Regularization

2020年9月30日 8:58

Minimize $\frac{1}{2N}\sum_i (w^T x_i - y_i)^2$. with $\|w\|$ constrained

$\searrow$ $\min_w \frac{1}{2N}\sum_i (w^T x_i - y_i)^2 + \frac{\lambda}{2}\|w\|_2^2$

called ridge regression

$\Rightarrow \nabla_w L = \frac{1}{N}\sum_i (w^T x_i - y_i) x_i + \lambda w$

Hessian $H = \frac{1}{N}\sum_i x_i x_i^T + \lambda I \Rightarrow \lambda_{min}(H) \geq \lambda$ Strongly Convex !

LASSO : sparse solutions.

Find "important" features from a large number of features

$\Rightarrow$ Minimize $\frac{1}{2N}\sum_i (w^T x_i - y_i)^2$, with $\|w\|_0 \leq C$

(find first c most important)

$\swarrow$

In practice, use L_1-loss for regularization

$\min_w \frac{1}{2N}\sum_i (w^T x_i - y_i)^2 + \lambda \|w\|_1$

Compressed Sensing

Nyquist Theorem : for a signal with frequency f, need $2f$ sampling rate to fully construct the signal.

- But sometimes we can compress (video / image ...) much smaller

Suppose x is a long, sparse vector

$A = [a_1, \cdots, a_n]^T \in \mathbb{R}^{n \times d}$ $(n << d)$. $\Leftarrow$ measurement matrix.

$\Rightarrow \boxed{y} = \boxed{\quad A \quad} \begin{bmatrix} \\ x \\ \\ \end{bmatrix}$

(sparse)

- Restricted Isoperimetric Property

$W \in \mathbb{R}^{n \times d}$ is (ε, S) - RIP if $\forall x \neq 0$. s.t. $\|x\|_0 \leq S$

Restricted Isoperimetric Property

$W \in \mathbb{R}^{n \times d}$ is (ε, S) - RIP if $\forall x \neq 0$. s.t. $\|x\|_0 \le S$

we have $(1-\varepsilon)\|x\|_2^2 \le \|Wx\|_2^2 \le (1+\varepsilon)\|x\|_2^2$

Theorem 1. $\varepsilon < 1$. W be $(\varepsilon, 2S)$ - RIP. $\|x\|_0 \le S$.

$y = Wx$. let $\tilde{x} \in \underset{v:\, Wv = y}{\operatorname{argmin}} \|V\|_0$ be a reconstructed vector of x.

$\Rightarrow$ Then, $\tilde{x} = x$.

Proof. If $\tilde{x} \neq x$.

$\because$ $Wx = y$, $\|\tilde{x}\|_0 \le \|x\|_0 \le S$

$\therefore$ $\|\tilde{x} - x_0\| \le 2S$

$\Rightarrow$ $(1-\varepsilon)\|\tilde{x} - x_0\|_2^2 \le \underbrace{\|W(\tilde{x} - x_0)\|_2^2}_{=0} \le (1+\varepsilon)\|\tilde{x} - x_0\|_2^2$

$\Rightarrow$ $\varepsilon \geq 1$ X.

Since $\underset{v:\, Wv = y}{\operatorname{argmin}} \|V\|_0$ is hard to compute, we approximate it using $\ell 1$.

Theorem 3. $\varepsilon < \frac{1}{1+\sqrt{2}}$. Let W be $(\varepsilon, 2S)$ - RIP. x any vector. $x_S \in \underset{V:\, \|V\|_0 \le S}{\operatorname{argmin}} \|x - V\|_1$.

x_S has x. s largest elements

let $y = Wx$.

$x^* \in \underset{V:\, Wv = y}{\operatorname{argmin}} \|V\|_1$.

$\Rightarrow$ $\underbrace{\|x^* - x\|_2}_{h} \le 2(1-\rho)^{-1} S^{\frac{1}{2}} \|x - x_S\|_1$, $\rho = \frac{\sqrt{2}\,\varepsilon}{1-\varepsilon}$.

intuition : h sparse $\Rightarrow$ easy $\Rightarrow$ $\|h\|_2 \le \|h'\|_2 + \|h''\|_2$

2S-sparse (d-2S) entries

Lemma 1. RIP $\Rightarrow$ Almost Orthogonality.

W. $(\varepsilon, 2S)$ - RIP. $\forall$ set I, J disjoint of size $\le S$

$\forall u$, $\langle Wu_I, Wu_J \rangle \le \varepsilon \|u_I\| \cdot \|u_J\|$.

e.x. $u = \begin{pmatrix} 5 \\ 6 \\ 7 \\ 8 \end{pmatrix}$. $I = \{1, 2\}$, $J = \{3, 4\}$

$\Rightarrow$ $u_I = \begin{pmatrix} 5 \\ 6 \\ 0 \\ 0 \end{pmatrix}$ $u_J = \begin{pmatrix} 0 \\ 0 \\ 7 \\ 8 \end{pmatrix}$

Proof. WLOG. assume $\|u_I\| = \|u_J\| = 1$.

$\langle Wu_I, Wu_J \rangle = \frac{\|Wu_I + Wu_J\|^2 - \|Wu_I - Wu_J\|^2}{4}$

$$\langle Wu_I, Wu_J \rangle = \frac{\|Wu_I + Wu_J\| - \|Wu_I - Wu_J\|}{4}$$

Since $|I \cup J| \le 2s$. by RIP

$$\|Wu_I + Wu_J\|^2 \le (1+\varepsilon)\left(\|u_I\|^2 + \|u_J\|^2\right) = 2(1+\varepsilon).$$

$$-\|Wu_I - Wu_J\|^2 \le -(1-\varepsilon)\left(\|u_I\|^2 + \|u_J\|^2\right) = -2(1-\varepsilon)$$

$$\Rightarrow \quad \langle Wu_I, Wu_J \rangle \le \varepsilon \|u_I\| \cdot \|u_J\|.$$

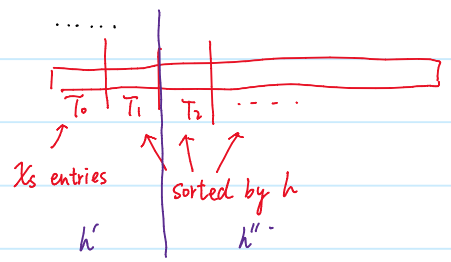
$[d] = T_0 \cup T_1 \cup \cdots \cup T_{d/s-1}$, $|T_i| = s$.

$d \% s = 0$

T_0 : s largest elem of $\underline{x}$ (all elems of x_s)
(abs)

$T_0^c = [d] \setminus T_0$.

$\Rightarrow$ T_1 : s largest of $\underline{h}$ on T_0^c



Proof (Main Theorem)

h'' ① $\|h_{T_{0,1}^c}\|_2 \le \|h_{T_0}\|_2 + 2s^{-1/2}\|x - x_s\|_1$

h' ② $\|h_{T_{0,1}}\|_2 \le \dfrac{\rho}{1-\rho} s^{-1/2} \|x - x_s\|_1$

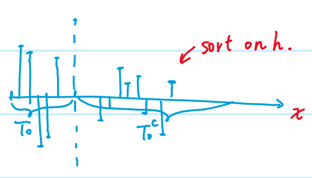
Then $\|h\|_2 \le \|h_{T_{0,1}}\|_2 + \|h_{T_{0,1}}^c\|_2$

$$\le \|h_{T_{0,1}}\|_2 + \|h_{T_0}\|_2 + 2s^{-1/2}\|x - x_s\|_1$$

$$\le 2\|h_{T_{0,1}}\|_2 + 2s^{-1/2}\|x - x_s\|_2$$

$$\le 2(1-\rho)^{-1} s^{-1/2} \|x - x_s\|_1$$

① $\|x\|_1 \ge \|x + h\|_1 = \|x^*\|_1$ $\left(\because\ x^* = \underset{v: Wv = y}{\operatorname{argmin}} \|v\|_1\right)$



split h into two parts

split h into two parts

$\|h_{T_0^c}\|$ cannot be too large —— I[+3] $\Rightarrow \|h\| \le 6$

Another rigourous proof to this:

$\|x\|_1 \ge \|x+h\|_1 = \sum_{i\in T_0} |x_i+h_i| + \sum_{i\in T_0^c} |x_i+h_i|$

$\ge \|x_{T_0}\|_1 - \|h_{T_0}\|_1 + \|h_{T_0^c}\| - \|x_{T_0^c}\|_1$

since $\|x_{T_0^c}\|_1 = \|x-x_s\|_1 = \|x\|_1 - \|x_{T_0}\|_1$

$\Rightarrow \|h_{T_0^c}\|_1 \le \|h_{T_0}\|_1 + 2\|x_{T_0^c}\|_1$.

$$\Rightarrow \underbrace{\|h_{T_0^c}\|_1 \le \|h_{T_0}\|_1 + 2\|x_{T_0^c}\|_1}$$

h在T_0上最大增量.

$\forall j>1 \quad \forall i \in T_j. \quad i' \in T_{j-1}. \quad \|h_i\| \le |h_i'\|.$

$\Rightarrow \|h_{T_j}\|_\infty \le \|h_{T_{j-1}}\|_1 / s$

$\Rightarrow \|h_{T_j}\|_2 \le s^{1/2} \|h_{T_j}\|_\infty \le s^{-1/2} \|h_{T_{j-1}}\|_1$

Triangle $\underbrace{\|h_{T_{0,1}^c}\|_2 \le \sum_{j=2} \|h_{T_j}\|_2 \le s^{-1/2} \sum_{j=1}^{d/s-2} \|h_{T_j}\|_2} \le \|h_{T_0^c}\|_1 s^{-1/2}.$

$\Rightarrow \|h_{T_{0,1}^c}\|_2 \le \|h_{T_0}\|_2 + 2s^{-1/2} \|x - x_s\|_1$

② $(1-\varepsilon)\|h_{T_{0,1}}\|_2^2 \le \|W h_{T_{0,1}}\|_2^2$

Since $W h_{T_{0,1}} = Wh - \sum_{j\ge 2} W h_{T_j} = -\sum_{j\ge 2} W h_{T_j}$.

$W(x^* - x) = 0$

$\Rightarrow \|W h_{T_{0,1}}\|^2 = -\sum_{j\ge 2} \langle W h_{T_0} + W h_{T_1}, W_{T_j} \rangle$

By almost-orthogonality,

$\le \quad \varepsilon \sum_{j\ge 2} (\|h_{T_0}\|_1 + \|h_{T_1}\|_1) \|h_{T_j}\|_2.$

$\le \quad \sqrt{2}\,\varepsilon \|h_{T_{0,1}}\|^2 \sum_{j\ge 2} \|h_{T_j}\|_2$

see ①

$\Rightarrow \|h_{T_{0,1}}\|_2^2 (1-\varepsilon) \le \sqrt{2}\,\varepsilon \quad s^{-1/2} \|h_{T_0^c}\|_1$

$\Rightarrow \|h_{T_{0,1}}\|_2^2 \le \frac{\sqrt{2}\,\varepsilon}{1-\varepsilon} s^{-1/2} \|h_{T_0^c}\|$

$\le \rho s^{-1/2} (\|h_{T_0}\|_1 + 2\|x_{T_0^c}\|_1)$

$\Rightarrow \|h_{T_{0,1}}\|_2^2 \le \frac{\rho s^{-1/2}}{1-\rho} 2\|x - x_s\|_1$

Q.E.D of Theorem 3.

Theorem 4. Construct RIP Matrix

Let $\mathcal{U}$ be arbitrary $n\times n$ orthonormal matrix. $\varepsilon, \delta \in (0,1)$. s integer in $[d]$

let n be an integer : $n \geq 100 \dfrac{s \ln(40d/(\delta\varepsilon))}{\varepsilon^2}$

$\Rightarrow$ Let $W \in \mathbb{R}^{n\times d}$ be a matrix s.t. each element of W distributed normally with $\mu = 0$. $\sigma^2 = \frac{1}{n}$.

Then w.p. of $\geq 1-\delta$, $W\mathcal{U}$ is (ε, s)-RIP.

Covering Balls. Make infinite into finite

Lemma 2. $\varepsilon \in (0,1)$. $\exists\ Q \subset \mathbb{R}^d$, $|Q| \leq (\frac{5}{\varepsilon})^d$ s.t.

$$\sup_{x:\ \|x\| \leq 1} \min_{v \in Q} \|x - v\| \leq \varepsilon.$$

Proof. $Q' = \{x \in \mathbb{R}^d : \forall j. \exists i \in \{-k, -k+1, \cdots, k-1, k\}$ s.t. $x_j = \frac{i}{k}\}$

$|Q'| = (2k+1)^d$, set $Q = Q' \cap B_2(1)$.

Lemma 5 (JL Lemma) Q: a finite set of vectors in $\mathbb{R}^d$. $\delta \in (0,1)$. $n \in \mathbb{Z}$

s.t. $\varepsilon = \sqrt{\dfrac{6 \ln(2|Q|/\delta)}{n}} \leq 3$

$\Rightarrow$ w.p. of $\geq 1-\delta$. a random $W \in \mathbb{R}^{n\times d}$ with elements iid. $N(0, \frac{1}{n})$.

$$\Rightarrow \sup_{x \in Q} \left| \frac{\|Wx\|_2^2}{\|x\|_2^2} - 1 \right| < \varepsilon.$$

Lemma 3. Let $\mathcal{U}$ be an $d\times d$ orthonormal matrix.

Index Set $I \subset [d]$, $|I| = s$.

$\mathcal{U}_i$: i^{th} column of $\mathcal{U}$ $\Rightarrow$ $S = \mathrm{span}\{\mathcal{U}_i, i \in I\}$. $\delta, \varepsilon \in (0,1)$.

$n \geq 24 \dfrac{\ln(2/\delta) + \varepsilon \ln(20/\varepsilon)}{\varepsilon^2}$

$\Rightarrow$ w.p. of $\geq 1-\delta$. a random $W \in \mathbb{R}^{n\times d}$ with elements iid. $N(0, \frac{1}{n})$.

$$\Rightarrow \sup_{x \in S} \left| \frac{\|Wx\|}{\|x\|} - 1 \right| < \varepsilon.$$

Apply Union Bound to all I $\Rightarrow n \geq 24 \dfrac{s \ln(40d/\delta\varepsilon)}{\varepsilon^2}$

So we prove theorem 4.

Proof. WLOG. $\|x\|_2 = 1$

Since $x \in \mathrm{span}\{\mathcal{U}_i\}$. $x = \mathcal{U}_I a$, $a \in \mathbb{R}$

$\|a\|_2 = 1$. so pick Q of size $|Q| \leq (20/\varepsilon)^s$ (by Lemma 2)

s.t. $\sup_{a:\ \|a\| = 1} \min_{v \in Q} \|a - v\| \leq \varepsilon/4$

($\mathcal{U}$ orthonormal)

$\alpha: \|\alpha\|=1 \quad v \in Q$

(U orthonormal)

$$\Rightarrow \sup_{\alpha: \|\alpha\|=1} \min_{v \in Q} \|U_I \alpha - U_I v\| \leq \frac{\varepsilon}{4}$$

by JL lemma, with prob. $\geq 1-\delta$

$$\sup_{v \in Q} \left| \frac{\|W(U_I v)\|}{\|U_I v\|} - 1 \right| \leq \varepsilon/2 .$$

Let α be the smallest number s.t. $\forall x \in S$, $\frac{\|Wx\|}{\|x\|} \leq 1+\alpha$

we want to show $\alpha \leq \varepsilon$

$$\|Wx\| \leq \|WU_I v\| + \|W(x - U_I v)\|$$

$$\leq 1 + \varepsilon/2 + (1+\alpha)\varepsilon/4$$

$$\Rightarrow \alpha \leq \varepsilon/2 + (1+\alpha)\varepsilon/4 \quad \Rightarrow \alpha \leq \varepsilon$$

The other side: $\|Wx\| \geq \|WU_I v\| - \|W(x - U_I v)\| \geq 1 - \varepsilon/2 - (1+\varepsilon)\varepsilon/4 \geq 1-\varepsilon$. Q.E.D

More general. what if W is non-linear ?

Neural Network : images are dense in pixel space,

but sparse in latent space (feature space).