Machine Learning

Lecture 3.1: Sparsity in Convex Optimization for Supervised Machine Learning

Marc Sebban and Amaury Habrard

Laboratoire Hubert Curien, UMR CNRS 5516
University of Jean Monnet Saint-Étienne (France)

Some materials used for this lecture:

- F. Bach "*Sparse methods for Machine Learning - Theory and Algorithms*" - Tutorial at NIPS'2009 and at ECML'2010.
- Y. Grandvalet "*Sparsity in Learning* - Tutorial at CAP'2013.
- G. Obozinski "*Sparse Methods in Statistical Learning Theory*" - 2010.
- F. Bach "*Learning with sparsity-inducing norms*" - MLSS 2008.

Outline

Regularized (penalized) supervised learning problem

Training data: a set of $S = \{z_i = (x_i, y_i)\}_{i=1}^{m}$ of m training data i.i.d. from an unknown joint distribution $\mathcal{D}_{\mathcal{Z}}$ over a space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

$$\min_h \sum_{i=1}^{m} \ell(y_i, h(x_i)) + \lambda ||h||$$

where $\lambda||h||$ is a regularization term which prevents the algorithm from overfitting.

The previous **penalized problem** can be rewritten as a **constrained problem**.

$$\min_h \sum_{i=1}^{m} \ell(y_i, h(x_i)) \text{ s.t. } ||h|| < c$$

Indeed, for any c in the constrained setting, there is a corresponding λ for which one can penalize the objective function.

Sparsity: a parsimonious use of data

What about sparsity?

We consider the set S composed of m examples in $\mathbb{R}^d$:

$$S = \begin{pmatrix} x_1 \\ . \\ . \\ . \\ x_i \\ . \\ . \\ . \\ x_m \end{pmatrix} = \begin{pmatrix} x^1 & \dots & x^j & \dots & x^d \end{pmatrix}$$

This set can be reduced:

- in columns $\Rightarrow$ deletion of features - useful when d is large compared to n.
- in rows $\Rightarrow$ deletion of examples (e.g. ℓ_1-SVM, CNN).
- in rank (e.g. PCA, LSA) $\Rightarrow$ Find the embedding space.

Why ignoring some variables?

- **Prevent from overfitting** - curse of dimensionality - Occam's razor principle.
- **Computational efficiency**
  - Fast evaluation at test time.
- **Interpretability**
  - Understanding the underlying phenomenon.

Three categories of methods

1. **Filter** approach
   - Variables "filtered" by a criterion (e.g. Fisher, Wilks, Mutual Information).
   - Learning proceeds after the treatment.
2. **Wrapper** approach
   - Heuristic search of subsets of variables.
   - Subset selection is done w.r.t. the learning algorithm performance.
3. **Embedded** approach: use of **sparsity-inducing norms**
   - Feature selection is part of the learning algorithm
   - All features processed during learning, only some influence the solution.
   - Example: LASSO in Linear Regression

$$\min_{\theta} \frac{1}{2m} \sum_{i=1}^{m} (y_i - \theta^T x_i)^2 + \lambda ||\theta||_1$$

Note that we would prefer to use directly the ℓ_0-norm to induce sparsity.

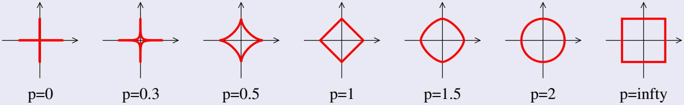
"Hard" subset selection with the ℓ_0-norm and Relaxation

ℓ_0 Norms in Linear Models

$$h(x) = \theta^T x$$

$$\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, h(x_i)) + \lambda ||\theta||_0$$

NP-hard problem



Relaxation

$$\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, h(x_i)) + \lambda ||\theta||_p$$

$$\left.\begin{array}{r} \text{Convex relaxation (if } \ell \text{ convex) for } p \geq 1 \\ \text{Sparse solution for } 0 < p \leq 1 \end{array}\right\} \Rightarrow \ell_1\text{-norm is a good trade-off}$$

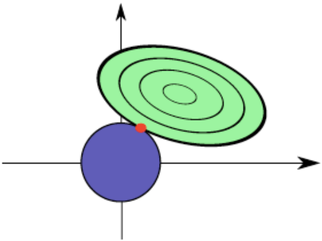
ℓ_2-norms

Regularization with ℓ_2-norm does not mean sparsity

Example: Let us consider the following problem:

$$\min_{\theta\in\mathbb{R}^d} \frac{1}{2}\theta^T\theta - \theta^T\mathbf{x} + \lambda||\theta||_2^2.$$

$$\frac{\partial \frac{1}{2}\theta^T\theta - \theta^T\mathbf{x} + \lambda||\theta||_2^2}{\partial\theta_j} = 0 \Leftrightarrow \forall j = 1..d,\ \theta_j{}^* = \frac{x_j}{1+2\lambda}$$



The ℓ_2 norm penalizes the larger components first

The gradient is linear in the magnitude of each component of the vector (indeed, $\frac{\partial x^2}{\partial x} = 2x$). Thus, small values are favored, but **it's more favorable to decrease a large value than a small one**.

Example

- The ℓ_2 norm of $\theta = (1, 3)$ is $||\theta||_2 = \sqrt{10}$.
- Decreasing the first component by 1 results in a vector $\theta = (0, 3)$ with $||\theta||_2 = \sqrt{9} = 3$.
- But decreasing the second component by 1 results in $\theta = (1, 2)$ with $||\theta||_2 = \sqrt{5} < 3$.
- Thus, its more favorable to decrease the larger components of the vector to minimize the norm of w. The ℓ_2 regularization is also called "**weight decay**"

(see demo)

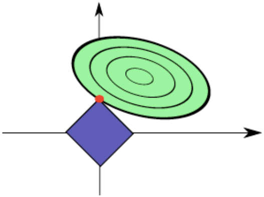
ℓ_1-norm Regularization

ℓ_1-norm Regularization performs **regularization** as well as feature **selection**.

ℓ_1-norm regularization results in sparse models

$$\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, h(x_i)) + \lambda ||\theta||_1$$

Increasing λ will cause more and more of the parameters of θ to be **driven to zero**.



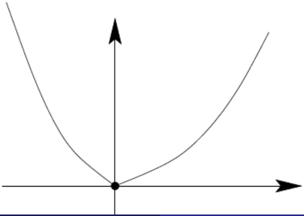
Impact of λ on the sparsity: an example

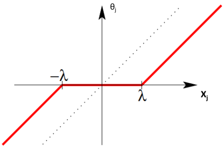
$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2}\theta^T\theta - \theta^T\mathbf{x} + \lambda||\theta||_1.$$

- If $\lambda = 0$ (i.e. no penalization) the zero gradient gives:

$$\theta^* = \mathbf{x} \Rightarrow \theta_j = 0 \text{ (i.e. sparsity) if and } \textbf{only if } \mathbf{x}_j = 0.$$

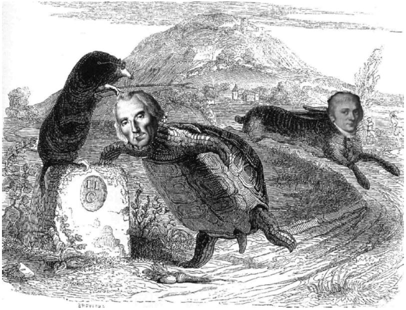
- If $\lambda \neq 0$, let us consider the partial derivative at $\theta_j = 0^+ : g^j_+ = \lambda - \mathbf{x}_j$ and at $\theta_j = 0^- : g^j_- = -\lambda - \mathbf{x}_j$. The solution is
  - ${\theta_j}^* = 0$ iff $g^j_+ \geq 0$ and $g^j_- \leq 0$.
  - So if $|\mathbf{x}_j| \leq \lambda$, the set of situations inducing sparsity is expanded!





Optimization methods

ℓ_2 versus ℓ_1 - Gaussian hare versus Laplacian tortoise



ℓ_1 is cool but ... which one is faster? ℓ_1 or ℓ_2?

- Gauss is in favor of ℓ_2 while Laplace is in favor of ℓ_1.
- Since ℓ_1 is not differentiable, it might look that it is harder to optimize. This is the tortoise.
- Since ℓ_2 usually leads to nice smooth convex optimization problem, it is supposed to be easier. This is the hare.

Optimization Methods

- $\min_{\mathbf{w}} \frac{1}{2m} \sum_{i=1}^{m} (y_i - \mathbf{w}^T x_i)^2 + \lambda \sum_{j=1}^{d} (\mathbf{w}_j{}^+ + \mathbf{w}_j{}^-)$ such that

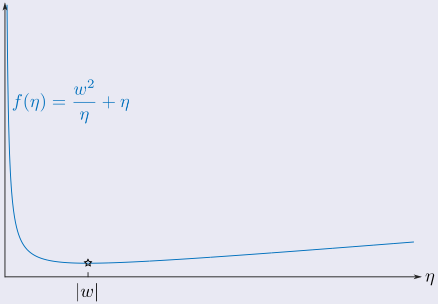
$$\mathbf{w} = \mathbf{w}_j{}^+ - \mathbf{w}_j{}^-$$

$$\mathbf{w}_j{}^+ \geq 0 \text{ and } \mathbf{w}_j{}^- \geq 0$$

  $\Rightarrow$ very slow.

- **Generic methods**: Interior points.
- **Active set methods**: LARS algorithm.

- η-**trick** (Micchelli and Pontil, 2006; Rakotomamonjy et al. 2008)
  - Notice that $||\mathbf{w}||_1 = \sum_{j=1}^{d} |\mathbf{w}_j| = \min_{\eta \geq 0} \frac{1}{2} \sum_{j=1}^{d} \left\{ \frac{\mathbf{w}_j{}^2}{\eta_j} + \eta_j \right\}$
  - Alternating minimization w.r.t. η (close-form) and $\mathbf{w}$ (weighted squared ℓ_2-norm regularized problem).

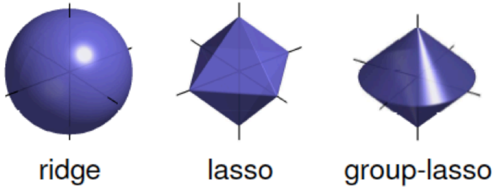


Group Sparsity in Linear Regression

Ball Crafting

Group Sparsity in Linear Regression

How to remove groups of features?
$\rightarrow$ based on the assumption that a group structure is known.



Group Lasso

$$\ell_1/\ell_2\text{-norm} = \sum_{g \in \mathcal{G}} ||\mathbf{w}_g||_2 = \sum_{g \in \mathcal{G}} \left(\sum_{j \in g} \mathbf{w}_j^2 \right)^{\frac{1}{2}}$$

where $\{\mathcal{G}_k\}_{k=1}^K$ forms a partition of $\{1, \ldots, d\}$.

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \mathbf{w}^T x_i) + \lambda \sum_{g \in \mathcal{G}} ||\mathbf{w}_g||_2$$

Sparse solution groupwise

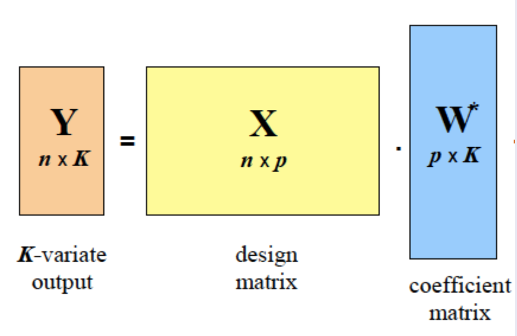
- Proximal methods.
- Blockwise coordinate descent.

Sparse Methods for Matrices

Learning on Matrices (1/2)

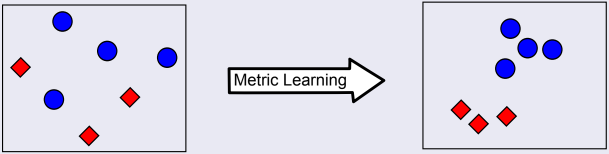
Multivariate Linear Regression

$$\min_{\mathbf{W}} \frac{1}{2}||Y - \mathbf{XW}||_{\mathcal{F}}^2 + \lambda||\mathbf{W}||_{\mathcal{F}}^2$$



Learning on Matrices (2/2)

Metric Learning



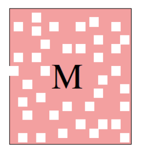
Mahalanobis Distance Learning: $\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$,

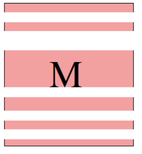
$$d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}') = \sqrt{(\mathbf{x} - \mathbf{x}')^T \mathbf{M} (\mathbf{x} - \mathbf{x}')},$$

where $\mathbf{M} \in \mathbb{R}^{d \times d}$ is a symmetric PSD matrix ($\mathbf{M} \succeq 0$).

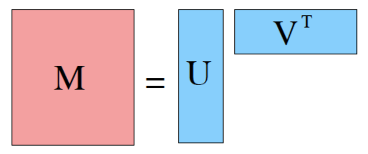
Two Types of Sparsity for matrices

- Directly on the elements of $\mathbf{M}$ using the ℓ_1-norm or the ℓ_1/ℓ_2-norm.





- Through a factorization of $\mathbf{M} = \mathbf{U}\mathbf{V}^T$ with low rank (k small), where $\mathbf{U} \in \mathbb{R}^{m \times k}$ and $\mathbf{V} \in \mathbb{R}^{d \times k}$ and $\mathbf{M} \in \mathbb{R}^{m \times d}$.



Rank constrained learning

Rank constrained learning

Given a matrix $\mathbf{M} \in \mathbb{R}^{m \times d}$

- $Rank(\mathbf{M}) = ||s||_0$ (non convex function) where $s \in \mathbb{R}^m_+$ are singular values.
- The rank of $\mathbf{M}$ is the minimum size m of all factorizations of $\mathbf{M}$ into $\mathbf{M} = \mathbf{U}\mathbf{V}^T$, $\mathbf{U} \in \mathbb{R}^{m \times k}$ and $\mathbf{V} \in \mathbb{R}^{d \times k}$

$$\min_{\mathbf{M} \in \mathbb{R}^{m \times p}} \ell(\mathbf{M}) \text{ s.t. } rank(\mathbf{M}) \leq m.$$

In general, NP-Hard

Solution: Convex Relaxation

Replace (relax) the rank objective function by a convex norm.

Trace Norm also known as Nuclear Norm or ... Ky-Fan-n-norm

Trace Norm $||\mathbf{M}||_{tr}$

- $Rank(\mathbf{M}) = ||s||_0 \xrightarrow{\text{relax}} ||s||_1 = ||\mathbf{M}||_{tr}$
- **Relaxation of the problem**:

$$\min_{\mathbf{M}\in\mathbb{R}^{m\times p}} \ell(\mathbf{M}) + \lambda||\mathbf{M}||_{tr}.$$

$\rightarrow$ Leads to convex optimization problems.

$\rightarrow$ **Algorithms**

- Proximal methods.
- Iterated reweighted Least-Square (Argyriou et al., 2009)
- Common bottleneck: requires iterative SVD.

Other convex relaxation

- **Log-det heuristic** [Fazel et al. 2003]
  - Uses the logarithm of the determinant as a smooth approximation for rank.

$$\min_{\mathbf{M}\succeq 0} \ell(\mathbf{M}) + \mathit{logdet}(\mathbf{M} + \lambda I).$$

  - **Interpretation**: the logdet corresponds to the log of the volume of an ellipsoid as the product of the eigenvalues of $\mathbf{M}$.