ԽՈՐԸ ՈՒՍՈՒՑՈՒՄ / ԴԱՍ II

Առնակ Դալալյան
18/08/2021

1) Հիշեցում

* $\underbrace{(X_1, Y_1), \dots, (X_n, Y_n)}_{\text{դիտարկումներ}}$ անկախ, $\underbrace{P^*}_{\text{անհայտ}}$ բաշխմամբ
* $f : \mathcal{X} \to \mathcal{Y}$ ֆունկցիայի արտապատկերումը՝ կանխատեսիչ
* $f(x, \theta) = \sigma_L(W_L \dots \sigma_1(W_1 x))$ – $(L-1)$ թաքնված շերտով նեյրոնային ցանց (ՆՑ), որտեղ

  $(L, d_0, \dots, d_L)$ – կառուցվածք

  $\theta = (W_1, \dots, W_L)$ – պարամետրեր
* Մարզման սխալանք (երբ $\mathcal{Y} \subset \mathbb{R}$)

$$\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

$$\hat{L}_n(\theta) := \hat{L}_n(f(\cdot, \theta))$$

* ERM (empirical risk minimizer)

$$\hat{\theta}_n \in \arg\min_{\theta} \hat{L}_n(\theta) \qquad (*)$$

* Դժվար է լուծել $(*)$-ը երբ $\theta \in \mathbb{R}^p$ շատ մեծ p-ով է, և $\hat{L}_n$-ը ուռուցիկ ֆունկցիա չէ:
* Ընդհանուր սխալանք՝ $L(f) = \mathbb{E}[(Y_i - f(X_i))^2]$:

2) Ի՞նչ է օվերֆիթինգը (overfitting)

* Իդեալական կանխատեսիչը՝

$$f^* \in \arg\min_{f} L(f)$$

որտեղ մինիմումը բոլոր հնարավոր կանխա-տեսիչների վրա է: $\mathcal{F} = \{ f: \mathcal{X} \to \mathcal{Y} \text{ չափելի} \}$

* Լավագույն ՄՈ՝

$f(\cdot, \theta^*)$ որտեղ $\theta^* \in \arg\min_{\theta \in \mathbb{R}^p} L(\theta)$

Դիտարկում՝ $\{ f(\cdot, \theta) : \theta \in \mathbb{R}^p \} \subsetneq \mathcal{F}$, հետևաբար $L(f^*) \leq L(\theta^*)$:

Սահմ. $\left(L(\theta^*) - L(f^*)\right)$-ը անվանում ենք մոտարկման սխալանք (approximation error):

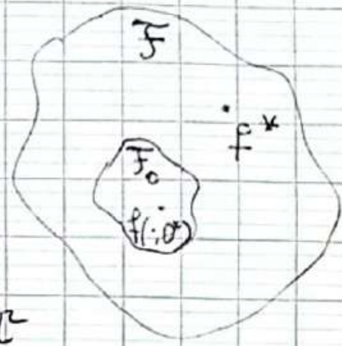
* ERM ՄՈ՝ $\hat{f}_n(\cdot) = f(\cdot, \hat{\theta}_n)$ որտեղ $\hat{\theta}_n$-ն $(*)$-ի լուծում է:

Դիտարկում՝ $L(\theta^*) \leq L(\hat{\theta}_n)$ և $\hat{L}_n(\hat{\theta}_n) \leq \hat{L}_n(\theta^*)$

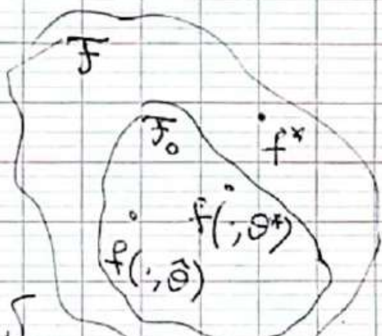
Սահմ. $\left(L(\hat{\theta}_n) - L(\theta^*)\right)$-ն անվանում ենք գնահատման կամ ուսուցման սխալանք:

Սահմ. $|\hat{L}_n(\hat{\theta}_n) - L(\hat{\theta}_n)|$-ը անվանում ենք ընդհանրացման սխալանք (generalisation error):

