ԴԱՍԸՆԹԱՑԻ ԲՈՎԱՆԴԱԿՈՒԹՅՈՒՆ

Գլ. I Ներածություն

Գլ. II Դասական ուսուցման տեսություն

Գլ. III Անբացահայտ կանոնավորում ԴԱՍ 6
(Implicit regularisation)

1) Ներածություն

$(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ iid P^* $i=1,\dots,n$

$\mathcal{F}_0 = \{ f(x,\theta) = \sigma_L(W_L \cdots \sigma_1(W_1 x)) : \theta \in \mathbb{R}^p \}$

Նշանակենք

$$\widehat{\Theta} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell\big(f(X_i,\theta), Y_i\big)$$

այսինքն $\widehat{\Theta}$-ն բոլոր մինիմումի կետերի բազմությունն է: Երբ $d > n$, $\widehat{\Theta}$-ը ~~գրեթե~~ բոլոր դեպքերում ունի անվերջ քանակով տարրեր: Որոշ տարրեր ունեն փոքր ընդհանրացման սխալանք, (1

որոշ տարրեր՝ մեծ: Դա կախված է նաև կողմնակի հանգամանքներից: Այն փաստը, որ գրադիենտի վայրէջքով հաշված $\hat{\theta}_n$-ը ունի փոքր ընդհանրացման սխալանք, բացատրվում է նրանով, որ գրադիենտի վայրէջքը գտնում է «կանոնավոր» մինիմումի կետ: Այս ուղղությամբ արված աշխատանքները հիմնականում կրում են հետևյալ տեսքը՝

$$\hat{\theta}_n^{GD} \approx \hat{\theta}_n^* \in \arg\min_{\theta \in \widehat{\Theta}_n} \|\theta\|_q$$

որտեղ $q \in [1; +\infty)$ և $\|\theta\|_q$ նորմը սահմանվում է հետևյալ ձևով՝

$$\|\theta\|_q = \Big(\sum_{j=1}^{p} |\theta_j|^q\Big)^{1/q} \quad \forall \theta \in \mathbb{R}^p:$$

Այս դասում կդիտարկենք այդպիսի աշխատանքների մի քանի օրինակներ: (2

2) Գրադիենտի վայրէջք գծային ցանցում

Դիտարկենք պարզագույն ցանցը՝
$L = 1$ (թաքնված շերտ չկա) և $\sigma(x) = x$:
Այս դեպքում
$$f(x;\Theta) = \langle \Theta; x \rangle = x^T \Theta :$$
Դիտարկենք քառակուսային կորուստը՝
$$\ell(y, y') = (y - y')^2 :$$
Ուրեմն,
$$\widehat{\Theta}_n = \arg\min_{\theta \in \mathbb{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_i - X_i^T \theta)^2}_{\|Y - \mathbb{X}\theta\|_2^2}$$

Ենթադրություն 1 $X_1, \dots, X_n \in \mathbb{R}^d$ վեկտորները գծորեն անկախ են:

Սա համարժեք է հետևյալ պնդումներին՝

(a) $\mathrm{Rank}(\mathbb{X}) = n$ (c) $\dim(\ker(\mathbb{X})) = d - n$

(b) $\dim(\mathrm{Im}(\mathbb{X})) = n$ (d) $\mathbb{X} \cdot \mathbb{X}^T$-ն հակադարձելի

Դիտողություն 1

Ենթադրություն 1-ից հետևում է, որ $d \geq n$:

Դիտողություն 2

Ենթադրություն 1-ից հետևում է, որ եթե $\widehat{\Theta}_n$-ը դատարկ չէ, ապա այն ափինային է $(d-n)$-չափանի ենթատարածություն: Իրոք, եթե $\widehat{\theta}_n \in \widehat{\Theta}_n$, ապա $\widehat{\theta}_n + u \in \widehat{\Theta}_n$ $\forall u \in \ker(\mathbb{X})$:

Դիտողություն 3 Ենթ. 1-ի տեղի ունենալու դեպքում $\widehat{\Theta}_n \neq \emptyset$:

Իրոք $(b) \Rightarrow \dim(\mathrm{Im}(\mathbb{X})) = n \Rightarrow$
$$\Rightarrow \mathrm{Im}(\mathbb{X}) = \mathbb{R}^n \Rightarrow \exists\, \theta_0 \text{ s.t. } \mathbb{X}\theta_0 = Y \Rightarrow \frac{1}{n}\|Y - \mathbb{X}\theta_0\|_2^2 = 0$$
$$\Rightarrow \theta_0 \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n}\|Y - \mathbb{X}\theta\|_2^2$$

Գրադիենտային վայրէջք $\theta_0 = 0$
$$\theta_{k+1} = \theta_k + \frac{2h_k}{n} \mathbb{X}^T (Y - \mathbb{X}\theta_k)$$

ԴԱՍ 6

$$\theta_{k+1} = \theta_k + \underbrace{\frac{2h_k}{n}\sum_{i=1}^{n}(Y_i - X_i^T\theta)\cdot X_i}_{X_1,\dots,X_n\text{-ի գծային կոմբինացիա}}$$

$$\hat\theta_K^{GD} = \theta_K$$

Դիտարկում $\hat\theta_K^{GD} \in \theta_0 + \mathrm{Im}(\mathbb{X}^T) = \mathrm{Im}(\mathbb{X}^T)$:

Թեորեմ 1. Եթե ճիշտ է Ենթ.1-ը և $\{h_k\}$ հաջորդականությունն այնպիսին է, որ

$$\hat\theta_K^{GD} \xrightarrow[K\to+\infty]{} \hat\theta_\infty^{GD},$$

ապա $\hat\theta_\infty^{GD} \in \arg\min\{\|\theta\|_2^2 : \theta \in \hat\Theta_n\}$
$= \arg\min\{\|\theta\|_2^2 : \mathbb{X}\theta = Y\}$:

Ապացույց $\hat L_n(\theta) = \frac{1}{n}\|Y - \mathbb{X}\theta\|_2^2 \geq 0$:

Ենթ.3-ից և (b)-ից $\Rightarrow \exists \theta_0 \in \hat\Theta_n$ s.t.
$\hat L_n(\theta_0) = 0$: Հետևաբար $\min_{\theta\in\mathbb{R}^d} \hat L_n(\theta) = 0$
և $\hat\Theta_n = \{\theta : \hat L_n(\theta) = 0\}$
$= \{\theta : \mathbb{X}\theta = Y\}$ (ինտերպոլյացիա)

Քանի որ $\hat L_n(\theta)$-ն ուռուցիկ ֆունկցիա է, գիտենք, որ $\hat\theta_\infty^{GD} \in \hat\Theta_n (\Leftrightarrow \mathbb{X}\hat\theta_\infty^{GD} = Y)$:

Ուրեմն, $\mathbb{X}\hat\theta_\infty^{GD} = Y$ և $\hat\theta_\infty^{GD} \in \mathrm{Im}(\mathbb{X}^T)$
$(\mathrm{Im}(\mathbb{X}^T) = \mathrm{span}(X_1,\dots,X_n))$:

Մնաց ապացուցել հետևյալ լեմման՝

Լեմմա $\tilde\theta \in \arg\min\{\|\theta\|_2^2 : \mathbb{X}\theta = Y\}$ (*)
այն և միայն այն դեպքում, երբ
$\mathbb{X}\tilde\theta = Y$ և $\tilde\theta \in \mathrm{Im}(\mathbb{X}^T)$:

Թեորեմ 1-ը ապացուցված է: ∎

Լեմմայի ապացույց Քանի որ $\theta \mapsto \|\theta\|_2^2$ ֆունկցիան ուռուցիկ է, իսկ $\mathbb{X}\theta = Y$ սահմանափակումները՝ գծային, $\tilde\theta$-ը կլինի (*)-ի լուծում այն և միայն այն դեպքում, երբ այն բավարարում է KKT պայմաններին:

Լագրանժյան՝ $\mathcal{L}(\theta,\mu) = \|\theta\|_2^2 + \sum_{i=1}^{n}\mu_i(X_i^T\theta - Y_i)$

$$\text{KKT}: \begin{cases} X\tilde\theta = Y \\ \exists \tilde\mu \in \mathbb{R}^n : \nabla_\theta \mathcal{L}(\tilde\theta,\tilde\mu) = 0 \end{cases}$$

$\Leftrightarrow X\tilde\theta = Y$ և $2\tilde\theta + \sum_{i=1}^{n} X_i\tilde\mu_i = 0$

$\Leftrightarrow X\tilde\theta = Y$ և $\tilde\theta = -\frac{1}{2}\mathbb{X}^T\tilde\mu \Leftrightarrow X\tilde\theta = Y$ և $\tilde\theta \in \mathrm{Im}(\mathbb{X}^T)$

