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

Մագիստրոսական Թեզ

Թեմա՝ Բառերի ներդրված վեկտորների ներկայացումների մասին

Ուսանող՝ Մինասյան Գևորգ

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

Երևան – 2019

Բովանդակություն

Ներածություն

Կառուցվածքային կամ համակարգված տվյալների (օրինակ՝ տվյալների բազաների աղյուսակները) հետ աշխատանքը արագ և էֆեկտիվ է համակարգչի միջոցով: Սակայն մարդիկ հաղորդակցվում են միմյանց հետ, օգտագործելով բառեր՝ ձևավորելով ոչ կառուցվածքային տվյալներ: Ոչ կառուցվածքային տվյալների մշակման համար չկան ստանդարտացված մեթոդներ: Առկա ծրագրավորման լեզուների միջոցով սահմանվում է որոշակի կանոների բազմություն, որով աշխատելու է ծրագիրը: Ոչ կառուցվածքային տվյալների համար այս կանոնների բազմությունը բավականին վերացական է և դժվար է այն հստակ սահմանել:

Հազարավոր տարիների ընթացքում մարդու ուղեղը ձեռք է բերել հսկայական փորձ հասկանալու բնական լեզուն: Մարդիկ կարողանում են հասկանալ կարդացած տեքստը և կապել այն իրական աշխարհի հետ, կարողանում են այնտեղ նկարագրվող օբյեկնտերի իրական տեսքի մասին պատկերացում կազմել, զգալ այն հույզերը որն առաջացնում է այդ տեքստի բովանդակությունը: Դեռևս համակարգիչը չի կարողանում բնական լեզուն հասկանալ այնպես ինչպես որ՝ մարդը:

Բնական լեզվի մշակումը արհեստական բանականության ենթաճյուղ է, որն ուսումնասիրում է մշակել և հասկանալ մարդկային լեզուն համակարգիչի միջոցով, մասնավորապես, ինչպես համակարգչային ծրագրերի միջոցով մշակել և վերլուծել մեծ քանակությամբ բնական լեզուների տվյալներ: Բնական լեզվի մշակումը հնարավորություն է տալիս համակարգիչներին կարդալ տեքստը, լսել խոսքը և մեկնաբանել այն:

Մեքենայական ուսուցման ալգորիթմները՝ հատկապես նեյրոնային ցանցերը, լայն տարածում ունեն բնական լեզվի մշակման ոլորտում, որոնցով կարելի է հասնել բարձր արդյունքների բնական լեզվի մշակման բազմաթիվ խնդիրներում, որոնցից են լեզվի մոդելավորումը, քերականական և իմաստաբանական վերլուծությունը, մեքենայական թարգմանությունը և այլն:

Բառերի ներդրված ներկայուցումներ

Բառերի էմբեդինգները ներկայացնում են բառի իմաստը վեկտորի միջոցով: Բազմաթիվ մոդելներով են կառուցվում այդ վեկտորները, որոնց միավորող փիլիսոփայությունը այն է, որ բառի իմաստը որոշվում է կոնտեքստից՝ այն բառերից որոնց հետ միաժամանակ հանդիպում է տեքստում: Ֆիրթի [1] կողմից լեզվաբանության մեջ առաջ է քաշված այսպես կոչված դիստրիբյուշնալ հիպոթեզը այն է՝ *նման կոնտեքստներում հանդիպող բառերը հակված են նման իմաստներ ունենալ:* Դիսկրետ մեծությունների՝ բառերի, արտապատկերումը դեպի էվկլիդյան տարածություն, հնարավոր է դարձնում բառերի միջև կատարել հանրհաշվական գործողություններ: Դիսկրետ բառերի համար գումարման գործողության սահմանում տալը հնարավոր չէ, բայց դիտարկելով դրանց վեկտորական ներկայացումները՝ գործողությունը դառնում է իմաստալից: Ունենալով վեկտորական ներկայացումները՝ հնարավոր է դառնում բառի վեկտորների համար կատարել գծային և ոչ գծային ձևափոխություններ, բառերի միջև հարաբերությունները արտահայտել սկալյար արտադրյալի և կոսինուսի միջոցով: Այս գործողությունների օգտագործմամբ կարելի է ստանալ բառերի միջև իմաստաբանական և ձևաբանական հարաբերությունները [2], դոկումենտների(փաստաթղթերի) միջև նմանությունները [3] և համեմատել տարբեր լեզուների բառարանները [4]: Բազմաթիվ կիրառական խնդիրներ մոդելավորելիս հիմքում օգտագործվում են բառերի վեկտորական ներկայացումները: Ռեկուրենտ նեյրոնային ցանցերը, LSTM [5] ցանցերը բառերի վեկտորական ներկայացումների հետ միասին օգտագործվում են լեզվի մոդելավորման [6], մեքենայական թարգմանության [7, 8], տեքստի համառոտ շարադրման (ամփոփմանշ eng. text summarization) [9] և նկարից տեքստ, վերնագիր ստեղծման (image caption generation) [10, 11] խնդիրներում: Այլ կարևոր կիրառություններ են հատուկ անունների ճանաչումը (named entity recognition) [12], սենտիմենտի որոշումը(sentiment analysis) [13] և գեներատիվ լեզվի մոդելները [14] և այլն: Դիսկրետ մեծությունների վեկտորական ներկայացումները ոչ միայն օգտագործվում են բնական լեզվի մշակման խնդիրներում այլ նաև տեղեկատվական որոնման մեջ (eng. information retrieval) [13, 15, 16], խորհրդատվական համակարգերում [17, 18], նկարների մշակման մեջ [19] և նույնիսկ կոդավորման տեսության խնդիրներում [20]:

Սահմանումներ և նշանակումներ

Աշխատանքում դիտարկելու ենք n բառերից բաղկացած $\mathcal{V} = \{1, 2, ..., n\}$ համարակալված բառարանը: Յուրաքանչյուր i բառի կհամապատասխանացնենք v_i վեկտոր էմբեդինգը: Բոլոր վեկտորները միասին վերցրած կստանանք $E \in \mathbb{R}^{n \times d}$ էմբեդինգ մատրիցան, որի i-րդ տողը i բառի $E_i, . = v_i$ էմբեդինգն է: Համակարգված մեծածավալ տեքստերի հավաքածուն կանվանենք կորպուս և կնշանակենք $\mathcal{C}$-ով, այն իրենից ներկայացնում է ինչ-որ $w_1, w_2, ..., w_T$ բառերի հաջորդականություն: E էմբեդինգ մատրիցայի ուսուցումը կատարվում է որևէ ալգորիթմով, որը մուտքում ստանում է $\mathcal{C}$ կորպուսը:

Ներդրված վեկտորներ և մատրիցային վերլուծություն

Վեկտոր Էմբեդինգների կառուցման սկզբնական պարզագույն մեթոդներում օգտագործվել է մատրիցային վերլուծությունը: Մատրիցային վերլուծությամբ ստացված ներդրված վեկտորները մեքենայական ուսուցման և բնական լեզվի մշակման ալգորիթմների մեծ դաս է, այդ թվում թաքնված սեմանտիկ վերլուծությունը կամ ինդեքսավորումը (LSA/LSI) : Համառոտ ներկայացնենք բացահայտ մատրիցային վերլուծությամբ ստացվող հայտնի մեթոդներից մի քանիսը:

Փաստաթղթի էմբեդինգներ թաքնված սեմանտիկ ինդեքսավորման օգտագործմամբ: Առաջին անգամ Դիրվեստերի [21] և այլոց կողմից ներկայացված LSI մեթոդը փաստաթղթերի վերլուծության և տեղեկատվական որոնման համար հզոր գործիք է: Դիցուք ունենք m հատ $d_1, d_2, ..., d_m$ փաստաթղթերի հավաքածուն, որը կազմված է n բառերից բաղկացած $\mathcal{V}$ բառարանից: Յուրաքանչյուր d_i փաստաթուղթ m_i երկարությամբ ինչ-որ $w_{i1}, w_{i2}, ..., w_{im_i}$ բառերի հաջորդականություն է, որտեղ $w_{ij} \in \mathcal{V}$ և $1 \leq i \leq m, 1 \leq j \leq m_i$: Փաստաթղթերի հավաքածուից կառուցվում է $n \times m$ չափանի բառ-փաստաթուղթ X մատրիցան, որի x_{ij} տարը ամենապարզ դեպքում կարող է լինել i-րդ բառի հանդես գալու քանակը d_j փաստաթղթում: Սակայն քանակների մատրիցան հաճախ կրկնվող բառերի նկատմամբ կարող է բայես պարունակել (անցանկալի առավելություններ տալ) (կողմնակալ կարող է լինել), այս թերությունից խոսափելու համար այսպես կոչված քանակների վերակշռման տարբեր եղանակներ են առաջարկվել, որոնցից ամենահայտնին $TF - IDF$ կոչված մատրիցան է [3, 16]:

$TF - IDF$ վերակշռման եղանակով նվազեցվում է հաճախ հանդիպող ոչ անհրաժեշտ բառերի կշիռները, նախորդ պարզագույն քանակների դեպքում այդ բառերին առավելություն էր տրվում: Այսպիսի բառերի առկայությունը աղավաղում է փաստաթղթերի միջև իրական տարբերությունների հայտնաբերմանը: Այս պատճառով ներմուծվում է նաև այն փաստաթղթերի քանակը, որոնք պարունակում են տվյալ բառը: Օգտագործելով այս լրացուցիչ մեծությունը՝ կարելի է հայտնաբերել, որ բառերն են շատ հաճախ հանդիպում ամբողջ փաստաթղթերում և դրանց տալ փոքր կշիռ:

Դիցուք n_{ij}-ն i-րդ բառի քանակն է d_j փաստաթղթում և n_i-ն այն փաստաթղթերի քանակը, որ պարունակում են i-րդ բառը: Ապա $TF - IDF$ մատրիցայի տարրը հետևյալն է՝

$$x_{ij} = n_{ij} \log_2 \frac{m}{n_i + 1} :$$

Այժմ դիտարկենք X մատրիցայի սինգուլյար վերլուծությունը՝

$$X = UDV^T:$$

Փաստաթղթերի ներդրման վեկտորների E մատրիցան ստացվում է հետևյալ կերպ՝

$$E = f_{\alpha,d}(X) = U_{1:d}D^{\alpha}_{1:d},$$

որտեղ α-ն սովորաբար ընտրվում է 0.5 կամ 1:

Բառերի էմբեդինգներ թաքնված սեմանտիկ վերլուծության օգտագործմամբ: Դիցուք ունենք $\mathcal{C}$ տեքստերի հավաքածուն կազմված $\mathcal{V}$ բառարանից, որտեղ $|\mathcal{V}| = n$: Ինչպես նաև տրված է m հզորությամբ կոնտեքստների $\mathcal{K}$ վերջավոր բազմությունը: Դիտարկենք բառ-կոնտեքստ $n \times m$ չափանի X մատրիցան, որի տողերը ինդեքսավորված են բառերով, սյուները՝ կոնտեքստներով: Ամենապարզ դեպքում x_{ij} տարը, i-րդ բառի և c_j կոնտեքստի համատեղ հայնտվելու քանակն է: Մասնավորապես $\mathcal{K}$ կոնտեքստների բազմությունը կարելի է վերցնել $\mathcal{V}$ բառարանը, այս դեպքում X-ը կլինի $n \times n$ չափանի սիմետրիկ մատրիցա, որի սյունները նույնպես ինդեքսավորված կլինեն բառերով, ինչպես՝ տողերը: Ֆիքսենք որևէ k բնական թիվ՝ անվանելով այն ֆիքսված պատուհանի չափ, որը դիտարկվող $\mathcal{C}$ կորպուսում k հատ հարևան բառերի հաջորդականություն է: Յուրաքանչյուր (i, j) բառազույգի համար x_{ij} տարը այն քանակն է, որ k պատուհանի ներսում միաժամանակ հայտնվում են i և j բառերը:

Պարզագույն քանակներով ստեղծված X մատրիցայի դեպքում, ինչպես թաքնված սեմանտիկ ինդեքսավորման մոդելում, այնպես էլ այս դեպքում հաճախ կրկնվող բառերին անցանկալի առավելություններ են տրվում: Վերոնշյալ խնդիրը վերանում է, երբ X մատրիցայի վերակշռված տարբերակներն են ընտրվում, որոնցից առավել հայտնի են PMI [22], դրական PMI [23] և տեղաշարժված PMI մատրիցաները [23]:

PMI(Pointwise Mutual Information) համեմատական չափ է, այն ցույց է տալիս x և y պատահույթների միաժամանակ հանդես գալու հավանականությունը՝ այն համեմատելով այդ մեծությունների անկախ լինելու դեպքում սպասվող հավանականության հետ՝

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}:$$

Դիտարկվող i բառի և c կոնտեքստի միջև PMI մեծությունը սահմանվում է հետևյալ կերպ՝

$$PMI(i, c) = \log_2 \frac{P(i, c)}{P(i)P(c)}:$$

Համարիչը ցույց է տալիս երկու բառերի տեքստում միասին հայտնվելը, իսկ հայտարարը՝ միասին հայտնվելու սպասվող քանակը, եթե ենթադրենք նրանց անկախությունը: Այսպիսով վերոնշյալ հարաբերությունը գնահատում է, թե ինչքան հաճախ են երկու բառեր միասին հայտնվում, քան կհայտնվեին պատահաբար: PMI-ի միջոցով կարելի է հայտնաբերել միմյանց հետ ուժեղ ասոցածված բառերը:

PMI արժեքները փոփոխվում են բացասական անվերջությունից դրական անվերջություն, սակայն բացասական PMI-ը, որը նշանակում է բառերը ավելի քիչ են միմյանց հետ հայտնվում, քան պատահաբար կհայտնվեին, պահանջում է աշխատել հսկայական չափի տեքստերի հավաքածուի՝ կորպուսի, հետ: Այս պատճառով սովորաբար օգտագործվում է $PPMI$ վերակշռման եղանակը՝

$$PPMI(i,c) = \max\left(\log_2 \frac{P(i,c)}{P(i)P(c)}, 0\right):$$

Դիցուք n_{ij}-ն i բառի c_j կոնտեքստում հայտվելու քանակն է, այդ դեպքում՝

$$p_{ij} = \frac{n_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} n_{ij}} \quad p_i = \frac{\sum_{j=1}^{m} n_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} n_{ij}} \quad q_j = \frac{\sum_{i=1}^{n} n_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} n_{ij}}:$$

Օգտագործելով p_{ij}, p_i և q_j թվերը՝ PMI և PPMI մատրիցաների էլեմենտները տրվում են հետևյալ կերպ՝

$$PMI_{ij} = \log_2 \frac{p_{ij}}{p_i q_j} \quad PPMI_{ij} = \max\left(\log_2 \frac{p_{ij}}{p_i q_j}, 0\right):$$

Այժմ ենթադրենք UDV^T պարզագույն քանակների X, վերակշռված PMI կամ PPMI մատրիցաներից որևէ մեկի սինգուլյար վերլուծությունն է, ապա ներդրված վեկտորների E մատրիցան տրված d չափողականության և α հիպերպարամետրի դեպքում հաշվվում է հետևյալ կերպ՝

$$E = f_{\alpha,d}(X) = U_{1:d} D_{1:d}^{\alpha}:$$

α հիպերպարամետրը կարող է կամայական թիվ լինել $[0,1]$ ինտերվալից, որը սիմետրիկության համար սովորաբար ընտրվում է 0.5, սակայն $\alpha = 0$ կամ $\alpha = 1$ դեպքերը նույնպես հաճախ են ընտրվում:

Ներդրված վեկտորների կառուցման արդի ալգորիթմներ

Նեյրոնային ցանցով լեզվի մոդելները [6, 25], որոնք ոչ գծային և ոչ ուռուցիկ մեթոդներ են, առաջարկում են ներդրված վեկտորների կառուցման մեկ այլ եղանակ՝ բառի ներդրված վեկտորը պարզապես ցանցի ներքին ներկայացումն է տվյալ բառի համար: Բառերի ներդրված վեկտորների և լեզվի վիճակագրական մոդելի ուսուցումը կատարվում է միաժամանակ: Այս մոտեցմամբ ներդրված վեկտորների կառուցման համար նեյրոնային ցանցերի տարբեր կառուցվածքներ են առաջարկվել՝ պարզագույն բազմաշերտ ցանցերից մինչև ռեկուրենտ ցանցեր: Սակայն 2013 թվականին Միքայիլովի և այլոց կողմից [26] առաջարկված լոգ-գծային երկու մոդելները արդյունավետությամբ գերազանցեցին նախորդ բոլոր բարդ կառուցվածք ունեցող ցանցերին և ամենակարևորը ժամանակային բարդությունը արմատապես նվազեցվեց: Հնարավոր եղավ օգտագործել շատ ավելի մեծ տեքստերի հավաքածու և ավելի ճշգրիտ ներդրված վեկտորներ ստանալ, որոնք հուսալիորեն կարիելի է օգտագործել տարբեր տեսակի բնական լեզվի մշակման մոդելների հիմքում: Ստորև համառոտ ներկայացնենք ներդրված վեկտորների կառուցման հիմնական մեթոդները:

Word2vec մեթոդների ընտանիք

Word2vec մեթոդների ընտանիքում մոդելների երկու տարբերակ է առաջարկվել: Երկու տարբերակում էլ բառերի ներդրված վեկտորները սկզբնարժեքավորվում են պատահականորեն: Այնուհետև հերթականորեն դիտարկվում է տեքստերի հավաքածուն՝ ֆիքսված պատուհանի չափով յուրաքանչյուր բառի համար մոտակա բառերը համարվում են դիտարկվող բառի կոնտեքստը: Ընդհանուր դեպքում ստոխաստիկ գրադիենտային վայրէջքի օպտիմիզացիոն մեթոդով մինիմիզացվում է հերթական բառի և կոնտեքստ բառերի միջև սկալյար արտադրյալը: Ամեն անգամ, երբ երկու բառեր հանդիպում են նման կոնտեքստում, սկալյար արտադրյալի փոքրացման շնորհիվ նրանց միջև կապը ուժեղանում է, հեռավորությունը՝ փոքրանում: Սակայն միայն փոքրացնելով նման կոնտեքստներում հանդիպող բառերի վեկտորների միջև հեռավորությունը՝ հանգում ենք հետևյալ խնդրին: Անսահմանափակ տեքստերի հավաքածուի դեպքում մինիմալ վիճակը կլինի այն, որ բոլոր վեկտորները հավասարվելու են կամ գտնվելու են միևնույն դիրքում, որը ակնհայտորեն ցանկալի չէ: Այս խնդիրը շրջանցելու նպատակով *word2vec*-ում սկզբում ներկայացվում է հիերարխիկ սոֆթմաքսը [27], իսկ ավելի ուշ negative sampling-ը:

Վերջինս ավելի պարզ և արդյունավետ է: Երբ յուրաքանչյուր անգամ նման կոնտեքստում հանդիպող բառերի միջև հեռավորությունները փոքրացվում են, նախապես տրված քանակով պատահական բառեր են ընտրվում և դիտարկվող բառից հեռավորությունները՝ մեծացվում: Այս կերպ ապահովվում է ոչ նման բառերի միմյանցից մեծ հեռավորության վրա գտնվելը: *word2vec*-ում առաջարկված մոդելների երկու տարբերակներն են՝

1. *Կոնտեքստից կանխատեսել դիտարկվող բառը:* w_t բառի կանխատեսումը կատարվում է օգտագործելով՝ այդ բառի ֆիքսված k պատուհանի շրջակայքում գտնվող բառերը՝ $w_{t-k}, w_{t-k+1}, ..., w_{t-1}, w_{t+1}, ..., w_{t+k-1}, w_{t+k}$:

2. *Դիտարկվող բառից կանխատեսել կոնտեքստը:* $w_{t-k}, w_{t-k+1}, ..., w_{t-1}, w_{t+1}, ..., w_{t+k-1}, w_{t+k}$ կոնտեքստի կանխատեսումը կատարվում է՝ օգտագործելով w_t դիտարկվող բառը:

Դիցուք ունենք $w_1, w_2, ..., w_T$ հաջորդական բառերով կազմված կորպուսը և որևէ ֆիքսված k պատուհանի չափ: c_t-ով նշանակենք w_t բառի կոնտեքստը՝

$$c_t \stackrel{\text{def}}{=} w_{t-k}, w_{t-k+1}, ..., w_{t-1}, w_{t+1}, ..., w_{t+k-1}, w_{t+k},$$

որտեղ $k+1 \leq t \leq T-k$: w բառի ներդրված վեկտորը նշանակենք՝ v_w-ով, իսկ կոնտեքստ վեկտորը՝ u_w-ով: c_t կոնտեքստին համապատասխան բառերի վեկտորների միջինը նշանակենք՝

$$u_{c_t} \stackrel{\text{def}}{=} \frac{1}{2k} \sum_{-k \leq j \leq k} u_{t+j}:$$

Վերոնշյալ կոնտեքստից բառի կանխատեսման դեպքում կոնտեքստի բառերի վեկտորները գումարվում են և օգտագործվում դիտարկվող բառի կանխատեսման համար: Նպատակային կորստի ֆունկցիան, որը պետք է մաքսիմիզացնել, հետևյալն է՝

$$\frac{1}{T-2k} \sum_{t=k+1}^{T-k} \log p(w_t|c_t),$$

որտեղ $p(w_t|c_t)$ հավանականությունը սահմանվում է սոֆթմաքս ֆունկցիայի միջոցով՝

$$p(w_t|c_t) = \frac{e^{v_{w_t}^T u_{c_t}}}{\sum_{w \in V} e^{v_w^T u_{c_t}}}:$$

Բառից կոնտեքստի կանխատեսման դեպքում կոնտեքստի յուրաքանչյուր բառ կանխատեսվում է անկախ, երբ տրված է դիտարկվող բառը՝

$$\frac{1}{T-2k} \sum_{i=k+1}^{T-k} \sum_{-k \leq j \leq k} \log p(w_{t+j}|w_t),$$

որտեղ $p(w_{t+j}|w_t)$ հավանականությունը սահմանվում է հետևյալ կերպ՝

$$p(w_{t+j}|w_t) = \frac{e^{u_{w_{t+j}}^T v_{w_t}}}{\sum_{w \in V} e^{u_w^T v_{w_t}}}:$$

Սակայն սոֆթմաքսի օգտագործումը կորստի ֆունկցիայում շատ ժամանակատար է՝ գրադիենտի հաշվման գործողությունների քանակը համեմատական է V բառարանի չափին, որը տարբեր կորպուսների դեպքում տասնյակ հազարաների կարող է հասնել, երբեմն նույնիսկ միլիոնների: Միքայիլովի աշխատանքում երկու տարբեր լուծումներ է առաջարկվում ալգորիթմի արագացման համար: Առաջինը հիերարխիկ սոֆթմաքսի օգտագործումն է, որի օգնությամբ գնահատվում է սոֆթմաքս ֆունկցիան և բարդությունը $O(\log_2 |V|)$ է: Երկրորդ առավել հայտնի լուծման այլընտրանքը *բացասական օրինակների ընտրության* [28] եղանակն է: Բացասական օրինակների ընտրության դեպքում կոնտեքստից բառի կանխատեսման մոդելի համար յուրաքանչյուր w_t ընթացիկ բառի համար հետևյալ նպատակային ֆունկցիան է մաքսիմիզացվում՝

$$\log \sigma(v_{w_t}^T u_{c_t}) + \sum_{i=1}^{s} \log \sigma(-v_{\tilde{w}_i}^T u_{c_t}), \quad \sigma(x) = \frac{e^x}{1+e^x}$$

որտեղ $\sigma(x)$-ը կոչվում է լոգիստիկ ֆունկցիա, s-ը բացասական օրինակների քանակն է և որպես հիպերպարամետր նախապես տրված է լինում, իսկ $\tilde{w}_i$ բացասական օրինակները ընտրվում են նախօրոք ֆիքսված $P(w)$ հավանականային բաշխումից, մասնավորապես այն կարող է լինել դիսկրետ հավասարահավանական բաշխումը: Բառից կոնտեքստի կանխատեսման մոդելի դեպքում w_t ընթացիկ բառի կորստի ֆունկցիան բացասական օրինակների ընտրության դեպքում հետևյալն է՝

$$\log \sigma(u_{w_{t+j}}^T v_{w_t}) + \sum_{i=1}^{s} \log \sigma(-u_{\tilde{w}_i}^T v_{w_t}):$$

Գլոբալ վեկտորներ բառերի ներկայացման համար

Բառերի վեկտորական ներկայացման հաջորդ հայտնի *glove* կոչված մոդելը ներկայացվել է 2014 թվականին Պենինգթոնի [28] և այլոց կողմից: Բառերի ներդրված վեկտորների ուսուցման մոդելները կարելի է բաժանել երկու ընտանիքների՝ գլոբալ մատրիցային վերլուծություն օգտագործող մեթոդներ (օրինակ. թաքնված սեմանտիկ վերլուծությունը) և լոկալ կոնտեքստ օգտագործող մեթոդներ (օրինակ. *word2vec*): Այս երկու ընտանիքներում առկա մեթոդները զգալի թերություններ ունեն: Լոկալ կոնտեքստ օգտագործող մեթոդները, ինչպիսին է *word2vec* մեթոդների ընտանիքը, անալոգ բառերի հայտնաբերման առաջադրանքում ճշգրիտ է աշխատում, սակայն կորպուսի բառերի համատեղ հայտվելու գլոբալ քանակները չի օգտագործում, փոխարենը այս մեթոդների ուսուցումը կատարվում է առանձին լոկալ պատուհաններում հայտնված բառերով: Մինչդեռ թաքնված սեմանտիկ վերլուծության նման մոդելները, որոնք օգտագործում են կորպուսի բառերի համատեղ հայտնվելու գլոբալ քանակները, համեմատաբար ճշգրտությամբ զիջում են *word2vec* ընտանիքի մեթոդներին: *Glove* մոդելով ներդրված վեկտորների կառուցման ալգորիթմը գործում է գլոբալ քանակների վրա, բայց ներդրված վեկտորական տարածությանը ունենում է *word2vec*-ի պես ճշգրիտ իմաստաբանական կառուցվածք: Բառերի վեկտորական ներկայացումներ կառուցելու համար կոնտեքստում բառերի միասին հայտնվելու քանակական մեծությունները հիմնական ինֆորմացիայի աղբյուրն է, սակայն թե ինչպես է քանակներից բառի իմաստը գեներացվում դեռևս բաց հարց է: Այս մոդելով բացատրվում է ներդրված վեկտորների իմաստաբանական կառուցվածք ունենալը՝ բացահայտ օգտագործելով կորպուսի գլոբալ քանակները որպես իմաստի գեներացման ինֆորմացիայի հիմնական աղբյուր: Դիցուք X-ը բառերի համատեղ հայտնվելու քանակների մատրիցան է, որի x_{ij} տարը j-րդ բառի հայնվելու քանակն է i-րդ բառի կոնտեքստում: $x_i = \sum_k x_{ik}$ այն քանակն է, որ կամայական բառ հայտնվում է i-րդ բառի կոնտեքստում: Այժմ՝

$$\hat{p}_{ij} = \frac{x_{ij}}{x_i}$$

i-րդ բառի կոնտրեքստում j-րդ բառի հայտնվելու հավանականության գնահատականն է: Դիտարկենք մի պարզ օրինակ, որը ցույց է տալիս, թե բառի իմաստային ինչպիսի առանձնահատկություններ են կրում համտեղ հայտնվելու հավականության գնահատականները: Վերցնենք $i = ice$, $j = steam$ բառազույգը, որոնց համար դիտարկելով տարբեր բառերի հետ համատեղ հայտվելու հավանականությունները՝ բառազույգի հարբերության մասին պատկերացում կարելի է կազմել: Այն բառերը, որոնք կապված են «ice» բառի հետ և կապված

չեն «steam» բառի հետ, օրինակի համար $k = solid$, այս դեպքում $\hat{p}_{ik}/\hat{p}_{jk}$ հարաբերությունը կլինի մեծ: Նմանապես $k = gas$ վերցնելով՝ որը «steam» բառի հետ է կապված, բայց ոչ «ice» բառի, վերոնշյալ հարաբերությունը կլինի փոքր: Այն բառերը որոնք կապված չեն ո՛չ «ice» բառի, ո՛չ «$steam$» բառի հետ, ինչպիսիք են «$water$» և «$fashion$» բառերը, հարաբերությունը մեկին մոտ է լինելու: Աղյուսակ 1-ում բերված են հավանականությունների գնահատականները և նկարագրված հարբերությունները, որոնք հաստատում են վերոնշյալ ենթադրությունները:

Հավանականություն և հարաբերություն	k = solid	k = gas	k = water	k = fashion
$p(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$p(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$p(k\|ice)/p(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Աղյուսակ 1: «ice» և «steam» բառերի և ընտրված կոնտեքստ բառերի հետ համատեղ հայտնվելու հավանականային գնահատականները ստացված շուրջ 6 միլիարդ բառեր պարունակող տեքստերի հավաքածուից:

Ուսումնասիրելով աղյուսակ 1-ի թվերը, պարզ է դառնում, որ բառերի համատեղ հայտնվելու հավանականաթյունների հարաբերությունները ավելի լավ են տարբերակում դիտարկվող բառի հետ կապված բառերը, ոչ կապված բառերից: Այս հանգամանքը հաշվի առնելով՝ բառերի ներդրված վեկտորների ուսուցման մոդելում ավելի նպատակահարմար է օգտագործել բառերի համատեղ հայտնվելու հավանականությունների հարբերությունները, այլ ոչ թե միայն հավանականությունները: Նկատենք $\frac{\hat{p}_{ik}}{\hat{p}_{jk}}$ հարաբերությունը կախված i, j և k բառերից, ընդհանուր դեպքում ներդրված վեկտորների ուսուցման մոդելը ունի հետևյալ տեսքը՝

$$F(v_i, v_j, \widetilde{v}_k) = \frac{\hat{p}_{ik}}{\hat{p}_{jk}}, \tag{1}$$

որտեղ $v_i, v_j \in \mathbb{R}^d$ համապատասխանաբար i և j բառերի ներդրված վեկտորներն են, իսկ $\widetilde{v}_k \in \mathbb{R}^d$ k բառի կոնտեքստ վեկտորն է: Վերոնշյալ հավասարումն մեջ դեռևս չսահմանված F ֆունկցիան պետք է կիրառել բառերի ներդրված վեկտորների և կոնտեքստ վեկտորների նկատմամբ, որի արդյունքը պետք է մոտարկի $\frac{\hat{p}_{ik}}{\hat{p}_{jk}}$ հարաբերությանը: F ֆունկցիայի ընտրության տարբերակները բազմաթիվ են, սակայն որոշ հատկություններ հաշվի առնելով՝ այդ ընտրությունը դառնում է միակը: Հատկություններից մեկը այն է, որ $\frac{\hat{p}_{ik}}{\hat{p}_{jk}}$ հարաբերության ինֆորմացիան պետք է ներառել բառերի ներդրված վեկտորների տարածությունում: Քանի որ վեկտորական տարածություններին

բնորոշ են գծային կառուցվածքները, ապա հարբերության ինֆորմացիան կարելի է ներառել վեկտորների տարբերության միջոցով: Այս նպատկով, կարելի է դիտարկել այն ֆունկցիաները որոնք կախված ենք դիտարկող երկու բառերի ներդրված վեկտորների տարբերությունից՝

$$F(v_i - v_j, \widetilde{v}_k) = \frac{\hat{p}_{jk}}{\hat{p}_{jk}}: \tag{2}$$

F ֆունկցիան կարելի է ընտրել բարդ պարամետրացված ֆունկցիաների ընտանիքից (օրինակ նեյրոնային ցանց), բայց այդ դեպքում ներդրված վեկտորական տարածությունում գծային կառուցվածքները չեն պահպանվելու: Գծային կառուցվածքների պահպանման նպատակով F ֆունկցիայի վեկտոր արգումենտները փոխարինվում է դրանց սկալյար արտադրյալով՝

$$F((v_i - v_j)^T \widetilde{v}_k) = \frac{\hat{p}_{ik}}{\hat{p}_{jk}}:$$

Եթե պահանջենք F-ը լինի հոմոմորֆիզմ $(\mathbb{R}, +)$ և $(\mathbb{R}_{>0}, \times)$ խմբերի միջև, ապա կունենանք հետևյալը՝

$$F((v_i - v_j)^T \widetilde{v}_k) = \frac{F(v_i^T \widetilde{v}_k)}{F(v_j^T \widetilde{v}_k)} \tag{3}$$

$$F(v_i^T \widetilde{v}_k) = \hat{p}_{ik} = \frac{x_{ik}}{x_i}: \tag{4}$$

Ակնհայտ է e^x ցուցչային ֆունկցիան հանդիսանում է վերոնշյալ հոմոմորֆիզմի լուծումը և այս դեպքում $v_i^T \widetilde{v}_k$ սկալյար արտադրյալը կունենա հետևյալ տեսքը՝

$$v_i^T \widetilde{v}_k = \log \hat{p}_{ik} = \log x_{ik} - \log x_i: \tag{5}$$

Նկատենք որ բառ-կոնտեքստ X մատրիցայում կարելի է ազատորեն փոխել բառերի և կոնտեքստների ինդեքսավորման առանցքները, քանի որ այդ մատրիցան սիմետրիկ է, հետևաբար նաև v բառ վեկտորի և $\widetilde{v}$ կոնտեքստ վեկտորների նշանակությունների փոփոխման նկակտմաբ պետք է ինվարիանտ լինի ներդված վեկտորների կառուցման մոդելը: 5 հավասարումը կարելի է ավելի պարզեցնել՝ քանի որ $\log x_i$ անդամը k բառից կախված չէ, ապա այն կարելի է փոխարնել b_i պարամետրով i բառի համար, ինչպես նաև հաշվի առնելով վերոնշյալ սիմետրիկությունը, ավելացնել նաև k կոնտեքստի համար $\widetilde{b}_k$ պարամետերը:

$$v_i^T \widetilde{v} + b_i + \widetilde{b}_k = \log x_{ik} \tag{6}$$

6 հավասարումը չի կարելի վերջնական համարել, քանի որ $\log x_{ik}$ որոշված չէ, երբ i և k բառազույգը կորպուսում ֆիքսված պատուհանի ներսում չի հայտնվել՝ $x_{ik} = 0$: Այս խնդիրը կարելի է շրջանցել կատարելով միվոր տեղաշարժ լոգարիթմում՝ $\log(x_{ik} + 1)$: Հաճախ հանդիպող

բառազույգերին անցանկալի առավելություններ չտալու նպատակով ներմուծվում վերակշռման $f(x_{ik})$ ֆունկցիան: f-ի ընտրության տարբերակները բազմաթիվ են, հեղինակները ընտրել են հետևյալ ֆունկցիան՝

$$f(x) = \begin{cases} \left(\frac{x}{x_{max}}\right)^{\alpha} & x < x_{max} \\ 1 & \text{հ.դ.} \end{cases} \tag{7}$$

Փորձնական ճանապարով հիպերպարամետրերը ընտրվել են՝ $x_{max} = 100$ և $\alpha = \frac{3}{4}$: $Glove$ մոդելում ներդրված վեկտորների կառուցման համար, 6 բանաձևի օգտագործմամբ մինիմիզացվում է կշռված փոքրագույների քառակուսացման հետևյալ կորստի ֆունկցիան՝

$$J = \sum_{i,j=1}^{n} f(x_{ij})(v_i^T \widetilde{v_j} + b_i + \widetilde{b_j} - \log(x_{ij} + 1))^2: \tag{8}$$

Բառերի ներդրված ներկայացումների որոշ տեսական հիմնավորումներ

[24] աշխատանքում փորձնական և որոշ տեսական հիմնավորումների միջոցով ցույց է տրվում, որ բառերի ներդրված վեկտորների արդի մեթոդները կապված են ավելի հին PMI մատրիցայի վրա հիմնված մոդելների հետ: PMI մատրիցայի պարզագույն տարբերակում դիտարկվում է սիմետրիկ մատրիցա, որի յուրաքանչյուր տող և սյուն ինդեքսավորված է բառերով: Այդ սիմետրիկ մատրիցայի (w, w') ինդեքսով տարրն է՝

$$PMI(w, w') = \log \frac{p(w, w')}{p(w)p(w')}, \tag{9}$$

որտեղ $p(w, w')$-ն այն էմպիրիկ հավանականությունն է, որ w և w' բառերը տեքստերի հավաքածուում կհայտնվեն որոշակի չափով պատուհանի ներսում, իսկ $p(w)$-ն այն հավանականությունն է որ տեքստերի հավաքածուից պատահական ընտրված բառը w-ն է: Այնուհետև բառերի ներդրված վեկտորները ստացվում են վերոնշյալ PMI մատրիցայի սինգուլյար վերլուծության միջոցով: Փորձնական եղանակով պարզվում է PMI մատրիցան մոտարկվում է փոքր ռանգի մատրիցայով: Գոյություն ունեն բառերի ներդրված վեկտորներ, որոնց չափը շատ փոքր է համեմատած դիտարկվող բառարանում պարունակող բառերի քանակի հետ և որոնց համար՝

$$v_w^T v'_w \approx PMI(w, w'): \tag{10}$$

[14] աշխատանքում առաջարվում է տեքստ գեներացնելու հավանականային նոր մոդել, որտեղ դուրս բերված բանաձևը, որը տեղի ունի որոշակի նախնական ենթադրությունների բավարարման դեպքում, ուղակիորեն բացատրում է 10 մոտարկումը: Կարևոր ենթադրություններից մեկն այն է որ, *բառերի ներդրված վեկտորները տարածական իզոտրոպ են, ինչը նշանակում է, որ վեկտորները չունեն նախընտրելի ուղղություն տարածությունում*: n վեկտորներ d չափանի տարածությունում իզոտրոպ լինելու համար անհրաժեշտ է $d << n$: Բացի այդ այս մոդելով տեսականորեն բացատրվում է անալոգ բառերի հայտնաբերման խնդրի արդյունավետ լուծումը բառերի ներդրված վեկտորների միջոցով՝ կատարելով գծային հանրահաշվի պարզագույն վեկտորական գործողություններ: Այն որոշակիորեն միավորում է ներդրված վեկտորների կառուզման բազմաթիվ մոդելներ և մեթոդների ընտանիքներ, մասնավորապես word2vec մեթոդների ընտանիքը, $glove$ մոդելը և այլն, բացի այդ բառերի ներդված վեկտորների օգտագործմամբ

բազմաթիվ խնդիրներ տեսական հիմնավորումներ են ստանում, որնցից են անալոգ բառերի հայտնաբերումը, բազմիմաստ բառերի ներդրված վեկտորների և դրանց տարբեր իմաստների ներդրված վեկտորների միջև կապը: Տեսականորեն բացատրվում է, թե ինչպես բազմիմաստ բառը կարել է ստանալ տարբեր իմաստներին համապատասխան ներդրված վեկտորների գծային սուպերպոզիցիայի միջոցով: Տեքստերի հավաքածուի՝ կորպուսի գեներացիան մոդելը դիտարկում է որպես դինամիկ պրոցես, որտեղ t-րդ բառը ծնվում է t-րդ քայլի ժամանակ: Պրոցեսը պայմանավորված է c_t դիսկուրս վեկտորի պատահական քայլով, որի կորդինատները ցույց են տալիս, թե ինչի մասին է խոսվում տվյալ պահին: Յուրաքանչյուր բառ ունի t ժամանակից անկախ, թաքնված մի վեկտոր, որը իր և դիսկուրս վեկտորի հետ կորելացիաներ է պարունակում: Նկարագրված ենթադրությունները մոդելավորում է լոգ-գծային գեներատիվ մոդելով՝

$$p\left(w_t|c_t\right) \propto e^{c_t^T v_w} \tag{11}$$

Վերոնշյալ c_t դիսկուրս վեկտոր կատարում է փոքր պատահական քայլ (c_{t+1} վեկտորը ստացվում է c_t-ին գումարելով փոքր տեղաշարժի վեկտոր՝ $c_{t+1} = c_t + \delta_t$), այնպես որ մոտիկ բառերը գեներացվեն նմանատիպ դիսկուրս վեկտորներից: Քանի որ մոդելը պահպանելու է կորպուսում բառազույգերի համատեղ հայտնվելու հավանականությունները, ապա դիսկուրս վեկտորին պատահական բայց ոչ մշտական մեծ տեղաշարժեր թույլատրված են, քանի որ դրանց ազդեցությունը այդ հավանականությունների վրա աննշան է լինելու:

n-ով նշանակենք բառերի քանակը և d-ով՝ դիսկուրս վեկտորի տարածության չափը, որտեղ $1 \leq d \leq n$: Նախնական ենթադրենք, որ ինչ որ տիրույթում բառերի ներդված վեկտորները հավասարաչափ են բաշխված, որը Բայեսյան վիջակագրությունում հայտնի նախնական ենթադրությունն է տվյալների հավանականային բաշխման վերաբերյալ: Ըստ այս նախնական ենթադրության՝ բառերի ներդրված վեկտորները գեներացվել են միմյանցից անկախ և նույնական $v = s\hat{v}$, պատահական մեծությունների արտադրյալին համապատասխան բաշխումից, որտեղ $\hat{v}$-ն սֆերային նորմալ բաշխման պատահական մեծությունն է, իսկ s-ը սկալյար պատահական մեծություն է, որի մաթսպասումն է τ և միշտ վերևից սահմանափակ է κ-ով: Հարկ է նշել, որ s-ի ընտրությունից է կախված մոդելավորման իրատեսական լինելը, սակայն այն այդքան էլ կարևոր չի տեսական հիմնավորումների համար: Դիսկուրս վեկտորի պատահական քայլը հավասարաչափ է բաշված $\mathcal{C}$ միավոր սֆերայում: Պատահական քայլի երկարությունը l_2 նորմով ամենաշատը $\frac{\epsilon_2}{d}$ է: Նկարագրված ենթադրությունների դեպքում 11 հավասարման նորմալաիզացիայի $Z_c = \sum_w e^{v_w^T c}$ գործակիցը կոնցենտրացվում է որևէ Z հաստատունի շրջակայքում, որն ամփոփված է հետևյալ լեմմայում՝

Լեմմա 1 ([14]). *Ներդրված վեկտորների վերոնշյալ Բայեսյան ենթադրությունների բավարարման դեպքում՝*

$$p((1-\epsilon_z)Z \leq Z_c \leq (1+\epsilon_z)Z) \geq 1-\delta, \tag{12}$$

որտեղ $\epsilon_z = \widetilde{O}(\frac{1}{\sqrt{n}})$ *և* $\delta = e^{-\Omega(\log^2 n)}$*:*

Z_c նորմալիզացիոն գործակցի հաստատունի շուրջ կոնցենտրացիայի դեպքում, հնարավոր է լինում բանաձևներ ստանալ կորպուսում w բառի $p(w)$ և w, w' բառերի ֆիքսված q չափանի պատուհանի ներսում հայտնվելու $p(w, w')$ հավանականությունների մոտարկման համար: Բանաձևերի տեսքը ձևակերպված է հիմնական աշխատանքի հիմնական թեորեմում՝

Թեորեմ 1 ([14]). *Եթե բառերի ներդրված վեկտորները բավարարում են 12 անհավասարությանը և ֆիքսված պատուհանի չափը* q *է, ապա*

$$\log p_q(w, w') = \frac{||v_w + v_{w'}||_2^2}{2d} - 2\log Z + \gamma \pm \epsilon \tag{13}$$

$$\log p(w) = \frac{||v_w||_2^2}{2d} - \log Z \pm \epsilon \tag{14}$$

$$PMI_q(w, w') = \frac{v_w^T v_{w'}}{d} + \gamma \pm O(\epsilon), \tag{15}$$

որտեղ $\epsilon = O(\epsilon_z) + \widetilde{O}(\frac{1}{d}) + O(\epsilon_2)$ *և* $\gamma = \log\left(\frac{q(q-1)}{2}\right)$*:*

Բառերի ներդրված վեկտորների բաշխման վերբերյալ կատարված Բայեզյան նախնական ենթադրությունները կարելի է մեղմացնել և փոխարինել այդ վեկտորների որոշակի հատկությունննով: Կարելի է ենթադրել, որ գոյություն ունեն բառերի ներդված վեկտորներ, որոնք գտնվում են որևէ տիրույթում, որտեղ այս վեկտորները տարածական իզոտրոպ են հետևյալ կերպ: *Գրեթե բոլոր միավոր* $c \in \mathcal{C}$ *վեկտորների համար* $Z_c = \sum_w e^{v_w^T c}$ *գումարը շատ մոտ է ինչ-որ* Z *հաստատունի:*

Ներդրված վեկտորների ստացման օպտիմիզացիոն նպատակային ֆունկնցիայի դուրս բերման համար օգտագործվում է 1 թերոեմի բանաձևերը: Դիցուք կորպուսում $x_{w,w'}$-ը w և w' բառազույգի նույն պատհունանի ներսում հայտնվելու քանակն է: Դիսկուրս վեկտորի հաջորդական քայլերի արդյունքում գեներացված բառերը իրարից անկախ չեն, սակայն եթե պատահական քայլը այնքան երկար լինի, որ դիսկուրս տարուծության միավոր սֆերայի որևէ ուղության վրա կետրոնացված դիսկուրսներ չլինեն, որոնցից գեներացվել են բառերը, ապա $x_{w,w'}$ քանակների բաշխումը շատ մոտ է լինելու ընդհանրացված բինոմական դիսկրետ բաշխմանը՝

$$Mul(\widetilde{L}, \{p(w, w')\}) = \prod_{(w,w')} \frac{\widetilde{L}!}{x_{w,w'}!} p(w, w')^{x_{w,w'}}: \tag{16}$$

Ենթադրելով այս մոտարկումը, [14] աշխատանքում ցույց է տրվում, որ տրված $x_{w,w'}$ քանակներով կորպուսը գեներացնելու մեծագույն հավանականություն կունենա այն մոդելը, որի ներդրված վեկտորները կբավարարեն հետևյալ մինիմիզացիայի խնդրին՝

$$\min_{\{v_w\},C} \sum_{w,w'} x_{w,w'} \left(\log(x_{w,w'}) - ||v_w + v_{w'}||_2^2 - C\right)^2: \tag{17}$$

Օգտագործելով 15 հավասարությունը՝ ճշմարտանմանության մաքսիմումի գնահատման միջոցով կարելի է ստանալ նմանատիպ օպտիմիզացիոն նպատակային ֆունկցիա՝

$$\min_{\{v_w\}} \sum_{w,w'} x_{w,w'} \left(PMI(w,w') - v_w^T v_{w'}\right)^2: \tag{18}$$

Այս երկու նպատակային ֆունկցիաների լուծումը գտնելը պահանջում է կատարել մատրիցայի կշռված սինգուլյար վերլուծության, որը NP բարդություն ունի, սակայն փորձնական եղանակով պարզվում է գրադիենտային վայերջքի իջեցման եղանակով հնարավոր է մինիմիզացնել վերոնշյալ երկու նպատակային ֆունկիցաները:

Համեմատելով $glove$ մոդելում փորձարարական ճանապարհներով դուրս բերված 8 օպտիմիզացիոն նպատակային ֆուկցիան 17-ի հետ՝ 8-ում մասնակցող անդամները, օգտագործելով թեորեմ 1 բանաձևերը, իմաստ են ստանում, մասնավորապես $b_w = ||v_w||_2^2$:

$Word2vec$ ընտանիքի կոնտեքստից բառ կանխատեսման մոդելի վերափոխված մի տարբերակ դիտարկենք: w_{k+1} բառի հայտնվելու հավանականությունը կախված նախորդ $w_1, w_2, ..., w_k$ հետևյալն է՝

$$p(w_{k+1}|\{w_i\}_{i=1}^k) \propto e^{\frac{1}{k}\sum_{i=1}^k v_{w_{k+1}}^T v_{w_i}}: \tag{19}$$

Ցույց տանք, որ հավանականության մեջ մասնակցող նախորդ k բառերի վեկտորական միջինը տեսականորեն կարելի է հիմնավորել: Դիտարկենք դիսկուրս վեկտորի պատահական քայլով պայմանավորված գեներատիվ մոդելի պարզեցված տարբերակը: Դիսկուրս c վեկտորի նմուշը վերցնել միավոր վեկտորների $\mathcal{C}$ տարածության հավասարահավանական բաշխումից, որից հետո k պատուհանի բառերի նմուշը վերցնել հետևյալ բաշխումից՝

$$(w_1, w_2, ..., w_k) \sim e^{\frac{\sum_{i=1}^k c^T v_{w_i}}{Z_c}}:$$

Բացի այդ ենթադրենք բոլոր c վեկտորների համար $Z_c = Z$ հաստատունի:

Պնդում 1. *Պարզեցված գեներատիվ մոդելի դեպքում՝*

$$\max_{c \in C} p(c|w_1, ..., w_k) = \frac{\sum_{i=1}^k v_{w_i}}{||\sum_{i=1}^k v_{w_i}||_2}:$$

Նկատենք որ, $p(c|w_1,...,w_k)$ հավանականության մաքսիմումը ըստ c-ի և $word2vec$-ի վերափոխված մոդելում մասնակցող k բառերի վեկտորական միջինը միմյանցից տարբերվում է հաստատուն գործակցով, որը փորձնական աշխատանքներում հաշվարկի նվազեցման նպատակներով հաջախ բաց է թողնվում:

Արդեն նշել ենք որ անալոգ բառերի հայտնաբերման խնդիրը բարձր արդյունավետությամբ հնարավոր է լուծել հանրահաշվի վեկտորոական պարզագույն գործողությունների շնորհիվ:

$$d = \underset{d}{\text{argmax}}\, v_d^T(v_c + v_b - v_a), \tag{20}$$

որտեղ բառերի ներդրված վեկտորները նորմալացված են այնպես որ $||v_d||_2^2 = 1$: 20-ը կարելի է գրել նաև հետևյալ կերպ՝

$$d = \underset{d}{\text{argmin}}\, ||v_a - v_b - v_c + v_d||_2^2: \tag{21}$$

20-ի և 21-ի հավասարությունը հետևում է $||v_a - v_b - v_c + v_d||_2^2 = ||v_a - v_b - v_c + v_d||_2^2 + ||v_d||_2^2 + 2v_d^T(v_c + v_b - v_a)$ և $||v_d||_2^2 = 1$ հավասարություններից:

Սեմանտիկ ներկայացումների համեմատական ուսուցում

Դիցուք տրված է $S = \{(x_j, x_j^+, x_{j1}^-, ..., x_{jk}^-)\}_{j=1}^M$ ուսուցման բազմությունը և ֆունկցիաների F բազմությունը, $f \in F$, $f : \mathcal{X} \to \mathcal{R}^d$ արտապատկերում է: F ֆունկնցաների դասի սահմանափակումը S բազմության վրա նշանակենք $\mathcal{F} \circ S$-ով որը հետևյալ բազմությունն է՝

$$\mathcal{F} \circ S = \{((f(x), f(x^+), f(x_1^-), ..., f(x_k^-)))| f \in \mathcal{F} \text{ և } (x, x^+, x_1^-, ..., x_k^-) \in S\}:$$

$\mathcal{F}$ ֆունկցիաների դասի Ռադեմախերի բարդությունը S բազմության նկատմամ հետևյալն է՝

$$R(\mathcal{F} \circ S) = \mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^{(k+2)dM}} \left[\sup_{f \in \mathcal{F}} \langle \sigma, f_{|S} \rangle \right]:$$

Դիտարկենք $g : \mathcal{X} \to \mathcal{R}^{Td}$ արտապատկերումը, որտեղ $g = (f^1, ..., f^T)$ և յուրանքանչյուր $f^t \in \mathcal{F}$: G-ով նշանակենք նկարգրված g արտապատեկերումների դասը:

Այժմ ցույց տանք հետևյալ անհավասարությունը՝ $R(\mathcal{G} \circ S) \leq TR(\mathcal{F} \circ S)$:

$$\begin{aligned} R(\mathcal{G} \circ S) = \mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^{(k+2)TdM}} \left[\sup_{g \in \mathcal{G}} \langle \sigma, g_{|S} \rangle \right] &\leq \mathop{\mathbb{E}}_{\substack{\sigma^1 \sim \{\pm 1\}^{(k+2)dM} \\ \vdots \\ \sigma^T \sim \{\pm 1\}^{(k+2)dM}}} \left[\sum_{t=1}^T \sup_{f^t \in \mathcal{F}} \langle \sigma, f_{|S}^t \rangle \right] \\ &= \mathop{\mathbb{E}}_{\substack{\sigma^1 \sim \{\pm 1\}^{(k+2)dM} \\ \vdots \\ \sigma^T \sim \{\pm 1\}^{(k+2)dM}}} \left[\sum_{t=1}^T \sup_{f^t \in \mathcal{F}} \langle \sigma, f_{|S}^t \rangle \right] = \sum_{t=1}^T \mathop{\mathbb{E}}_{\sigma^t \sim \{\pm 1\}^{(k+2)dM}} \left[\sup_{f^t \in \mathcal{F}} \langle \sigma, f_{|S}^t \rangle \right] \\ &= \sum_{t=1}^T R(\mathcal{F} \circ S) = TR(\mathcal{F} \circ S) \end{aligned}$$

Վերահսկվող ուսուցմամբ ստացվող ներկայացումների կիրառելիությունը այլ առաջադրանքներում

$\mathcal{X}$-ով նշանակենք բոլոր հնարավոր տվյալների օրինակները, իսկ $\mathcal{C}$-ով նշանակենք բոլոր պիտակների կամ դասերի բազմությունը: Յուրաքանչյուր $c \in \mathcal{C}$ դասին համապատասխանում 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 $\mathcal{X}$ տվյալների բազմությունը արտապատկերում d-չափանի էվկլիդյան $\mathcal{R}^d$տարածություն՝ $f : \mathcal{X} \to \mathcal{R}^d$, 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 ընտրված 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) *կույզը, 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՝

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

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

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

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՝*

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

Լեմմա 2 (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 3. *Դիցուք ունենք միմյանցից անկախ $\mathcal{P}(\mathcal{T} \,|\, |\mathcal{T}| = k+1)$ 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:*

$$\hat{L}(f) \geq L(f) - B\sqrt{\frac{(k+1)\log\left(\frac{1}{\delta}\right)}{2n}} \tag{22}$$

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 հետևյալը՝

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

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

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

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՝ կունենանք հետևյալը՝

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

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

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

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: □

Պնդում 2. *Կամայական $v \in \mathbb{R}^d$ վեկտորի համար տեղի ունի հետևյալը՝*

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

Թեորեմ 2. *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՝*

$$\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^n}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i h_i(f(x_i))\right] \leq \sqrt{2}L \mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^{nd}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sum_{j=1}^{d}\sigma_{ij} f_j(x_i)\right] \tag{23}$$

Թեորեմ 3-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 3. *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՝

$$\mathop{\mathbb{E}}_{\sigma\sim\{\pm1\}^n}\left[\sup_{\substack{f\in\mathcal{F}\\y\in\mathcal{Y}}}\sum_{i=1}^{n}\sigma_i h_i(f(x_i),y)\right]\le\sqrt{2}L\mathop{\mathbb{E}}_{\sigma\sim\{\pm\}^{nd}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sum_{j=1}^{d}\sigma_{ij}f_j(x_i)\right] \tag{24}$$

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՝

$$\mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}}\sup_{\substack{f\in\mathcal{F}\\y\in\mathcal{Y}}}\epsilon h_i(f(x_i),y)+g(f,y)\le\sqrt{2}L\mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}^d}\sup_{\substack{f\in\mathcal{F}\\y\in\mathcal{Y}}}\sum_{j=1}^{d}\epsilon_j f_j(x_i)+g(f,y) \tag{25}$$

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

$$2\mathop{\mathbb{E}}_{\epsilon\sim\{\pm1\}}\sup_{\substack{f\in\mathcal{F}\\y\in\mathcal{Y}}}\epsilon h_i(f(x_i),y)-\delta=\sup_{\substack{f,\bar f\in\mathcal{F}\\y\in\mathcal{Y}}}h_i(f(x_i),y)+g(\bar f,y)-h_i(\bar f(x_i),y)+g(\bar f,y)-\delta$$

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

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

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

Օգտագործելով h_i ֆունկցաի Լիպշիցի հատկությամբ օժտված լինելը կունենանք՝

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

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

Պնդում 2-ը կիրառելով կստանանք՝

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

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

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

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 $\sup_{y\in\mathcal{Y}} g(f,y)+g(\bar{f},y)$ ինվարիանտ է $f,\bar{f}$ ֆունկցիաների փոփոխման նկատմամբ, կունենանք՝

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

Հաշվի առնելով Ռադեմախերի ϵ_j փոփոխականների սիմետրիկությունը կստանանք՝

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

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

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

Քանի որ վերջինս տեղի ունի ցանկացած δ-ի համար, այստեղից անմիջապես հետևում է 25 անհավասարությունը:

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

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

24 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 $m \le n$:

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

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 25 անհավասարությունը կստանանք՝

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

□

Թեորեմ 4. *Դիցուք $\mathcal{G}$ ֆունկցիաների բազմությունը, որի յուրաքանչյուր ֆունկցիա Z-ը արտապատկերում է $[0,1]$ և $S = \{z_i\}_{i=1}^m$ m հպորությամբ 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՝*

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(\mathcal{G}) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}} \tag{26}$$

և

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_S(\mathcal{G}) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2m}} \tag{27}$$

Դիցուք ունենք միմյանցից անկախ $\mathcal{P}(\mathcal{T} \,|\, |\mathcal{T}| = k+1)$ 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՝

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

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 4. *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 4 թեորեմը G ֆունկցիաների բազմության համար կունենանք՝

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

Այժմ ցույց տանք, որ ցանկացած $W \in \mathcal{V}$ և $i \in [M]$ համար $h_i(f, W) = g_{f,W}(z_i)$ ֆունկցիան ըստ f-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 $\mathcal{T} \setminus \{y\} = \{y'_1, y'_2, ..., y'_n\}$: $\Phi_y : \mathbb{R}^d \times \mathcal{T} \to \mathbb{R}^n$ որի տեսքը հետևյալն է՝

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

□

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.