* Դիտարկենք $\mathcal{F}_0 = \{f(\cdot,\theta) : \theta \in \mathbb{R}^p\}$:



a) Երբ p-ն փոքր է, $\mathcal{F}_0$ բազմությունը փոքր է, և θ^*-ը "հեշտ" է գնահատել: Սակայն, այդ դեպքում գնահատման սխալանքը փոքր է, բայց կարող է մեծ լինել մոտարկման սխալանքը: Այս երևույթը անվանում են «under-fitting» (թերուսուցում):



b) Երբ p-ն մեծ է, θ^*-ը ավելի «դժվար է» գնահատել, բայց f^*-ը ավելի լավ է մոտարկվում $f(\cdot,\theta^*)$-ով: Այս երևույթն անվանում են «overfitting» (գերուսուցում):

c) Գերուսուցման դեպքում, քանի որ մեծ է պարամետրերի քանակը (θ-ի չափողականությունը), փոքր է ճշգրտման սխալանքը: Հաճախ՝

$$\hat{L}_n(\hat{\theta}_n) = \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(X_i;\theta)\right)^2 = 0:$$

(« կատարյալ ուսուցում - perfect fit »)

Երբ դրան զուգահեռ՝ փոքր է նաև ընդհանուր սխալանքը (կամ ընդհանրացման սխալանքը), ապա գերուսուցումը կանվանենք «բարորակ»:

Կարևոր՝ Ստոխաստիկ գրադիենտի մեթոդով մարզված ԽՈՐԸ ՆՑ-երի դեպքում տեղի է ունենում բարորակ գերուսուցում:

3) Օպտիմիզացիա

Ինչպես լուծել $\hat{\theta}_n \in \underset{\theta \in \mathbb{R}^p}{\arg\min}\ \hat{L}_n(\theta)$ օպտիմիզացիայի խնդիրը:

* Գրադիենտային վայրէջք՝ եթե $\hat{L}_n$-ը դիֆերենցելի է

- ընտրել $\theta_0 \in \mathbb{R}^p$ և $h \in (0, +\infty)$
- հաշվել $\theta_{k+1} = \theta_k - h \nabla \hat{L}_n(\theta_k)$, $k = 0, 1, \dots, K-1$
- Վերադարձնել $\hat{\theta}_n^{GD} = \theta_K$

GD = Gradient Descent ; h = ուսուցման քայլ learning rate

* Ստոխաստիկ GD

Նկատենք, որ

$$\nabla \hat{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \left(Y_i - f(X_i, \theta)\right)^2 :$$

Մեկ անգամ $\nabla \hat{L}_n$-ը հաշվելու համար անհրաժեշտ է հաշվել $\nabla_\theta f(\cdot, \theta)$-ն n անգամ:

- ընտրել $\theta_0 \in \mathbb{R}^p$ և $h > 0$
- $k = 0, 1, \dots, K$-ի համար
  - ընտրել պատահականորեն $i_k \in \{1, \dots, n\}$
  - հաշվել $\theta_{k+1} = \theta_k - h \nabla_\theta \left(Y_{i_k} - f(X_{i_k}, \theta_k)\right)^2$
- Վերադարձնել $\hat{\theta}_n^{SGD} = \theta_K$

Նկատառում $\hat{\theta}_n^{GD}$-ն և $\hat{\theta}_n^{SGD}$-ն չեն համընկնում (*)-ի լուծում $\hat{\theta}_n$-ի հետ, այլ միայն (լավագույն դեպքում) մոտարկված են այն:

$|L(\hat{\theta}_n^{GD}) - L(\hat{\theta}_n)|$-ը կոչվում է օպտիմիզացիայի սխալանք:

Դիտողություն Երեք տիպի սխալանք՝

1) մոտարկման՝ $L(f(\cdot,\theta^*)) - L(f^*)$
2) գնահատման՝ $L(f(\cdot,\hat{\theta}_n)) - L(f(\cdot,\theta^*))$
3) օպտիմիզացիայի՝ $|L(f(\cdot,\hat{\theta}_n^{GD})) - L(f(\cdot,\hat{\theta}_n))|$

Մենք հիմնականում կուսումնասիրենք գնահատման սխալանքը:

ԳԼՈՒԽ II ԴԱՍԱԿԱՆ ՈՒՍՈՒՑՄԱՆ ՏԵՍՈՒԹՅՈՒՆ

1) Մեծ թվերի հավասարաչափ օրենք

Ունենք $(X_1,Y_1),\dots,(X_n,Y_n) \overset{iid}{\sim} \mathbb{P}^*$ ($\mathcal{X}\times\mathcal{Y}$-ի վրա)

iid = independent and identically distributed.

$\mathcal{F} = \{ f : \mathcal{X} \to \mathcal{Y}$ չափելի$\}$.

$\mathcal{F}_0 \subset \mathcal{F}$ ($\mathcal{F}_0$-ն կարող է լինել ՆՑ-երի բազմություն)

$l : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ չափելի ֆունկցիա, որը կանվանենք կորուստի ֆունկցիա: Օրինակ՝ $l(y,y') = (y-y')^2$ կամ $l(y,y') = |y-y'|$: