Երևանի Պետական Համալսարան
Ինֆորմատիկայի և Կիրառական Մաթեմատիկայի Ֆակուլտետ
Թվային անալիզի և մաթեմատիկական մոդելավորման ամբիոն

Մագիստրոսական Թեզ

Թեմա՝ Տրանսֆերային ուսուցման որոշակի մեթոդի ընդհանրացման սխալանքի գնահատման մասին

Ուսանող՝ Մինասյան Գևորգ

Ղեկավար՝ ֆիզ. մաթ. գիտ. թեկնածու Հ.Է. Դանոյան

Երևան – 2019

Բովանդակություն

Նշանակումներ և սահմանումներ

$\mathcal{X}$-ով նշանակենք բոլոր հնարավոր տվյալների օրինակները, իսկ $\mathcal{C}$-ով նշանակենք բոլոր պիտակների կամ դասերի բազմությունը: Յուրաքանչյուր $c \in \mathcal{C}$ դասին համպատասխանում է $\mathcal{X}$ բազմության վրա որոշված ինչ-որ $\mathcal{D}_c(x)$ բաշխում, այն ցույց է տալիս, թե x օրինակը ինչքանով է c դասին համապատասխան: Ուսուցումը կատարվում է $\mathcal{F}$ ներկայացումների ֆունկցիաների դասի վրա: $\forall f \in \mathcal{F}$ ֆունկցիա $\mathcal{X}$ տվյաների բազմությունը արտապատկերում d-չափանի էվկլիդյան $\mathcal{R}^d$տարծություն՝ $f : \mathcal{X} \to \mathcal{R}^d$, բացի այդ կդիտարկենք միայն սահմանափակ ֆունկցիաները՝

$$||f(x)|| \leq R \, \forall x \in \mathcal{X} \text{ և } R > 0:$$

Վերահսկվող առաջադրանքներ

Այժմ կնկարագրենք այն առաջադրանքները, որոնց միջոցով փորձարկվելու է ներկայացումների f ֆուկցիան: $k+1$ դասերից բաղկացած $\mathcal{T}$ վերահսկվող առաջադրանքը, բաղկացած է

$$\{c_1, ..., c_{k+1}\} \subseteq \mathcal{C}$$

միմյանցից տարբեր դասերից: Կենթադրենք որ վերահսկվող առաջադրանքները ունեն $\mathcal{P}(\mathcal{T})$ բաշխում, որը բնութագրում է այդ առաջադրանքը դիտարկվելու հավանականությունը: $k+1$ դասերից բաղկացած վերահսկվող առաջադրանքների բաշխումը հետևյալն է՝

$$\mathcal{P}(\mathcal{T} \mid |\mathcal{T}| = k+1)$$

Պիտակավորված տվյալների բազմությունը $\mathcal{T}$ առաջադրանքի համար բաղկացած է m հատ միմյանցից անկախ և միևնույն բաշխումից ընտրրված օրինակներից: Այդ օրինակները ընտրվում են ստորև նկարագրված պրոցեսով:

$c \in \{c_1, ..., c_{k+1}\}$ *դասը ընտրվում է ըստ* $\mathcal{D}_\mathcal{T}$ *բաշխման, որից հետո* x *օրինակը ընտրվում է* $\mathcal{D}_c$ *բաշխումից: Դրանք միասին ձևավորում են պիտակավորված* (x, c) *կույգը, որը ունի հետևյալ բաշխումը՝*

$$\mathcal{D}_\mathcal{T}(x, c) = \mathcal{D}_c(x)\mathcal{D}_\mathcal{T}(c) :$$

Վերահսկվող ներկայացումների գնահատման չափը

f ներկայացումների ֆունկցիաի որակի գնահատումը կատարվում է $\mathcal{T}$ բազմադաս դասակարգման առաջադրանքի միջոցով՝ օգտագործելով գծային դասակարգիչ: Ֆիքսենք $\mathcal{T} = \{c_1, ..., c_{k+1}\}$ առաջադրանքը: $\mathcal{T}$ առաջադրանքի բազմադաս դասակարգիչը ֆուկցիա է՝ $g : \mathcal{X} \to \mathcal{R}^{k+1}$, որի արժեքի կորդինատները ինդեքսավորված են $\mathcal{T}$ առաջադրանքի դասերով: $(x, y) \in \mathcal{X} \times \mathcal{T}$ կետում g դասակարգիչով պայմանավորված կորուստը սահմանենք հետևյալ կերպ՝

$$l(\{g(x)_y - g(x)_{y'}\}_{y \neq y'}),$$

որը ֆունկցիա կախված k չափանի վեկտորից, այն ստացվում է $k+1$ չափանի $g(x)$ վեկտորի կորդինատների տարբերությունից, բացի այդ $\{g(x)_y - g(x)_{y'}\}_{y \neq y'}$ վեկտորի կոմպոնենտները կամայական հերթականությամբ կարելի է համարակալել և l-ի արժեքը կախված չէ վեկտորի կոմպոնենտների համարակալման հերթականությունից: Պրակտիկայում մեծ կիրառություն ունեցող երկու կորուստի ֆունկցիաներ ենք դիտարկելու աշխատանքում՝ ստանդարտ հինջ կորստի ֆունկցիան որը սահմանվում է հետևյալ կերպ՝

$$l(v) = \max\{0, 1 + \max_i\{-v_i\}\}$$

և լոգիստիկ կորստի ֆունկցիան՝

$$l(v) = \log_2(1 + \sum_i e^{-v_i}),$$

որտեղ $v \in \mathcal{R}^k$: $\mathcal{T}$ առաջադրանքի համար g դասակարգիչի կորուստը հետևյալն է՝

$$L(\mathcal{T}, g) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{(x,c)\sim\mathcal{D}_\mathcal{T}} \left[l(\{g(x)_c - g(x)_{c'}\}_{c \neq c'})\right]$$

f ներկայացումների ֆունկցիան օգտագործելու նպատակով, $g(x) = Wf(x)$ տեսքի դասակարգիչներն ենք դիտարկելու, որտեղ $W \in \mathcal{R}^{(k+1)\times d}$, որը ունի սահմանափակ նորմ՝ $||W|| \leq Q$ և $Q > 0$: $\mathcal{W}$-ով նշանակենք սահմանափակ նորմ ունեցող մատրիցաների բազմությունը՝

$$\mathcal{V} = \{W : ||W|| \leq Q \text{ և } Q > 0\}$$

$\mathcal{T}$ առաջադրանքի համար $g(x) = Wf(x)$ ներկայացումից կախված գծային դասակարգչի կորստի ֆունկցիան հետևյալն է՝

$$L(\mathcal{T}, f, W) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{(x,c)\sim\mathcal{D}_\mathcal{T}} \left[l(\{Wf(x)_c - Wf(x)_{c'}\}_{c \neq c'})\right]$$

Ֆիքսելով որևէ f ներկայացում կարելի լավագույն W գտնել, այնպես որ f-ից կախված գծային դասակարգչի կորուստը լինի ամենափոքրը, ուստի f ներկայացման վերահսկիչ կորուստը $\mathcal{T}$ առաջադրանքի համար կսահմանենք, այն կորուստը, երբ լավագույն W ենք ընտրել f-ի համար՝

$$L(\mathcal{T}, f) \overset{\text{def}}{=} \inf_{W \in \mathcal{V}} L(\mathcal{T}, f, W)$$

Սահմանում 1 (վերահսկիչ միջին կորուստ)**.** *$k + 1$ դասերից բաղկացած առաջադրանքների վերահսկիչ միջին կորուստը f ներկայացման համար սահմանվում է որպես՝*

$$L(f) \overset{\text{def}}{=} \mathop{\mathbb{E}}_{\mathcal{T} \sim \mathcal{P}} [L(\mathcal{T}, f) \mid |\mathcal{T}| = k + 1]$$

Սահմանում 2 (էմպիրիկ վերահսկիչ միջին կորուստ)**.** *Դիցուք ունենք միմյանցից անկախ $\mathcal{P}(\mathcal{T} \mid |\mathcal{T}| = k+1)$ բաշխումից ընտրված N հատ առաջադրանքներ՝ $\mathcal{T}_1, ..., \mathcal{T}_N$: Էմպիրիկ վերահսկիչ միջին կորուստը f ներկայացման համար հետևյալն է՝*

$$\hat{L}(f) \overset{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} L(\mathcal{T}_i, f)$$

Օժանդակ արդյունքներ

Լեմմա 1 (Հոֆդինգի անհավասարություն)**.** *Դիցուք $Z_1, ..., Z_m$ անկախ և միևնույն բաշխման պատահական մեծություններ են և $\bar{Z} = \frac{1}{m}\sum_{i=1}^{m} Z_i$: Ենթադրենք $\mathbb{E}[\bar{Z}] = \mu$ և յուրաքանչյուր i-ի համար $\mathbb{P}[a \leq Z_i \leq b] = 1$: Այդ դեպքում ցանկացած $\epsilon > 0$ թվի համար տեղի ունի հետևյալը՝*

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu > \epsilon\right] \leq e^{\frac{-2m\epsilon^2}{(b-a)^2}}$$

և

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu < -\epsilon\right] \leq e^{\frac{-2m\epsilon^2}{(b-a)^2}}$$

Լեմմա 2. *Դիցուք ունենք միմյանցից անկախ $\mathcal{P}(\mathcal{T} \mid |\mathcal{T}| = k+1)$ բաշխումից ընտրված N հատ առաջադրանքներ՝ $\mathcal{T}_1, ..., \mathcal{T}_N$ և ֆիքսենք կամայական $f \in \mathcal{F}$ ներկայացում: $\hat{L}(f)$ էմպիրիկ վերահսկիչ միջին կորուստն է f ներկայացման համար, իսկ $L(f)$-ը վերահսկիչ միջին կորուստը և դիցուք $| \cup_{i=1}^{N} T_i| = n$: Այդ դեպքում առնվազն $1-\delta$ հավանականությամբ տեղի ունի հետևյալ անհավասարությունը:*

$$\hat{L}(f) \geq L(f) - B\sqrt{\frac{(k+1)\log\left(\frac{1}{\delta}\right)}{2n}} \tag{1}$$

որտեղ B ինչ-որ դրական հաստատուն է:

Ապացույց. Օգտվելով $L(\mathcal{T}_i, f)$ սահմանումից և օգտագործելով f-ի սահմանափակությունը հեշտ է համոզվել որ գոյություն ունի B դրական թիվ այնպես որ կամայական $i \in [N]$ տեղի ունի հետևալը՝

$$0 \leq L(T_i, f) \leq B$$

Այժմ նկատենք որ Հոֆդինգի լեմմայի պայմանները բավարարված են և օգտվելով այդ լեմմայի անհավասարությունից կունենաք՝

$$\mathbb{P}[\hat{L}(f) - L(f)] < -\epsilon] \leq e^{\frac{-2N\epsilon^2}{B^2}}$$

որտեղից և հավանականության $\mathbb{P}[A] = 1 - \mathbb{P}[\bar{A}]$ հատկությունը օգտագործելով՝

$$\mathbb{P}[\hat{L}(f) - L(f) \geq -\epsilon] \geq 1 - e^{\frac{-2N\epsilon^2}{B^2}}$$

$e^{\frac{-2N\epsilon^2}{B^2}}$ հավասարեցնենք δ-ի՝

$$\delta = e^{\frac{-2N\epsilon^2}{B^2}}$$

և լուծելով այն ϵ-ի նկատմամբ՝ կունենաք հետևյալը՝

$$\epsilon = B\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2N}}$$

Այսպիսով առնվազն $1-\delta$ հավանականությամբ տեղի ունի հետևյալ անահավսարությունը՝

$$\hat{L}(f) \geq L(f) - B\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2N}}$$

Նկատենք որ $n \leq (k+1)N$, որտեղից անմիջապես հետևում է հետևյալ անհավասարությունը՝

$$\sqrt{\frac{k+1}{n}} \geq \sqrt{\frac{1}{N}}$$

Օգտագործելով վերջին անհավասարությունը կունենանք, որ առնվազն $1-\delta$ հավանականությամբ տեղի ունի

$$\hat{L}(f) \geq L(f) - B\sqrt{\frac{(k+1)\log\left(\frac{1}{\delta}\right)}{2n}}$$

անհավասարությունը: $\square$

Պնդում 1. *Կամայական $v \in \mathbb{R}^d$ վեկտորի համար տեղի ունի հետևյալը՝*

$$||v|| \leq \sqrt{2} \underset{\sigma\sim\{\pm1\}^d}{\mathbb{E}} \left|\sum_{i=1}^{d} \sigma_i v_i\right|$$

Թեորեմ 1. *Դիցուք $\mathcal{X}$-ը որևէ բազմություն է և $(x_1, x_2, ..., x_n) \in X^N$: Տրված է նաև $\mathcal{F}$ ֆունկցիաների բազմություն, որի կամայական $f \in \mathcal{F}$ ֆունկցիա $\mathcal{X}$ բազմությունը արտապատկերում է $\mathbb{R}^d$ էվկլիդյան տարածություն՝ $f : \mathcal{X} \to \mathbb{R}^d$: Դիցուք h_i ֆունկցիաներ ունենք որոնք $\mathbb{R}^d$ էվկլիդյան տարածությունը արտապատկերում են իրական թվերի $\mathbb{R}$ տարածություն՝ $h_i : \mathbb{R}^d \to \mathbb{R}$, կամայական $i \in [n]$ համար: Կենթադրենք, որ բոլոր h_i ֆունկցիաները, ինչ-որ L դրական հաստատունով Լիպշից հատկությամբ օժտված ֆունկցիաներ են: Այդ դեպքում տեղի ունի հետևյալ անհավասարությունը՝*

$$\underset{\sigma\sim\{\pm1\}^n}{\mathbb{E}} \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i h_i(f(x_i))\right] \leq \sqrt{2}L \underset{\sigma\sim\{\pm1\}^{nd}}{\mathbb{E}} \left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n}\sum_{j=1}^{d} \sigma_{ij} f_j(x_i)\right] \tag{2}$$

Թեորեմ 2-ը կարելի է ընդհանրացնել $h_i(v, y) \in \mathbb{R}$ ֆունկցիաների համար, որտեղ $v \in \mathbb{R}^d$, $y \in \mathcal{Y}$ և h_i ֆունկցիաները ըստ v փոփոխականի L հաստատունով Լիպշից հատկությամբ օժտված ֆունկցիաներ են կամայական $y \in \mathcal{Y}$ համար:

Թեորեմ 2. *Դիցուք $\mathcal{X}$-ը և $\mathcal{Y}$-ը որևէ բազմություններ են և $(x_1, x_2, ..., x_n) \in X^N$: Տրված է նաև $\mathcal{F}$ ֆունկցիաների բազմություն, որի կամայական $f \in \mathcal{F}$ ֆունկցիա $\mathcal{X}$ բազմությունը արտապատկերում է $\mathbb{R}^d$ էվկլիդյան տարածություն՝ $f : \mathcal{X} \to \mathbb{R}^d$: Դիցուք h_i ֆունկցիաներ ունենք՝*

$$h_i : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$$

կամայական $i \in [n]$ համար: Կենթադրենք, որ բոլոր $h_i(v, y)$ ֆունկցիաները, ինչ-որ L դրական հաստատունով Լիպշից հատկությամբ օժտված ֆունկցիաներ են ըստ v-ի կամայական $y \in \mathcal{Y}$ համար: Այդ դեպքում տեղի ունի հետևյալ անհավասարությունը՝

$$\mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^n} \left[\sup_{\substack{f \in \mathcal{F} \\ y \in \mathcal{Y}}} \sum_{i=1}^{n} \sigma_i h_i(f(x_i), y) \right] \leq \sqrt{2} L \mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^{nd}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sum_{j=1}^{d} \sigma_{ij} f_j(x_i) \right] \tag{3}$$

Ապացույց. Սկզբում ցույց տանք, որ բոլոր $i \in [n]$-երի համար և կամայական $g : \mathcal{F} \times \mathcal{Y} \to \mathbb{R}$ ֆունկցիոնալի համար տեղի ունի հետևյալ անհավասարությունը՝

$$\mathop{\mathbb{E}}_{\epsilon \sim \{\pm 1\}} \sup_{\substack{f \in \mathcal{F} \\ y \in \mathcal{Y}}} \epsilon h_i(f(x_i), y) + g(f, y) \leq \sqrt{2} L \mathop{\mathbb{E}}_{\epsilon \sim \{\pm 1\}^d} \sup_{\substack{f \in \mathcal{F} \\ y \in \mathcal{Y}}} \sum_{j=1}^{d} \epsilon_j f_j(x_i) + g(f, y) \tag{4}$$

Դիցուք $\delta > 0$ կամայական դրական թիվ է: Այդ դեպքում համաձայն Ռադեմախերի փոփոխականի սահմանաման կունենանք՝

$$2 \mathop{\mathbb{E}}_{\epsilon \sim \{\pm 1\}} \sup_{\substack{f \in \mathcal{F} \\ y \in \mathcal{Y}}} \epsilon h_i(f(x_i), y) - \delta = \sup_{\substack{f, \bar{f} \in \mathcal{F} \\ y \in \mathcal{Y}}} h_i(f(x_i), y) + g(\bar{f}, y) - h_i(\bar{f}(x_i), y) + g(\bar{f}, y) - \delta$$

Օգտվելով սուպրեմումի սահմանումից՝ գոյություն ունեն $f*, \bar{f}^* \in \mathcal{F}$ ֆունկցիաներ, որ տեղի ունի հետևյալը՝

$$\sup_{\substack{f, \bar{f} \in \mathcal{F} \\ y \in \mathcal{Y}}} h_i(f(x_i), y) + g(\bar{f}, y) - h_i(\bar{f}(x_i), y) + g(\bar{f}, y) - \delta \leq$$
$$\leq \sup_{y \in \mathcal{Y}} h_i(f^*(x_i), y) - h_i(\bar{f}^*(x_i), y) + g(f^*, y) + g(\bar{f}^*, y)$$

Օգտագործելով h_i ֆունկցաի Լիպշիցի հատկությամբ օժտված լինելը կունենանք՝

$$\sup_{y \in \mathcal{Y}} h_i(f^*(x_i), y) - h_i(\bar{f}^*(x_i), y) + g(f^*, y) + g(\bar{f}^*, y) \leq$$
$$\leq L||f^*(x_i) - \bar{f}^*(x_i)|| + \sup_{y \in \mathcal{Y}} g(f^*, y) + g(\bar{f}^*, y)$$

Պնդում 1-ը կիրառելով կստանանք՝

$$L||f^*(x_i) - \bar{f}^*(x_i)|| + \sup_{y \in \mathcal{Y}} g(f^*, y) + g(\bar{f}^*, y) \leq$$
$$\leq \sqrt{2} L \mathop{\mathbb{E}}_{\epsilon \sim \{\pm 1\}^d} \left| \sum_{j=1}^{d} \epsilon_j (f_j^*(x_i) - \bar{f}_j^*(x_i)) \right| + \sup_{y \in \mathcal{Y}} g(f^*, y) + g(\bar{f}^*, y) \leq$$
$$\leq \sqrt{2} L \mathop{\mathbb{E}}_{\epsilon \sim \{\pm 1\}^d} \sup_{f, \bar{f} \in \mathcal{F}} \left| \sum_{j=1}^{d} \epsilon_j f_j(x_i) - \sum_{j=1}^{d} \epsilon_j \bar{f}_j(x_i) \right| + \sup_{y \in \mathcal{Y}} g(f, y) + g(\bar{f}, y)$$

Հեշտ է նկատել, որ կամայական ֆիքսված ϵ-ի դեպքում

$$\sup_{f,\bar{f}\in\mathcal{F}}\left|\sum_{j=1}^{d}\epsilon_j f_j(x_i)-\sum_{j=1}^{d}\epsilon_j \bar{f}_j(x_i)\right| = \sup_{f,\bar{f}\in\mathcal{F}}\sum_{j=1}^{d}\epsilon_j f_j(x_i)-\sum_{j=1}^{d}\epsilon_j \bar{f}_j(x_i)$$

և քանի որ $\text{sup}_{y\in\mathcal{Y}}\, g(f,y)+g(\bar{f},y)$ ինվարիանտ է $f,\bar{f}$ ֆունկցիաների փոփոխման նկատմամբ, կունենանք՝

$$\sqrt{2}L \mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^d}\left|\sum_{j=1}^{d}\epsilon_j(f_j^*(x_i)-\bar{f}_j^*(x_i))\right| + \sup_{y\in\mathcal{Y}} g(f^*,y)+g(\bar{f}^*,y) \leq$$

$$\leq \sqrt{2}L \mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^d}\sup_{f,\bar{f}\in\mathcal{F}}\sum_{j=1}^{d}\epsilon_j f_j(x_i)-\sum_{j=1}^{d}\epsilon_j\bar{f}_j(x_i)+\sup_{y\in\mathcal{Y}} g(f,y)+g(\bar{f},y) =$$

$$= \sqrt{2}L \mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^d}\sup_{f\in\mathcal{F}}\sum_{j=1}^{d}\epsilon_j f_j(x_i)+\sup_{y\in\mathcal{Y}} g(f,y)+\mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^d}\sup_{\bar{f}\in\mathcal{F}}-\sum_{j=1}^{d}\epsilon_j\bar{f}_j(x_i)+\sup_{y\in\mathcal{Y}} g(\bar{f},y)$$

Հաշվի առնելով Ռադեմախերի ϵ_j փոփոխականների սիմետրիկությունը կստանանք՝

$$\sqrt{2}L \mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^d}\sup_{f\in\mathcal{F}}\sum_{j=1}^{d}\epsilon_j f_j(x_i)+\sup_{y\in\mathcal{Y}} g(f,y)+\mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^d}\sup_{\bar{f}\in\mathcal{F}}-\sum_{j=1}^{d}\epsilon_j\bar{f}_j(x_i)+\sup_{y\in\mathcal{Y}} g(\bar{f},y) =$$

$$= 2\left(\sqrt{2}L \mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^d}\sup_{f\in\mathcal{F}}\sum_{j=1}^{d}\epsilon_j f_j(x_i)+\sup_{y\in\mathcal{Y}} g(f,y)\right) =$$

$$= 2\left(\sqrt{2}L \mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^d}\sup_{\substack{f\in\mathcal{F}\\ y\in\mathcal{Y}}}\sum_{j=1}^{d}\epsilon_j f_j(x_i)+g(f,y)\right)$$

Այսպիսով կամայական $\delta > 0$ դրական թվի համար՝

$$\mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}}\sup_{\substack{f\in\mathcal{F}\\ y\in\mathcal{Y}}}\epsilon h_i(f(x_i),y)-\delta \leq \sqrt{2}L \mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^d}\sup_{\substack{f\in\mathcal{F}\\ y\in\mathcal{Y}}}\sum_{j=1}^{d}\epsilon_j f_j(x_i)+g(f,y)$$

Քանի որ վերջինս տեղի ունի ցանկացած δ-ի համար, այստեղից անմիջապես հետևում է 4 անհավասարությունը:

Այժմ ինդուկցիայի միջոցով ցույց տանք, որ ցանկացած $m \in \{0,...,n\}$ համար տեղի ունի հետևյալ անհավասարությունը:

$$\mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^n}\left[\sup_{\substack{f\in\mathcal{F}\\ y\in\mathcal{Y}}}\sum_{i=1}^{n}\epsilon_i h_i(f(x_i),y)\right]\leq\sqrt{2}L\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{md}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sum_{j=1}^{d}\sigma_{ij}f_j(x_i)\right]+$$
$$+\mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^{n-m}}\left[\sup_{\substack{f\in\mathcal{F}\\ y\in\mathcal{Y}}}\sum_{i=m+1}^{n}\epsilon_i h_i(f(x_i),y)\right]$$

3 անհավասարությունը անմիջապես հետևում է՝ վերցնելով $m = n$: Երբ $m = 0$ անհավասարության երկու կողմերում նույն արտահայտությունն է գրված և հետևաբար տեղի ունի անհավասարությունը: Կատարենք ինդուկցիոն ենթադրություն և համարենք անհավասարությունը տեղի ունի $(m-1)$-ի համար, որտեղ $m \leq n$:

$$\mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^n}\left[\sup_{\substack{f\in\mathcal{F}\\ y\in\mathcal{Y}}}\sum_{i=1}^{n}\epsilon_i h_i(f(x_i),y)\right]\leq\sqrt{2}L\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{(m-1)d}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m-1}\sum_{j=1}^{d}\sigma_{ij}f_j(x_i)\right]+$$
$$+\mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^{n-m+1}}\left[\sup_{\substack{f\in\mathcal{F}\\ y\in\mathcal{Y}}}\sum_{i=m}^{n}\epsilon_i h_i(f(x_i),y)\right]=$$
$$=\mathop{\mathbb{E}}_{\substack{\epsilon\sim\{\pm1\}^{n-m}\\ \sigma\sim\{\pm1\}^{(m-1)d}}}\mathop{\mathbb{E}}_{\epsilon_m\sim\{\pm1\}}\left[\sup_{\substack{f\in\mathcal{F}\\ y\in\mathcal{Y}}}\left(\epsilon_m h_m(f(x_m),y)+\sqrt{2}L\sum_{i=1}^{m-1}\sum_{j=1}^{d}\sigma_{ij}f_j(x_i)+\sum_{i=m+1}^{n}\epsilon_i h_i(f(x_i),y)\right)\right]$$

Սահմանենք

$$g(f,y)=\sqrt{2}L\sum_{i=1}^{m-1}\sum_{j=1}^{d}\sigma_{ij}f_j(x_i)+\sum_{i=m+1}^{n}\epsilon_i h_i(f(x_i),y)$$

և տեղադրելով այն վերջինիս մեջ և օգտագործելով 4 անհավասարությունը կստանանք՝

$$\mathop{\mathbb{E}}_{\substack{\epsilon\sim\{\pm1\}^{n-m}\\ \sigma\sim\{\pm1\}^{(m-1)d}}}\mathop{\mathbb{E}}_{\epsilon_m\sim\{\pm1\}}\left[\sup_{\substack{f\in\mathcal{F}\\ y\in\mathcal{Y}}}\left(\epsilon_m h_m(f(x_m),y)+g(f,y)\right)\right]\leq$$
$$\leq\mathop{\mathbb{E}}_{\substack{\epsilon\sim\{\pm1\}^{n-m}\\ \sigma\sim\{\pm1\}^{(m-1)d}}}\mathop{\mathbb{E}}_{\sigma_m\sim\{\pm1\}^d}\left[\sup_{\substack{f\in\mathcal{F}\\ y\in\mathcal{Y}}}\left(\sum_{j=1}^{d}\sigma_{mj}f(x_m)+g(f,y)\right)\right]=$$
$$=\sqrt{2}L\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{md}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sum_{j=1}^{d}\sigma_{ij}f_j(x_i)\right]+\mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^{n-m}}\left[\sup_{\substack{f\in\mathcal{F}\\ y\in\mathcal{Y}}}\sum_{i=m+1}^{n}\epsilon_i h_i(f(x_i),y)\right]$$

□

Թեորեմ 3. *Դիցուք $\mathcal{G}$ ֆուկցիաների բազմությունը, որի յուրաքանչյուր ֆունկցիա Z-ը արտապատկերում է $[0,1]$ և $S = \{z_i\}_{i=1}^m$ m հզորությամբ միմյանցից անկախ և միևնույն բաշխումից ընտրված օրինակների բազմություն է: Այդ դեպքում ցանկացած δ դրական թվի համար առնվազն $1-\delta$ հավանականությամբ բոլոր $g \in \mathcal{G}$ ֆունկցիաների համար տեղի ունի հետևյալ անհավասարությունները՝*

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(\mathcal{G}) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}} \quad (5)$$

և

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_S(\mathcal{G}) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2m}} \quad (6)$$

Դիցուք ունենք միմյանցից անկախ $\mathcal{P}(\mathcal{T} \mid |\mathcal{T}| = k+1)$ բաշխումից ընտրված N հատ առաջադրանքներ՝ $\mathcal{T}_1, ..., \mathcal{T}_N$ և $\mathcal{T} = \cup_{i=1}^{N}\mathcal{T}_i$: Միավորված առաջադրանքի հզորությունը n է՝ $|\mathcal{T}| = n$: Այժմ ենթադրենք միավորված $\mathcal{T}$ առաջադրանքի համար ունենք միմյանցին ակախ և $D_{\mathcal{T}}$ բաշխումից ընտրված M օրինակներ՝

$$S = \{(x_1, y_1), (x_2, y_2), ..., (x_M, y_M) | x_i \in \mathcal{X}, y_i \in \mathcal{T} \text{ և } i \in [M]\}$$

$\mathcal{T}$ առաջադրանքի համար դիցուք $g(x) = Wf(x)$ գծային դասակարգչն է ըստ $f \in \mathcal{F}$ ներկայացման, որտեղ W-ն $(n+1) \times d$ չափանի մատրիցա է և $W \in \mathcal{V}$: $g(x)$ դասակարգչի էմպիրիկ սխալանքը S բազմության վրա սահմանենք հետևյալ կերպ՝

$$\hat{L}(\mathcal{T}, f, W) = \frac{1}{M}\sum_{i=1}^{M} l(\{(Wf(x_i))_{y_i} - (Wf(x_i))_{y_j}\}_{y_i \neq y_j})$$

Ալգորիթմը որով սովորելու ենք ներկայցման ֆունկցիա $\mathcal{F}$ դասից հետևյալն է՝

$$(\hat{f}, \hat{W}) = \underset{\substack{f \in \mathcal{F} \\ W \in \mathcal{V}}}{\text{argmin}}\, \hat{L}(\mathcal{T}, f, W)$$

որտեղ $\hat{f}$ փնտրվող ներկայացումն է: Այսպիսով ալգորթմը ըստ f ներկայացման և գծային դասկարգիչի W մատրիցայի մինիմիզացնում է $\mathcal{T}$ առաջադրանքի վերահսկիչ էմպիրիկ սխալանքը S օրինակների բազմության վրա:

Լեմմա 3. *Դիցուք δ-ն կամայական դրական թիվ է: Այդ դեպքում առնվազն $1-\delta$ հավանականությամբ կամայական $f \in \mathcal{F}$ ներկայացման և կամայական $W \in \mathcal{V}$ մատրիցայի համար տեղի ունի հետևյալ անհավասարությունը՝*

$$L(\mathcal{T}, \hat{f}, \hat{W}) \leq L(\mathcal{T}, f, W) + Gen_M$$

,

Ապացույց. Սահմանենք G ֆունկցիաների բազմությունը հետևյալ կերպ՝

$$G = \left\{(x,y) \mapsto g_{f,W}(x,y) = \frac{1}{B} l(\{[Wf(x)]_y - [Wf(x)]_{y'}\}_{y \neq y'}) | f \in \mathcal{F}, W \in \mathcal{V}\right\}$$

Վերցնենք $Z = \mathcal{X} \times \mathcal{T}$ և $S = \{z_i = (x_i, y_i)\}_{i=1}^M$, կիրառելով 3 թեորեմը G ֆունկցիաների բազմության համար կունենանք՝

$$\mathbb{E}[g(z)] \leq \frac{1}{m} \sum_{i=1}^m g(z_i) + \frac{2}{M} \mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^M} \sup_{\substack{f \in \mathcal{F} \\ W \in \mathcal{V}}} \sum_{i=1}^M \sigma_i g_{f,W}(z_i) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2m}} \tag{7}$$

Այժմ ցույց տանք, որ ցանկացած $W \in \mathcal{V}$ և $i \in [M]$ համար $h_i(f(x_i), W) = g_{f,W}(z_i)$ ֆունկցիան ըստ $f(x_i)$-ի ինչ-որ L հաստատունով օժտված է Լիպշիցի հատկությամբ: Ներմուծենք $\Phi_y(f(x), W)$ ֆունկցիան, այնպես որ $h_i = \frac{1}{B} l \circ \Phi_{y_i}$: Ֆիքսենք որևէ $y \in \mathcal{T}$ դաս և մնացած n դասերը համարակալենք $\mathcal{T} \setminus \{y\} = \{y'_1, y'_2, ..., y'_n\}$: $\Phi_y : \mathbb{R}^d \times \mathcal{V} \to \mathbb{R}^n$ որի տեսքը հետևյալն է՝

$$\Phi_y(x, W) = (w_y x - w_{y'_i} x)_{i \in [n]}$$

Ըստ x փոփոխականի Φ_y ֆունկցիայի Յակոբյանը նշանակենք J_{Φ_y}-ով:

$$J_{\Phi_y} = \begin{pmatrix} \frac{\partial \Phi_{y1}}{\partial x_1} & \frac{\partial \Phi_{y1}}{\partial x_2} & \cdots & \frac{\partial \Phi_{y1}}{\partial x_d} \\ \frac{\partial \Phi_{y2}}{\partial x_1} & \frac{\partial \Phi_{y2}}{\partial x_2} & \cdots & \frac{\partial \Phi_{y2}}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \Phi_{yn}}{\partial x_1} & \frac{\partial \Phi_{yn}}{\partial x_2} & \cdots & \frac{\partial \Phi_{yn}}{\partial x_d} \end{pmatrix} = \begin{pmatrix} w_{y1} - w_{y'_1 1} & w_{y2} - w_{y'_1 2} & \cdots & w_{yd} - w_{y'_1 d} \\ w_{y1} - w_{y'_2 1} & w_{y2} - w_{y'_2 2} & \cdots & w_{yd} - w_{y'_2 d} \\ \vdots & \vdots & \ddots & \vdots \\ w_{y1} - w_{y'_n 1} & w_{y2} - w_{y'_n 2} & \cdots & w_{yd} - w_{y'_n d} \end{pmatrix}$$

$$\begin{aligned} ||J_{\Phi_y}||_F &= \sqrt{\sum_{i=1}^n \sum_{k=1}^d \left(w_{yk} - w_{y'_i k}\right)^2} = \sqrt{n \sum_{k=1}^d w_{yk}^2 - 2 \sum_{i=1}^n \sum_{k=1}^d w_{yk} w_{y'_i k} + \sum_{i=1}^n \sum_{k=1}^d w_{y'_i k}^2} \\ &\leq \sqrt{nQ^2 + 2nQ^2 + nQ^2} = 2Q\sqrt{n} \end{aligned}$$

Այսպիսով Φ_y ֆունկցիան ըստ x-ի փոփոխականի $2Q\sqrt{n}$ հաստատունով Լիպշիցի հատկությամբ օժտված ֆունկցիա է և քանի որ l-ը η հաստատունով Լիպշիցի հատկությամբ էր օժտված, ապա կունենաք որ h_i ֆունկցիաները բոլոր $i \in [M]$ համար $\frac{2\eta Q\sqrt{n}}{B}$ հաստատունով ըստ $f(x_i)$-ի Լիպշիցի հատկություն ունի ցանկացած $W \in \mathcal{V}$ մատրիցայի համար:

Նկատենք որ թեորեմ 2-ի պայմանները բավարարված են և կիրառելով այն կունենանք՝

$$\mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^M} \sup_{\substack{f \in \mathcal{F} \\ W \in \mathcal{V}}} \sum_{i=1}^M \sigma_i g_{f,W}(z_i) \leq \frac{2\sqrt{2}\eta Q\sqrt{n}}{B} \mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^{Md}} \sup_{f \in \mathcal{F}} \sum_{i=1}^M \sum_{j=1}^d \sigma_{ij} f(x_i)$$

Վերջինս տեղադրենք [7](#)-ի մեջ և անհավասարության երկու կողմը բազմապատկենք B-ով, ցանկացած $g \in G$ համար կունենանք՝

$$\mathbb{E}[Bg(z)] \leq \frac{1}{M}\sum_{i=1}^{M} Bg(z_i) + \frac{4\sqrt{2}\eta Q\sqrt{n}}{M} \mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{Md}} \sup_{f\in\mathcal{F}} \sum_{i=1}^{M}\sum_{j=1}^{d} \sigma_{ij} f(x_i) + 3B\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2M}}$$

որտեղից էլ՝

$$L(\mathcal{T}, f, W) \leq \hat{L}(\mathcal{T}, f, W) + \frac{4\sqrt{2}\eta Q\sqrt{n}}{M} \mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{Md}} \sup_{f\in\mathcal{F}} \sum_{i=1}^{M}\sum_{j=1}^{d} \sigma_{ij} f(x_i) + 3B\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2M}} \tag{8}$$

որը տեղի ունի $\forall f \in \mathcal{F}$ և $\forall W \in \mathcal{V}$: Քանի որ 8-ը տեղի ունի $\forall f \in \mathcal{F}$ և $\forall W \in \mathcal{V}$, հետևաբար այն տեղի ունի նաև $\hat{f}$ և $\hat{W}$-ի համար՝

$$L(\mathcal{T}, \hat{f}, \hat{W}) \leq \hat{L}(\mathcal{T}, \hat{f}, \hat{W}) + \frac{4\sqrt{2}\eta Q\sqrt{n}}{M} \mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{Md}} \sup_{f\in\mathcal{F}} \sum_{i=1}^{M}\sum_{j=1}^{d} \sigma_{ij} f(x_i) + 3B\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2M}} \tag{9}$$

Դիցուք $f^*, W^* = \text{argmin}_{f\in\mathcal{F}, W\in\mathcal{V}} L(\mathcal{T}, f, W)$: Կիրառելով Հոֆդինգի անհավասարությունը առնվազն $1 - \frac{\delta}{2}$ հավանականությամբ տեղի ունի հետևյալը՝

$$\hat{L}(\mathcal{T}, f,^* W^*) \leq L(\mathcal{T}, f,^* W^*) + B\sqrt{\frac{\log\frac{2}{\delta}}{2M}}$$

Հաշվի առնելով որ $\hat{L}(\mathcal{T}, \hat{f}, \hat{W}) \leq \hat{L}(\mathcal{T}, f,^* W^*)$՝ 9 անհավասարությունը կարող ենք գրել հետևյալ կերպ՝

$$L(\mathcal{T}, \hat{f}, \hat{W}) \leq L(\mathcal{T}, f,^* W^*) + \frac{4\sqrt{2}\eta Q\sqrt{n}}{M} \mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{Md}} \sup_{f\in\mathcal{F}} \sum_{i=1}^{M}\sum_{j=1}^{d} \sigma_{ij} f(x_i) + 4B\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2M}} \tag{10}$$

Հեշտ է նկատել որ 10 տեղի ունի $\forall f \in \mathcal{F}$ և $\forall W \in \mathcal{V}$ համար՝

$$L(\mathcal{T}, \hat{f}, \hat{W}) \leq L(\mathcal{T}, f, W) + \frac{4\sqrt{2}\eta Q\sqrt{n}}{M} \mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{Md}} \sup_{f\in\mathcal{F}} \sum_{i=1}^{M}\sum_{j=1}^{d} \sigma_{ij} f(x_i) + 4B\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2M}} \tag{11}$$

Կիրառելով պատահույթների միավորման բանաձևը 11 տեղի ունի առնվազն $1 - \delta$ հավանականությամբ և լեմման ապացուցված է:

□

Թեորեմ 4. *Դիցուք δ կամայական դրական թիվ է, այդ դեպքում առնվազն $1-\delta$ հավանականությամբ տեղի ունի հետևյալ անհավասարությունը՝*

$$L(\hat{f}) \leq L(T, f, W) + Gen_{M,n} \; \forall f \in \mathcal{F} \text{ և } \forall W \in \mathcal{V}$$

Որտեղ M ուսուցման օրինակների քանակն է, իսկ N-ը առաջադրանքների քանակը:

Ապացույց. Առաջին հերթին կարելի է հեշտությամբ համոզվել որ դիտարկվող հինջ և լոգիստիկ կորստի ֆունկցիաները բավարարում են հետևյալ հատկությանը՝

$$\forall I \subseteq [t]\; l(\{v_i\}_{i \in I}) \leq l(\{v_i\}_{i \in [t]}) \tag{12}$$

Դիցուք ունենք N հատ միմյանցից անկախ $\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_N$ առաջադրանքները ընտրված

$$\mathcal{P}(\mathcal{T} \mid |\mathcal{T}| = k+1)$$

բաշխումից և $\mathcal{T} = \cup_{i=1}^{N} \mathcal{T}_i$, որի հզորությունը հավասար է $n+1$-ի: □

Գրականություն

[1] John R Firth. *A synopsis of linguistic theory,* 1930-1955. Studies in linguistic analysis, 1957.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. *Distributed representations of words and phrases and their compositionality.* In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 26, pages 3111–3119. Curran Associates, Inc., 2013c.

[3] Karen Sparck Jones. *A statistical interpretation of term specificity and its application in retrieval.* Journal of documentation, 28(1):11–21, 1972.

[4] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. *Exploiting similarities among languages for machine translation.* arXiv preprint arXiv:1309.4168, 2013b.

[5] Sepp Hochreiter and J¨urgen Schmidhuber. *Long short-term memory.* Neural computation, 9(8): 1735–1780, 1997.

[6] Yoshua Bengio, R´ejean Ducharme, Pascal Vincent, and Christian Jauvin. *A neural probabilistic language model.* Journal of machine learning research, 3(Feb):1137–1155, 2003.

[7] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. *Sequence to sequence learning with neural networks.* In Advances in neural information processing systems, pages 3104–3112, 2014.

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural machine translation by jointly learning to align and translate.* arXiv preprint arXiv:1409.0473, 2014.

[9] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, C¸ a glar Gul¸cehre, and Bing Xiang. *Abstractive text summarization using sequence-to-sequence rnns and beyond.* CoNLL 2016, page 280, 2016.

[10] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: *Neural image caption generation with visual attention.* In International Conference on Machine Learning, pages 2048–2057, 2015.

[11] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: *A neural image caption generator.* In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3156–3164, 2015.

[12] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. *Neural architectures for named entity recognition.* In Proceedings of NAACL-HLT, pages 260–270, 2016.

[13] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. *Recursive deep models for semantic compositionality over a sentiment treebank.* In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1631–1642, 2013.

[14] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Rand-walk: *A latent variable model approach to word embeddings.* arXiv preprint arXiv:1502.03520, 2015

[15] Gerard Salton. *The smart retrieval systemexperiments in automatic document processing.* 1971.

[16] Gerard Salton and Christopher Buckley. *Term-weighting approaches in automatic text retrieval. Information processing and management,* 24 (5):513–523, 1988.

[17] John S Breese, David Heckerman, and Carl Kadie. *Empirical analysis of predictive algorithms for collaborative filtering.* In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, pages 43–52. Morgan Kaufmann Publishers Inc., 1998.

[18] Zi Yin, Keng-hao Chang, and Ruofei Zhang. *Deepprobe: Information directed sequence understanding and chatbot design via recurrent neural networks.* In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 2131–2139. ACM, 2017.

[19] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas olov, et al. *Devise: A deep visual-semantic embedding model.* In Advances in neural information processing systems, pages 2121–2129, 2013.

[20] Eliya Nachmani, Elad Marciano, Loren Lugosch, Warren J Gross, David Burshtein, and Yair Beery. *Deep learning methods for improved decoding of linear codes.* arXiv preprint arXiv:1706.07043, 2017.

[21] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. *Indexing by latent semantic analysis. Journal of the American society for information science,* 41(6):391, 1990.

[22] Kenneth Ward Church and Patrick Hanks. *Word association norms, mutual information, and lexicography.* Computational linguistics, 16(1):22–29, 1990.

[23] Yoshiki Niwa and Yoshihiko Nitta. *Co-occurrence vectors from corpora vs. distance vectors from dictionaries.* In Proceedings of the 15th conference on Computational linguistics-Volume 1, pages 304–309. Association for Computational Linguistics, 1994.

[24] Omer Levy and Yoav Goldberg. *Neural word embedding as implicit matrix factorization.* In Advances in neural information processing systems, pages 2177–2185, 2014.

[25] Ronan Collobert and Jason Weston. 2008. *A uni- fied architecture for natural language processing: Deep neural networks with multitask learning.* In Proceed- ings of the 25th International Conference on Machine Learning.

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. *Efficient estimation of word representations in vector space.* ICLR Workshop, 2013.

[27] X. Rong. *word2vec parameter learning explained.* arXiv:1411.2738, 2014. https://arxiv.org/abs/1411.2738

[28] Gutmann, M. and Hyvarinen, A. (2010). *Noise-contrastive estimation: A new estimation principle for unnormalized statistical models.* In Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS'10).

[29] J. Pennington, R. Socher, and C. D. Manning. *GloVe: Global vectors for word representation.* In EMNLP, 2014.