Advanced Topics in Machine Learning (Part II)

4. Sparsity Methods

Massimiliano Pontil

Today's Plan

- Sparsity in linear regression
- Formulation as a convex program – Lasso
- Group Lasso
- Matrix estimation problems (Collaborative Filtering, Multi-task Learning, Inverse Covariance, Sparse Coding, etc.)
- Structure Sparsity
- Dictionary Learning / Sparse Coding
- Nonlinear extension

L1-regularization

Least absolute shrinkage and selection operator (LASSO):

$$\min_{\|w\|_1 \le \alpha} \frac{1}{2}\|y - Xw\|_2^2$$

where $\|w\|_1 = \sum_{j=1}^{d} |w_j|$

- equivalent problem: $\min\limits_{w \in \mathbb{R}^d} \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1$

- can be rewritten as a QP:

$$\min_{w^+, w^- \ge 0} \frac{1}{2}\|y - X(w^+ - w^-)\|_2^2 + \lambda e^\top (w^+ + w^-)$$

L1-norm regularization encourages sparsity

Consider the case $X = I$:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w - y\|_2^2 + \lambda \|w\|_1$$

Lemma: Let $H_\lambda(t) = (|t| - \lambda)_+ \text{sgn}(t),\ t \in \mathbb{R}$. The solution $\hat{w}$ is given by

$$\hat{w}_i = H_\lambda(y_i), \quad i = 1, \ldots, d$$

Proof: First note that the problem decouples: $\hat{w}_i = \text{argmin}\left\{\frac{1}{2}(w_i - y_i)^2 + \lambda |w_i|\right\}$. By symmetry $\hat{w}_i y_i \geq 0$, thus w.l.o.g. we can assume $y_i \geq 0$. Now, if $\hat{w}_i > 0$ the objective function is differentiable and setting the derivative to zero gives $\hat{w}_i = y_i - \lambda$. Since the minimum is unique we conclude that $\hat{w}_i = (y_i - \lambda)_+$.

Minimal norm interpolation

If the linear system $Xw = y$ of equations admits a solution, when $\lambda \to 0$ the L1-regularization problem reduces to:

$$\min \{\|w\|_1 : Xw = y\} \qquad \text{(MNI)}$$

which is a linear program (exercise)

- the solution is in general not unique

- suppose that the $y = Xw^*$; under which condition w^* is also the unique solution to (MNI)?

Restricted isometry property

Without further assumptions there is no hope that $\hat{w} = w^*$

The following condition are sufficient:

- Sparsity: $\mathrm{card}\{j : |w_j^*| \neq 0\} \leq s$, with $s \ll d$

- X satisfies the restricted isometry property (RIP): there is a $\delta_s \in (0,1)$ such that, for every $w \in \mathbb{R}^d$ with $\mathrm{card}\{j : w_j \neq 0\} \leq s$, it holds that

$$(1-\delta_s)\|w\|_2^2 \leq \|Xw\|_2^2 \leq (1+\delta_s)\|w\|_2^2$$

Optimality conditions

Directional derivative of a function $f : \mathbb{R}^d \to \mathbb{R}$ at w in the direction d:

$$D^+ f(w; d) := \lim_{\epsilon \to 0^+} \frac{f(w + \epsilon d) - f(w)}{\epsilon}$$

- when f is convex, the limit is always well defined and finite

Theorem 1: $\hat{w} \in \arg \min_{w \in \mathbb{R}^d} f(w)$ iff $D^+ f(\hat{w}; d) \geq 0 \ \forall d \in \mathbb{R}^d$

- if f is differentiable at w then $D^+ f(w; d) = d^\top \nabla f(w)$ and Theorem 1 says that $\hat{w}$ is a solution iff $\nabla f(\hat{w}) = 0$

Optimality conditions (cont.)

If f is convex its subdifferential at w is defined as

$$\partial f(w) = \{u : f(v) \geq f(w) + u^\top (v - w),\ \forall v \in \mathbb{R}^d\}$$

- a set-valued function!
- always a closed convex set
- the elements of $\partial f(w)$ are called the subgradients of f at w
- intuition: $u \in \partial f(w)$ if the affine function $f(w) + u^\top (v - w)$ is a global underestimator of f

Theorem 2: $\hat{w} \in \arg \min_{w \in \mathbb{R}^d} f(w)$, iff $0 \in \partial f(\hat{w})$ (easy to proof)

Optimality conditions (cont.)

Theorem 2: $\hat{w} \in \arg\min_{w \in \mathbb{R}^d} f(w)$, iff $0 \in \partial f(\hat{w})$

- if f is differentiable then $\partial f(w) = \{\nabla f(w)\}$ and Theorem 2 says that $\hat{w}$ is a solution iff $\nabla f(\hat{w}) = 0$

Some properties of gradients are still true for subgradients, e.g:

- $\partial(af)(w) = af(w)$, for all $a \geq 0$
- If f and g are convex then $\partial(f+g)(w) = \partial f(w) + \partial g(w)$

Optimality conditions for Lasso

$$\min \|y - Xw\|_2^2 + \lambda \|w\|_1$$

- by Theorem 2 and the properties of subgradients, w is a optimal solution iff
$$X^\top (y - Xw) \in \lambda \partial \|w\|_1$$

- to compute $\partial \|w\|_1$ use the sum rule and the subgradient of the absolute value: $\partial |t| = \{\mathrm{sgn}(t)\}$ if $t \neq 0$ and $\partial |t| = \{u : |u| \leq 1\}$ if $t = 0$

Case $X = I$: $\hat{w}$ is a solution iff, for every $i = 1, \ldots, d$, $y_i - \hat{w}_i = \lambda \mathrm{sgn}(\hat{w}_i)$ if $\hat{w}_i \neq 0$ and $|y_i - \hat{w}_i| \leq \lambda$ otherwise (verify that these formulae yield the soft thresolding solution on page 4)

General learning method

In generally we will consider optimization problems of the form

$$\min_{w \in \mathbb{R}^d} F(w), \quad \text{where } F(w) = f(w) + g(w)$$

Often f will be a data term: $f(w) = \sum_{i=1}^{m} E(w^\top x_i, y_i)$, and g a convex penalty function (non necessarily smooth, e.g. the L1-norm)

Next week we will discuss a general and efficient method to solve the above problem under the assumptions that f has some smoothness property and g is "simple", in the sense that the following problem is easy to solve

$$\min_{w} \frac{1}{2}\|w - y\|^2 + g(w)$$

Group Lasso

Enforce sparsity across a-priori known groups of variables:

$$\min_{W \in \mathbb{R}^d} f(w) + \lambda \sum_{\ell=1}^{N} \|w_{|J_\ell}\|_2$$

where $J_1, \dots, J_N$ are prescribed subsets of $\{1, \dots, d\}$

- In the original formulation (Yuan and Lin, 2006) the groups form a partition of the index set $\{1, \dots, n\}$

- Overlapping groups (Zhao et al. 2009; Jennaton et al. 2010): hierarchical structures such as DAGS

  Example: $J_1 = \{1, 2, \dots, d\}, J_2 = \{2, 3, \dots, n\}, \dots, J_n = \{n\}$

Multi-task learning

- Learning multiple linear regression or binary classification tasks simultaneously
- Formulate as a matrix estimation problem ($W = [w_1, \ldots, w_T]$)

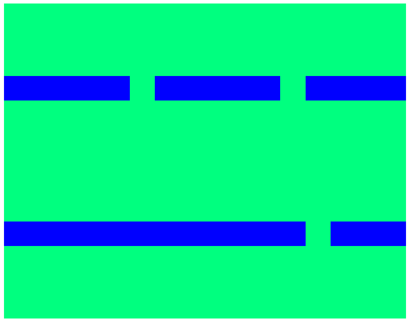
$$\min_{W \in \mathbb{R}^{d \times T}} \sum_{t=1}^{T} \sum_{i=1}^{m} E(w^\top x_{ti}, y_{ti}) + \lambda g(W)$$

- Relationships between tasks modeled via sparsity constraints on W
- Few common important variables (special case of Group Lasso):

$$g(W) = \sum_{j=1}^{d} \|w^j\|_2$$

Structured Sparsity

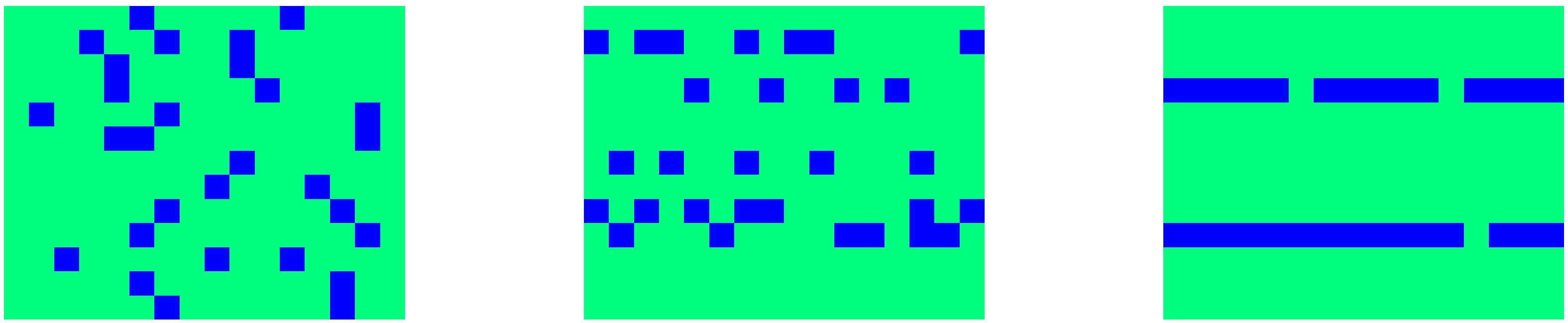
- The above regularizer favors matrices with many zero rows (few features shared by the tasks)



$$g(W) = \sum_{j=1}^{d} \sqrt{\sum_{t=1}^{T} w_{tj}^2}$$

2. Structured Sparsity (cont.)

Compare matrices W favored by different norms $(\text{green} = 0,\ \text{blue} = 1)$:



$\#\text{rows} =$	13	5	3
$g(W) =$	19	12	8
$\sum_{tj} \lvert w_{tj} \rvert =$	29	29	29

Estimation of a low rank matrix

$$\min_{W \in \mathbb{R}^{d \times T}} \left\{ \sum_{i=1}^{m} (y_i - \langle W, X_i \rangle)^2 : \text{rank}(W) \leq k \right\}$$

- Multi-task learning: choose $X_i = x_i e_{c_i}^\top$, hence $\langle W, X_i \rangle = w_{c_i}^\top x_i$
- Collaborative filtering: choose $X_i = e_{r_i} e_{c_i}^\top$, hence $\langle W, X_i \rangle = W_{r_i c_i}$, where $r_i \in \{1, \ldots, d\}$ and $c_i \in \{1, \ldots, T\}$ (rows / columns indices)

Relax the rank with the trace (or nuclear) norm: $\|W\|_* = \sum_{i=1}^{\min(d,T)} \sigma_i(W)$

Trace norm regularization

$$\min_{W \in \mathbb{R}^{d \times T}} \sum_{i=1}^{m} (y_i - \langle W, X_i \rangle)^2 + \lambda \|W\|_*$$

- complete data case: $\displaystyle \min_{W \in \mathbb{R}^{d \times T}} \|Y - W\|_{\mathrm{Fr}}^2 + \lambda \|W\|_*$
- if $Y = U \mathrm{diag}(\sigma) V^\top$ then the solution is (recall H_λ from page 4):

$$\hat{W} = U \mathrm{diag}(H_\lambda(\sigma)) V^\top$$

Proof uses *von Neumann's Theorem:* $\mathrm{tr}(Y^\top W) \leq \sigma(Y)^\top \sigma(W)$ and equality holds iff Y and W have the same ordered system of singular vectors

Sparse Inverse Covariance Selection

Let $x_1, \ldots, x_m \sim p$, where $p(x) = \frac{1}{(2\pi)^d \det(\Sigma)} e^{-(x-\mu)^\top \Sigma^{-1}(x-\mu)}$

Maximum likelihood estimate for the covariance

$$\begin{aligned}
\hat{\Sigma} &= \arg\max_{\Sigma \succ 0} \prod_{i=1}^{d} p(x_i) = \arg\max_{\Sigma \succ 0} \prod_{i=1}^{d} \log p(x_i) \\
&= \arg\max_{\Sigma \succ 0} \left\{ -\log\det(\Sigma) - \langle S, \Sigma^{-1} \rangle \right\}
\end{aligned}$$

where $S = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)(x_i - \mu)^\top$

- The solution is $\hat{\Sigma} = S$ (show it as an exercise)

Sparse Inverse Covariance Selection (cont.)

Inverse covariance provides information about the relationship between variables: $\Sigma_{ij}^{-1} = 0$ iff x^i and x^j are conditionally independent

$$\hat{W} = \arg\max_{W \succ 0} \{\log\det(W) - \langle S, W\rangle\} = \arg\min_{W \succ 0} \{\langle S, W\rangle - \log\det(W)\}$$

If we expect many pairs of variables to be conditionally independent we could solve the problem

$$\min\{\langle S, W\rangle - \log\det(W) : W \succ 0,\ \operatorname{card}\{(i,j) : |W_{ij}| > 0\} \le k\}$$

which can be relaxed to the convex program

$$\min\{\langle S, W\rangle - \log\det(W) : W \succ 0,\ \|W\|_1 \le k\}$$

Dictionary Learning / Sparse Coding

Given $x_1, \ldots, x_m \sim p$ find $d \times k$ matrix W which minimize the average reconstruction error:

$$\sum_{i=1}^{m} \min_{z \in Z} \|x_i - Wz\|_2^2$$

Can be seen as a constrained matrix factorization problem

$$\min \left\{ \|X - WZ\|_{\mathrm{F}}^2 : W \in \mathcal{W}, Z \in \mathcal{Z} \right\}$$

where $X = [x_1, \ldots, x_m]$ and $\mathcal{W} \subseteq \mathbb{R}^{d \times k}$, $\mathcal{Z} \subseteq \mathbb{R}^{k \times m}$

Interpretation: the columns of W are some basis vectors (could be linearly dependent) and the columns of Z are the codes / coefficients used to reconstruct the inputs as a linear combination of the basis vectors

Examples

- PCA: $\mathcal{W} = \mathbb{R}^{d \times k}$, $\mathcal{Z} = \mathbb{R}^{k \times m}$
- k-means clustering: $\mathcal{W} = \mathbb{R}^{d \times k}$, $\mathcal{Z} = \{Z : z_i \in \{e_1, \ldots, e_k\}\}$
- Nonnegative matrix factorization

$$\min_{W,Z \geq 0} \|X - WZ\|_{\mathrm{F}}^2$$

- Sparse coding: $\mathcal{W} = \mathbb{R}^{d \times k}$, $\mathcal{Z} = \{Z : \|z_i\|_0 \leq s\}$

  Can be relaxed to the problem: $\min \|X - WZ\|_{\mathrm{Fr}}^2 + \lambda \|Z\|_1$

Nonlinear extension

The methods we have seen so far can be extended to a RKHS setting; for example the Lasso extends to the problem

$$\min \sum_{i=1}^{m} E\left(\sum_{\ell=1}^{N} f_\ell(x_i), y_i\right) + \lambda \sum_{\ell=1}^{N} \|f_\ell\|_{K_\ell} \quad (*)$$

- minimum is over functions $f_1, \dots, f_N$, with $f_\ell \in H_{K_\ell}$, with $K_1, \dots, K_N$ some prescribed kernels

- feature space formulation (recall $K_\ell(x,t) = \langle \phi_\ell(x), \phi_\ell(t) \rangle$)

$$\min \sum_{i=1}^{m} E\left(\sum_{\ell=1}^{N} w_\ell^\top \phi_\ell(x_i), y_i\right) + \lambda \sum_{\ell=1}^{N} \|w_\ell\|_2$$

Connection to Group Lasso

Two important "parametric" versions of the above formulation:

- **Lasso:** choose $f_j(x) = w_j x_j$, $K_j(x,t) = x_j t_j$

$$\sum_{i=1}^{m} E(w^\top x_i, y_i) + \gamma \sum_{j=1}^{d} |w_j|$$

- **Group Lasso**: choose $f_j(x) = \sum_{j \in J_\ell} w_j x_j$, $K_j(x,t) = \langle x_{|J_\ell}, t_{|J_\ell} \rangle$, where $\{J_\ell\}_{\ell=1}^{n}$ is a partition of index set $\{1, \ldots, d\}$

$$\sum_{i=1}^{m} E(w^\top x_i, y_i) + \gamma \sum_{\ell=1}^{N} \|w_{|J_\ell}\|_2$$

Representer theorem

Two reformulations of (*) as a finite dimension optimization problem

- Using the representer theorem:

$$\min \sum_{i=1}^{m} E\left(\sum_{\ell=1}^{N}\sum_{j=1}^{m} K_\ell(x_i, x_j)\alpha_{\ell,j}, y_i\right) + \lambda \sum_{\ell=1}^{N} \sqrt{\alpha_\ell^\top K_\ell \alpha_\ell}$$

- Using the formula $\sum_\ell |t_\ell| = \inf\limits_{z>0} \frac{1}{2}\sum_\ell \frac{t_\ell^2}{z_\ell} + z_\ell$, rewrite the problem as

$$\inf_{z>0} \min \sum_{i=1}^{m} E(f(x_i), y_i) + \frac{\lambda}{2}\|f\|^2_{\sum\limits_\ell z_\ell K_\ell} + \sum_\ell z_\ell$$

Some references

- **L1-regularization / L1-MNI:**

  - P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.
  - E. J. Candès. The restricted isometry property and its implications for compressed sensing. Compte Rendus de l'Academie des Sciences, Paris, Serie I, 346 589-592.
  - E. J. Candès, J. Romberg and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. Comm. Pure Appl. Math., 59 1207-1223.
  - R. Tibshirani. Regression Shrinkage and Selection via the Lasso, *J. Royal Statistical Society B*, 58(1):267–288, 1996.

- **Group Lasso:**

  - M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables *Journal of the Royal Statistical Society, Series B*, 68(1):49–67, 2006.
  - P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.
  - R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. arXiv:0904.3523v2, 2009.

- **Multi-task learning:**
  - A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.
  - G. Obozinski, B. Taskar, and M.I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):1–22, 2010.
- **Low rank matrix estimation:**
  - V. Koltchinskii, A.B. Tsybakov, K. Lounici. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *arXiv:1011.6256*, 2011.
  - N. Srebro, J.D.M. Rennie, T.S. Jaakkola. Maximum-Margin Matrix Factorization. *Advances in Neural Information Processing Systems 17*, pages 1329–1336, 2005.
  - E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. of Comput. Math.*, 9 717-772.
- **Nonlinear Group Lasso / Multiple kernel learning:**
  - A. Argyriou, C. A. Micchelli and M. Pontil. Learning convex combinations of continuously parameterized basic kernels. COLT 2005
  - F. R. Bach, G. R. G. Lanckriet and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. ICML 2004

- G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui and M. I. Jordan. Learning the kernel matrix with semidefinite programming. JMLR 2004
- C.A. Micchelli and M. Pontil. Learning the kernel function via regularization. JMLR 2005
- A. Rakotomamonjy, F. R. Bach, S. Canu, Y. Grandvalet, SimpleMKL, JMLR, 2008.

- **Sparse Coding:**

  - B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.
  - D. Lee and H. Seung, Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13, pages 556-562, 2001.