Նկատառում 1. (*) օպտիմիզացիայի խնդիրը ունի լուծում և այն միակն է:
Ավելին $\tilde{\theta} = \arg\min \{\|\theta\|_2^2 : X\theta = Y\}$

$$= X^{+} \cdot Y$$

որտեղ X^{+}-ը X-ի պսևդոհակադարձն է:
Այն սահմանվում է հետևյալ կերպ: Եթե $X = U \cdot D \cdot V^T$-ը X-ի SVD-ն է, որտեղ

$$D = \Big[\underbrace{\begin{matrix} \sigma_1 & & 0 \\ 0 & & \sigma_n \end{matrix}}_{n} \; \underbrace{\begin{matrix} 0 & \dots & 0 \\ 0 & \dots & 0 \end{matrix}}_{d-n}\Big] \Big\} n \quad \sigma_i > 0 \quad i=1,\dots,n$$

ապա

$$X^{+} = V \cdot D^{+} \cdot U^T \quad \text{որտեղ} \quad D^{+} = \underbrace{\begin{bmatrix} \sigma_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \sigma_n^{-1} \\ 0 & \dots & 0 \\ & \ddots & \\ 0 & \dots & 0 \end{bmatrix}}_{n} \Big\} d-n$$

Նկատառում 2.
$\tilde{\theta}$-ն կարելի է ստանալ նաև որպես ռիջ-ռեգրեսիայի սահման՝

$$\hat{\theta}(\lambda) = \arg\min_{\theta \in \mathbb{R}} \left(\frac{1}{n}\|Y - X\theta\|_2^2 + \lambda \|\theta\|_2^2\right):$$

Կարելի է ցույց տալ, որ

$$\lim_{\lambda \to 0} \hat{\theta}(\lambda) = \tilde{\theta}:$$

3. GD բաժանելի դեպքում

Դիտարկենք գծային մոդելի դեպքը՝

$$f(x;\theta) = \theta^T x$$

Բայց փոխենք կորստի ֆունկցիան՝

$$\ell(y, y') = \exp(-y \cdot y'):$$

Այս ֆունկցիան օգտագործվում է բուստինգ մեթոդի շրջանակներում, հատկապես երբ $Y_i \in \{-1, +1\}$:

$$\hat{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \exp\left(-Y_i \cdot (X_i^T \theta)\right)\}$$

Կարելի է հեշտ համոզվել, որ

$$\hat{L}_n(\theta) > 0 \quad \forall\, \theta \in \mathbb{R}^d$$

և

$$\hat{L}_n(\alpha \cdot \theta^*) \xrightarrow[\alpha \to \infty]{} 0$$

կամայական θ^*-ի համար, այնպիսին, որ

$$Y_i = \mathrm{sign}(X_i^T \theta^*) \quad i = 1, \dots, n \quad (1):$$

Եթե ենթ. 1-ը ճիշտ է, ապա $\exists$ անվերջ

համասեռ θ^{*} որոնց համար (1)-ը ճիշտ է:
Հետևաբար

$$\min_{\theta\in\mathbb{R}^d} \hat{L}_n(\theta)$$

խնդիրը լուծում չունի, բայց

$$\hat{\Theta}_n = \left\{\theta\in\mathbb{R}^d : \lim_{\alpha\to+\infty} \hat{L}_n(\alpha\theta) = 0\right\} \neq \emptyset:$$

<u>Թեորեմ 2</u> Եթե $\hat{\theta}^{CGD}_K \to \hat{\theta}^{CGD}_\infty$,

ապա $$\hat{\theta}^{CGD}_\infty = \arg\min_\theta \left\{\|\theta\|_1 : y_i\cdot(\theta^T x_i) \geq 1 \quad \forall i=1,\dots,n\right\}$$

լավագույն մաքսիմալացված
«կարծրային» դասակարգող

<u>Նշանակում</u> $\hat{\theta}^{CGD}_K$ = coordinatewise gradient descent

<u>Նշանակում</u> Եթե CGD-ի փոխարեն օգտագործենք GD, ճիշտ է հետևյալը

$$\theta_K = \tilde{\theta} + O\left(\frac{\log\log K}{\log K}\right),$$ որտեղ

$$\tilde{\theta} = \arg\min_\theta \left\{\|\theta\|_2 : Y_i\cdot(\theta^T X_i) \geq 1 \quad \forall i=1,\dots,n\right\}$$

և $$h \leq \frac{8}{\lambda_{max}\left(\frac{1}{n}\mathbb{X}^T\mathbb{X}\right)}$$: