Regularized Linear Regression

Friday, January 21, 2022 9:26 AM

Recall the standard linear model:

$$y = X\theta^* + w \quad ; \quad y, w \in \mathbb{R}^n \quad ; \quad X \in \mathbb{R}^{n \times d} \quad ; \quad \theta^* \in \mathbb{R}^d$$

Previously, we assumed $n \gg d$ and key to deriving the estimation error is

$$\underbrace{\|X\Delta\|_2^2}_{\hat{\Delta} = \hat{\theta} - \theta^*} \geq \underbrace{\lambda_{min}(X^T X)}_{> 0 \text{ when } X \text{ has full col rank}} \cdot \|\Delta\|_2^2 \qquad \forall \Delta \in \mathbb{R}^d$$

Consider the case where $d \gg n$.

Typically, we assume θ^* has some low-dimensional structure. (For example, sparsity, group sparsity, low-rankness, etc.) Then, a natural family of estimators is

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \mathcal{L}(\theta; \{z_i\}_{i=1}^n) + \lambda R(\theta) \right\}$$

where $z_1, \ldots, z_n$ are observations from a space $\mathcal{Z}$;

$\mathcal{L} : \mathbb{R}^d \times \mathcal{Z}^n \to \mathbb{R}$ is a smooth convex loss function;

$R : \mathbb{R}^d \to \mathbb{R}_+$ is a norm regularizer that aims to induce the desired low-dimensional structure;

$\lambda > 0$ is a regularization parameter.

e.g.: In standard linear model:

$$z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} \triangleq \mathcal{Z}$$

where $x_i^T = i^{th}$ row of X;

$$\mathcal{L}(\theta; \{z_i\}_{i=1}^n) = \frac{1}{2} \|y - X\theta\|_2^2 ;$$

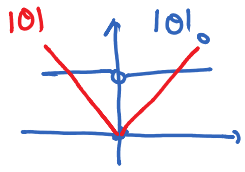
Under the assumption that θ^* is sparse,

$$R(\theta) = \|\theta\|_1$$

<u>Remark:</u> A more direct measure of sparsity is

$$\|\theta\|_0 = \# \text{ of non-zero entries in } \theta$$

However, this is not continuous:



One uses a convex surrogate to replace it. $\|\theta\|_1$ is a typical choice.

Decomposability of the Regularizer

Let $\mathcal{M} \subseteq \bar{\mathcal{M}} \subseteq \mathbb{R}^d$ be a pair of subspaces

- $\mathcal{M}$ is the model subspace and intends to capture the constraints specified by the model
- $\bar{\mathcal{M}}^\perp \triangleq \{ v \in \mathbb{R}^d : u^T v = 0 \quad \forall u \in \bar{\mathcal{M}} \}$ is the perturbation space and captures the deviations from $\mathcal{M}$.
- For simplicity, one can take $\mathcal{M} = \bar{\mathcal{M}}$, but the case where $\mathcal{M} \subsetneq \bar{\mathcal{M}}$ can be useful in some settings.

Definition: The regularizer R is <u>decomposable</u> wrt $(\mathcal{M}, \bar{\mathcal{M}}^{\perp})$

if $\quad R(\theta+\gamma) = R(\theta) + R(\gamma) \quad \forall \theta \in \mathcal{M},\ \gamma \in \bar{\mathcal{M}}^{\perp}$.

Q: How should we choose $\mathcal{M}, \bar{\mathcal{M}}$?

In general, there are many choices; e.g., $\mathcal{M} = \mathbb{R}^d$, $\bar{\mathcal{M}}^{\perp} = \{0\}$.

This is not desirable because $\mathcal{M}$ is too "big".

To have an effective decomposition, we want

(i) $\theta^*_{\mathcal{M}} \triangleq \Pi_{\mathcal{M}}(\theta^*)$ should approximately equal θ^*.

(ii) $\mathcal{M}$ should be relatively "small".

Example 1: Sparse vectors and ℓ_1-regularization

- model: s-sparse vectors in $\mathbb{R}^d$; $R(\theta) = \|\theta\|_1$
- model subspace: For any $S \subseteq \{1, \cdots, d\}$ of cardinality s,

  define $\mathcal{M}(S) = \{\theta \in \mathbb{R}^d : \theta_j = 0 \ \forall j \notin S\}$

  Note: If θ^* is supported on S, then $\Pi_{\mathcal{M}(S)}(\theta^*) = \theta^*$.

  $\hookrightarrow \mathrm{Supp}(\theta^*) = \{i : \theta^*_i \neq 0\}$.

- The decomposability of R wrt $(\mathcal{M}(S), \underbrace{\mathcal{M}(S)^{\perp}}_{\bar{\mathcal{M}}(S)^{\perp}})$ follows from

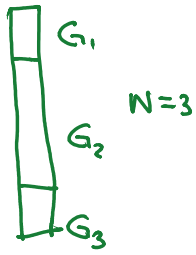
  $$\mathcal{M}(S)^{\perp} = \{\theta \in \mathbb{R}^d : \theta_j = 0 \ \forall j \in S\}$$

Example 2: Group-sparse vectors and mixed-norm regularizer

- Motivation: Groups of efficients likely to be zero or non-zero together

- Model: Partition $\{1, \cdots, d\}$ into N disjoint groups $G_1, \cdots, G_N$.

  Number of selected groups s should be small.



$$R(\theta) = \sum_{i=1}^{N} \|\theta_{G_i}\|_p \quad , \quad p \in [1, \infty] \quad \text{(mixed } \ell_{1,p}\text{-norm)}$$

↳ 1-norm of $(\|\theta_{G_1}\|_p, \cdots, \|\theta_{G_N}\|_p)$

- model subspace: For any $S \subseteq \{1, \cdots, N\}$ of cardinality s,

  define $\mathcal{M}(S) \triangleq \{ \theta \in \mathbb{R}^d : \theta_{G_i} = 0 \quad \forall i \notin S \}$

  Take $\bar{\mathcal{M}}(S) = \mathcal{M}(S)$. Then, the decomposability of R wrt $(\mathcal{M}(S), \mathcal{M}(S)^\perp)$ follows from

$$\mathcal{M}(S)^\perp = \{ \theta \in \mathbb{R}^d : \theta_{G_i} = 0 \quad \forall i \in S \}$$

Next:

<u>Proposition</u>: Suppose that $\mathcal{L}$ is smooth and convex,

$\lambda \geq 2 R^*(\nabla \mathcal{L}(\theta^*))$ and R is decomposable wrt $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$.

↳ dual norm

Then,

$$\hat{\Delta} = \hat{\theta} - \theta^* \in \mathcal{C} \triangleq \{ \Delta \in \mathbb{R}^d : R(\Delta_{\bar{\mathcal{M}}^\perp}) \leq 3 R(\Delta_{\bar{\mathcal{M}}}) + 4 R(\theta^*_{\mathcal{M}^\perp}) \}$$