leads to loading vectors with many zeros, and thus performs model selection on top of regularization. Recent work (e.g., [7, 8]) has looked precisely at the model consistency of the Lasso, i.e., if we know that the data were generated from a sparse loading vector, does the Lasso actually recover the sparsity pattern when the number of observations grows? Moreover, how many irrelevant variables could we consider while still being able to infer correctly the relevant ones?

The objective of the tutorial is to give a unified overview of the recent contributions of sparse convex methods to machine learning, both in terms of theory and algorithms. The course will be divided in three parts: in the first part, the focus will be on the regular L1-norm and variable selection, introducing key algorithms [3, 4, 5, 6] and key theoretical results [7, 8, 9]. Then, several more structured machine learning problems will be discussed, on vectors (second part) and matrices (third part), such as multi-task learning [10, 11], sparse principal component analysis [12], multiple kernel learning [13, 14], structured sparsity [15, 16] and sparse coding [17]. Throughout the tutorial, applications to data from various domains (computer vision, image processing, bioinformatics, speech processing, recommender systems) will be considered.

[1] R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of The
Royal Statistical Society Series B, 58(1):267–288, 1996.
[2] S. S. Chen, D L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. SIAM Review, 43(1):129–159, 2001.
[3] W. Fu. Penalized regressions: the bridge vs. the Lasso. Journal of Computational
and Graphical Statistics, 7(3):397–416, 1998).
[4] J. Friedman, T. Hastie T, and R. Tibshirani. Pathwise coordinate

optimization.
Annals of Applied Statistics, 1(2):302–332, 2007.
[5] M. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. Journal
of Computational and Graphical Statistics, 9(2):319–337, 2000.
[6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression.
Annals of Statistics, 32:407, 2004.
[7] P. Zhao and B. Yu. On model selection consistency of Lasso. Journal of Machine
Learning Research, 7:2541–2563, 2006.
[8] M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using l1-constrained quadratic programming. Technical Report 709, Department of Statistics, UC Berkeley, 2006.
[9] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. Annals of Statistics, 2008. To appear.
[10] M. Pontil, A. Argyriou, and T. Evgeniou. Multi-task feature learning. In Advances
in Neural Information Processing Systems, 2007.
[11] G. Obozinski, B. Taskar, and M.I. Jordan. Joint covariate selection and joint
subspace selection for multiple classification problems. Statistics and Computing,
pages 1–22, 2009.
[12] A. D'aspremont, El L. Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct
formulation for sparse PCA using semidefinite programming. SIAM Review, 49(3):434–48, 2007.
[13] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic
duality, and the SMO algorithm. In Proceedings of the International Conference on Machine Learning (ICML), 2004.
[14] F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning.
In Advances in Neural Information Processing Systems (NIPS), 2008.
[15] L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlaps and graph Lasso. In Proceedings of the 26th International Conference on Machine Learning (ICML), 2009.
[16] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component
analysis. Technical report, arXiv:0909.1440, 2009.
[17] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Research, 37:3311–3325, 1997.