FODAVA-Lead:

Dimension Reduction and Data Reduction: Foundations for Visualization

SPARSE RECOVERY IN MACHINE LEARNING

Vladimir Koltchinskii

School of Mathematics

Georgia Institute of Technology

vlad@math.gatech.edu

FODAVA Kick-Off Meeting

September, 2008

Research Goals

- Sparse Recovery and Feature Selection in Machine Learning
- Manifold Learning and Nonlinear Dimension Reduction

Manifold Learning

- Spectral theory of empirical graph laplacians;
- Regularized estimators of spectral characteristics of Laplace-Beltrami operator based on manifold data;
- Data-driven choice of regularization parameters;
- Penalized empirical risk minimization in manifold learning problems

Prediction Problem

(X, Y) a random couple

The risk of prediction rule f :

$$L(f) := \mathbb{E}\ell(Y; f(X)),$$

where ℓ is a loss function.

Target Function: Optimal Prediction Rule

$$f_* := \text{argmin}_{f:S\mapsto\mathbb{R}} L(f)$$

Special Cases

- regression
- large margin classification: boosting, kernel machines

Sparse Recovery Problem

$h_1, \ldots, h_N : S \mapsto \mathbb{R}$ a dictionary;
$f_\lambda := \sum_{j=1}^N \lambda_j h_j, \ \lambda \in \mathbb{R}^N$

Suppose there exists a sparse vector λ such that f_λ is a good approximation of the target function f_*.

- How to recover this sparse approximating function based on the training data?

- How to find the subset of the dictionary needed to approximate f_* ("feature selection")?

Examples of Dictionaries

- the union of several orthonormal systems used to approximate the target function f_* (Fourier basis, wavelet bases, etc);
- a set of features defined on an image;
- a set of statistical estimates of the target function to be aggregated in a more complex estimate with a better generalization performance.

Penalized Empirical Risk Minimization

$(X_1, Y_1), \ldots, (X_n, Y_n)$ training data (consists of i.i.d. random couples);

$\Lambda \subset \mathbb{R}^N$ is a convex set;

$p(\lambda)$ a convex complexity penalty;

$\varepsilon > 0$ a regularization parameter;

ℓ a convex loss function;

$L_n(f) := n^{-1} \sum_{j=1}^{n} \ell(Y_j, f(X_j))$ empirical risk.

$$\hat{\lambda}^{\varepsilon} := \operatorname{argmin}_{\lambda \in \Lambda} \left[L_n(f_\lambda) + \varepsilon p(\lambda) \right]$$

Examples of Complexity Penalties in Sparse Recovery

- LASSO: $p(\lambda) = \|\lambda\|_{\ell_1}$, $\Lambda = \mathbb{R}^N$;
- ℓ_p : $p(\lambda) = \|\lambda\|_{\ell_p}^p, p > 1$, p close to 1, $\Lambda = \mathbb{R}^N$;
- entropy: $p(\lambda) = \sum_{j=1}^N \lambda_j \log \lambda_j$,

$$\Lambda := \left\{ \lambda \in \mathbb{R}^N : \lambda_j \geq 0, \sum_{j=1}^N \lambda_j = 1 \right\}$$

Typical Mathematical Results

• sparsity inequalities show that in "sparse" problems the empirical solution $\hat{\lambda}^{\varepsilon}$ is "approximately sparse" with a high probability and its "sparsity pattern" mimicks the sparsity pattern of "sparse oracles";

• oracle inequalities show that, with a high probability, the empirical solution $\hat{\lambda}^{\varepsilon}$ provides the same approximation of the target function f_* as optimal "sparse oracles" up to an error term that depends on the degree of sparsity of the problem.

Oracle Inequalities

Bunea, Tsybakov and Wegkamp (2007), Koltchinskii (2008), van de Geer (2008)

With a high probability,

$$L(f_{\hat{\lambda}^\varepsilon}) - L(f_*) \leq C \inf_{\lambda \in \Lambda} \left[L(f_\lambda) - L(f_*) + B(\lambda) d(\lambda) \varepsilon^2 \right],$$

where

$$d(\lambda) = \text{card}(\text{supp}(\lambda))$$

and $B(\lambda)$ is a quantity that characterizes geometric properties of the dictionary.

Other Methods of Sparse Recovery

- Dantzig Selector (Candes and Tao (2007))
- Greedy Approximation

More General Models of Sparse Recovery

- sparse additive models;
- sparse recovery in large ensembles of kernel machines (Koltchinskii and Yuan (2008));
- sparse recovery in linear spans or convex hulls ("sparse mixtures") of infinite dictionaries (Koltchinskii and Minsker, in progress)