Տվյալների նորմավորում

Հայկ Կարապետյան

1 Անհրաժեշտություն

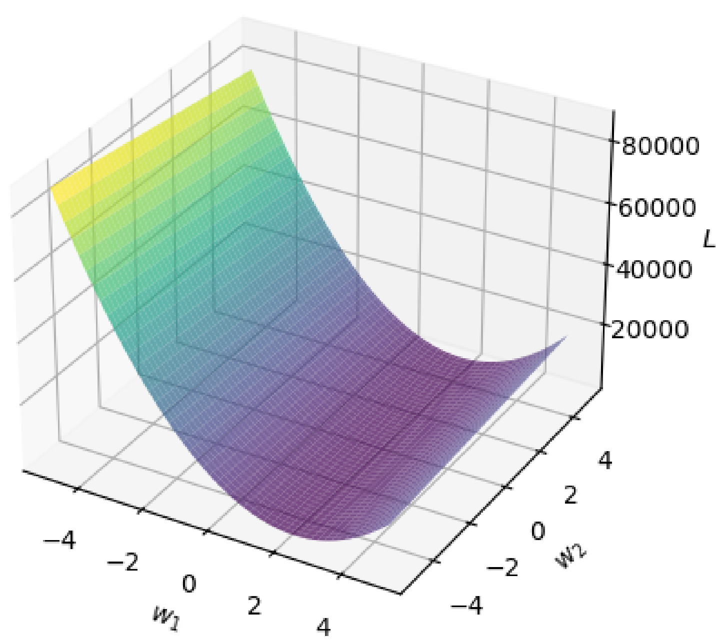
Ունենք բնակարանների տվյալներ։ Բնակարանների բնութագրիչներն են տարածքը (x_1) և սենյակների քանակը (x_2)։ Առկա տվյալների բնութագրիչները փոփոխվում են հետևյալ միջակայքերում.

մակերես (x_1)՝ 60-120, սենյակների քանակ (x_2)՝ 1-4։

Կարող ենք տեսնել, որ տարածքի միջակայքը շատ է տարբերվում սենյակների միջակայքից և դա կարող է առաջացնել խնդիր ուսուցման ընթացքում։ Ուսումնասիրենք այդ խնդիրը։ Ենթադրենք իրական տան գինը ստացվում է հետևյալ բանաձևով.

$$\text{գին} = 2 \times \text{մակերես} + \text{սենյակների քանակ}$$

Փորձենք մոտարկել այս ֆունկցիան օգտագործելով w_1 և w_2 պարամետրեր։ Վերցնենք $f(x) = w_1x_1 + w_2x_2$, կորստի ֆունկցիան քառակուսային՝ $L = (w_1x_1 + w_2x_2 - y)^2$։ Լավագույն դեպքում w_1-ը պետք է հավասարվի 2-ի, իսկ w_1-ը 1-ի։ Կորստի ֆունկցիային նայելիս կարող ենք հասկանալ, որ w_1-ի արժեքը փոքր չափով փոփոխելը կորստի արժեքի վրա կունենա մեծ ազդեցություն, քանի որ բազմապատկվում է x_1-ով, իսկ w_2-ի արժեքը փոքր չափով փոփոխելը գրեթե ազդեցություն չի ունենա կորստի արժեքի վրա։ Պատկերենք կորստի ֆունկցիայի գրաֆիկը (Նկար 1)։

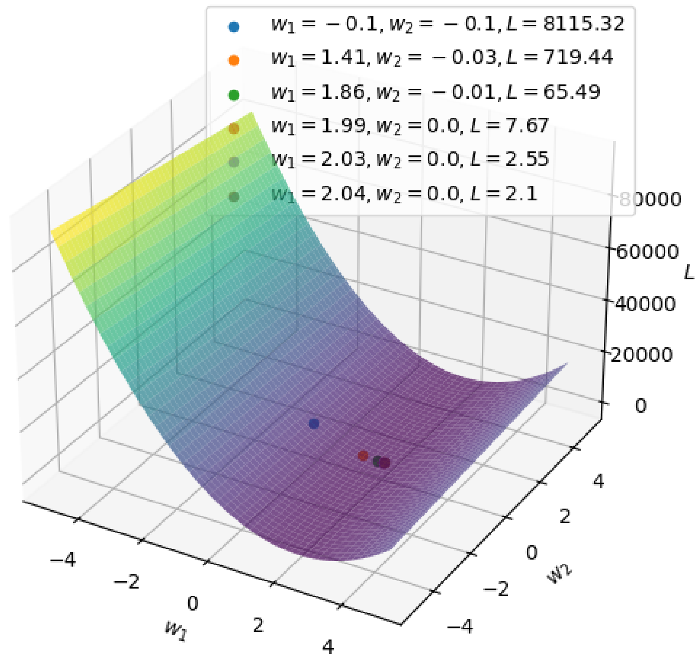


Նկար 1: Կորստի ֆունկցիայի գրաֆիկը առանց տվյալների նորմավորման[1]։
w_2-ի արժեքը փոփոխելիս կորստի արժեքը գրեթե չի փոփոխվում։
Կորստի արժեքը փոփոխվում է 0-80000 միջակայքում

Նույն եզրահանգմանը կարող էինք գալ նայելով կորստի ֆունկցիայի մասնակի ածանցյալները.

$$\nabla L = \left[\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}\right]$$
$$\frac{\partial L}{\partial w_1} = 2x_1(w_1x_1 + w_2x_2 - y)$$
$$\frac{\partial L}{\partial w_2} = 2x_2(w_1x_1 + w_2x_2 - y)$$

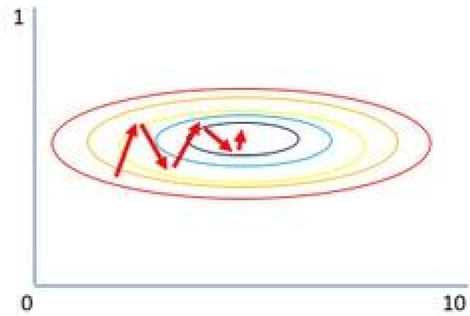
Մասնակի ածանցյալներում մասնակցում են x_1 և x_2 արժեքները: Այսինքն $\frac{\partial L}{\partial w_1}$-ը ընդունում է մեծ արժեքներ, իսկ $\frac{\partial L}{\partial w_2}$-ը փոքր արժեքներ, այդ իսկ պատճառով կշիռները թարմացնելիս w_2-ի արժեքը շատ քիչ է փոփոխվելու (Նկար 2):



Նկար 2: 5 քայլ գրադիենտային վայրէջք առանց տվյալների նորմավորման

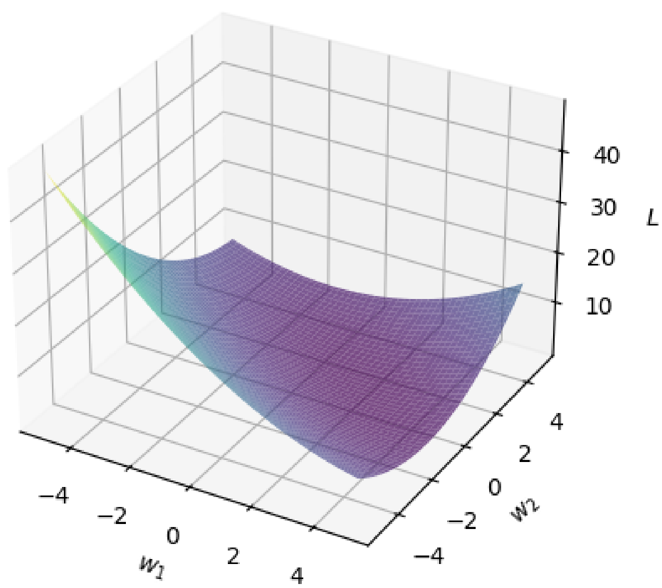
Նկար 2-ում երևում է, որ 5 քայլ գրադիենտային վայրեջքից հետո w_1-ը գրեթե հասել է իր լավագույն արժեքին, իսկ w_2-ը դեռ երկար ճանապարհ ունի անցնելու: Կորստի արժեքը հավասար է 2.1: Առանց տվյալների նորմավորման, կորստի ֆունկցիայի մակարդակի գծերը[2] ունեն էլիպսների տեսք (Նկար 3): Ինչքան փոքր է էլիպսը նշանակում է ավելի ցածր տեղից ենք կտրել ֆունկցիայի գրաֆիկը:

1. Տվյալների նորմավորումը, բնութագրիչների նույն միջակայք բերելու գործընթացն է:
2. Եռաչափ ֆունկցիան, երկչափ տարածությունում պատկերելու համար կտրենք այն մի քանի հարթություններով: Օրինակ՝ վերցնենք բոլոր այն x և y արժեքները, որտեղ z=5, կստանանք z=5 մակարդակի գիծը: Ստացված x և y զույգերը պատկերենք երկչափ տարածությունում: Ստացված պատկերը կոչվում է ֆունկցիայի մակարդակի գծեր:



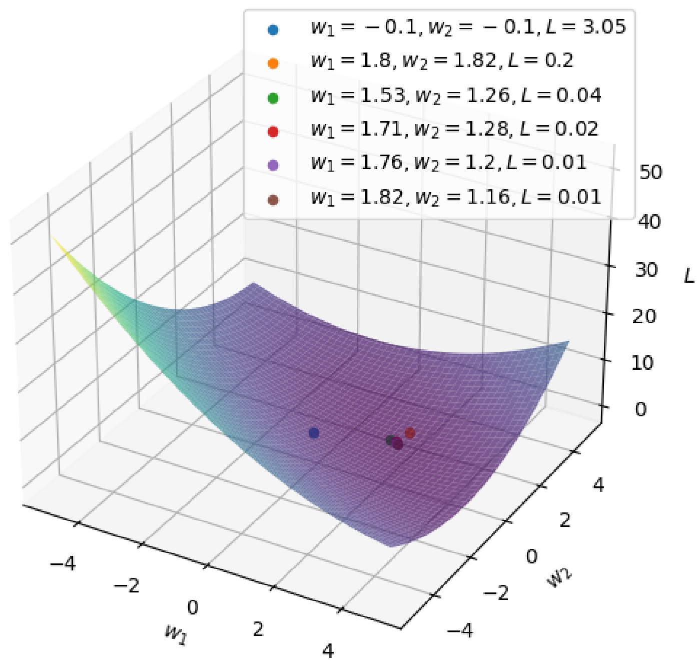
Նկար 3: Կորստի ֆունկցիայի մակարդակի գծերը առանց տվյալների նորմավորման

Իսկ ինչ տեսք կունենան կորստի ֆունկցիան և մակարդակի գծերը, եթե մեր տվյալները լինեն նույն միջակայքից, մակերես՝ 0-10, սենյակների քանակ՝ 0-10: Կորստի արժեքը փոփոխվում է w_2-ի արժեքը փոփոխելիս (Նկար 4):



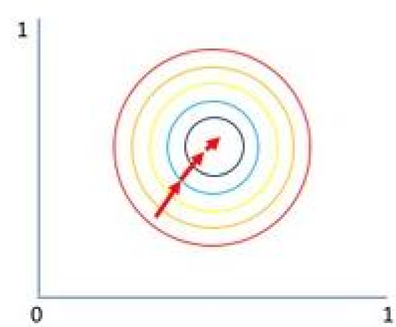
Նկար 4: Կորստի ֆունկցիայի գրաֆիկը տվյալների նորմավորումից հետո: w_2-ի արժեքը փոփոխելիս կորստի արժեքը զգալիորեն փոփոխվում է: Կորստի արժեքը փոփոխվում է 0-40 միջակայքում

5 քայլ գրադիենտային վայրէջքից հետո կորստի արժեքը հասնում է 0.01, իսկ w_1-ը և w_2-ը մոտ են իրենց լավագույն արժեքներին (Նկար 5)։



Նկար 5: 5 քայլ գրադիենտային վայրէջք տվյալների նորմավորումից հետո

Մակարդակի գծերը կունենան շրջանանման տեսք և մինիմումի կետը գտնելու ժամանակը ավելի կարճ կլինի։



Նկար 6: Կորստի ֆունկցիայի մակարդակի գծերը տվյալների նորմավորումից հետո

Ինչպես տեսանք տվյալների նորմավորումից հետո նույն քանակի գրադիենտային քայլերից հետո կորուստը հասավ ավելի փոքր արժեքի, ինչը հետևում է նաև մակարդակի գծերից։ Տվյալ օրինակում կորստի ֆունկցիան ուռուցիկ է և միանշանակ չէ ոչ ուռուցիկ ֆունկցիայի դեպքում տվյալների նորմավորում գործածելը կարագացնի ուսուցման գործընթացը, թե ոչ։ Բայց շատ փորձարկումներ ցույց են տալիս, որ տվյալների նորմավորում օգտագործելը օգնում է ուսուցմանը և կարող է կրճատել ուսուցման ժամանակը մի քանի անգամ։ Այդ պատճառով տարբեր միջակայքերի բնութագրիչներ ունենալիս, նորմավորեք դրանք ներքևում նշված եղանակներից մեկն օգտագործելով։

2 Ստանդարտ նորմավորում

Նորմավորման մեթոդներից է ստանդարդ նորմավորումը (standard normalization): Հետևյալ նորմավորման համար անհրաժեշտ է հաշվել տվյալների միջինը և ստանդարդ շեղումը: Դրանք հաշվելուց հետոմ ամեն տվյալից ահանում ենք միջինը և բաժանում ստանդարդ շեղման բրա: Արդյունքում ստանում ենք տվյալներ, որոնց միջինը հավասար է զրոյի, իսկ ստանդարդ շեղումը հավասար է մեկի: Արդյունքում տվյալների միջակայքը կդառնա (-1; 1): Ունենք $(x_i, y_i)_{i=1}^n,\ x_i \in R^k,\ y_i \in R^m$ տվյալներ:

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \dots & x_1^k \\ x_2^1 & x_2^2 & \dots & x_2^k \\ \vdots & \vdots & \vdots & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^k \end{bmatrix}$$

Վերցնենք $a^j = \dfrac{1}{n} \sum_{i=1}^n x_i^j$, $\sigma^j = \sqrt{\dfrac{1}{n} \sum_{i=1}^n (x_i^j - a^j)^2}$

$$z_i^j = \frac{x_i^j - a^j}{\sigma^j}$$

Մեր նոր տվյալները արդեն բաղկացած կլինեն $(x_i, y_i)_{i=1}^n,\ z_i \in R^k,\ y_i \in R^m$ զույգերից: Միջին արժեքը և ստանդարդ շեղումը հաշվելու ենք ուսուցման տվյալների միջոցով: Կարգավորման և փորձարկման տվյալների հավաքածուները նորմավորելիս, արդեն օգտագործելու ենք ուսուցման տվյալների համար հաշված միջինը և ստանդարդ շեղումը:

3 Փոքր-Մեծ նորմավորում

Երկրորդ նորմավորման մեթոդը փոքր-մեծ նորմավորումն է (min-max normalization): Հետևյալ նորմավորման ժամանակ անհրաժեշտ է ամեն տվյալից հանել ամենափոքր արժեքը և բաժանել մեծագույն ու փոքրագույն արժեքների տարբերության վրա: Ունենք $(x_i, y_i)_{i=1}^n,\ x_i \in R^k,\ y_i \in R^m$ տվյալներ:

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \dots & x_1^k \\ x_2^1 & x_2^2 & \dots & x_2^k \\ \vdots & \vdots & \vdots & \vdots \\ x_n^1 & x_n^2 & \dots & x_n^k \end{bmatrix}$$

$$z_i^j = \frac{x_i^j - min_i x_i^j}{max_i x_i^j - min_i x_i^j}$$

Մեր նոր տվյալները արդեն բաղկացած կլինեն $(x_i, y_i)_{i=1}^n,\ z_i \in R^k,\ y_i \in R^m$ զույգերից: Այստեղ նույնպես փոքրագույն և մեծագույն արժեքները ընտրում ենք ուսուցման տվյալներից: Կարգավորման և փորձարկման տվյալների հավաքածուները այս մեթոդով նորմավորելիս, նույնպես օգտագործում ենք ուսուցման տյալների փոքրագույն և մեծագույն արժեքները: