Վիճակագրական տվյալների տեսակներ: Սոֆթմաքս դասակարգիչ

Հայկ Կարապետյան

1 Վիճակագրական տվյալների տեսակներ

Վիճակագրական տվյալները բաժանվում են 3 հիմնական խմբի.

1. թվային (numerical)
2. կատեգորիկ (categorical)
3. հերթական (ordinal)

1. Թվային տվյալները ներառում են ինչպես դիսկրետ (դասարանում սովորողների քանակ), այնպես էլ անընդհատ արժեքներ (մարդու հասակ):
2. Կատեգորիկ տվյալները բաժանվում են առանձին խմբերի: Օրինակ՝ գույներ, մեքենաների տեսակներ: Կատեգորիկ տվյալները թվայինի վերածելիս չեն կարող հերթականորեն համարակալվել: Օրինակ՝ ունենք 3 գույն՝ դեղին, կարմիր, կանաչ: Չենք կարող դեղին գույնին վերագրել 0, կարմիրին՝ 1, կանաչին՝ 2, քանի որ նրանց մեջ ոչ մի հերթականություն չկա և մոդելը ուսուցանելիս այն կարող է մտածել, որ կանաչ գույնը ավելի մեծ կարևորություն է ցույց տալիս քան դեղինը: Այդ պատճառով օգտագործվում են one-hot վեկտորները: 3 կատեգորիայի դեպքում մենք կունենանք 3 վեկտոր.

$$\text{դեղին} = \begin{bmatrix}1\\0\\0\end{bmatrix}, \quad \text{կարմիր} = \begin{bmatrix}0\\1\\0\end{bmatrix}, \quad \text{կանաչ} = \begin{bmatrix}0\\0\\1\end{bmatrix}$$

Տվյալների one-hot վեկտորների փոխակերպելու գործընթացը կոչվում է one-hot encoding:

3. Հերթական տվյալները նման են կատեգորիկ տվյալներին, բայց ի տարբերություն կատեգորիկի կարող ենք համարակալել, քանի որ նրանք ցույց են տալիս ինչ որ հերթականություն: Օրինակ՝ վաճառողուհու սպասարկումը կարող ենք գնահատել 3 տարբերակով՝ "վատ", "միջին", "լավ": Հետևյալ 3 կատեգորիաները կարող ենք փոխարինել համապատասխանաբար 0, 1, 2 թվերով, քանի որ "լավ" գնահատականը ավելի մեծ է, քան "վատ" գնահատականը:

2 Սոֆթմաքս դասակարգիչ

Ներմուծենք հավանականային վեկտոր գաղափարը: Վեկտորը կոչվում է հավանականային, եթե նրա բոլոր կոորդինատները մեծ են 0-ից և կոորդինատների գումարը հավասար է մեկի:

$$P = \begin{bmatrix}P_1\\P_2\\....\\P_m\end{bmatrix}, \quad P_i > 0, \quad \sum_{i=1}^{m} P_i = 1$$

m չափանի հավանականային վեկտոր

Ունենք $(x_i, y_i)_{i=1}^{n}, \quad x_i \in R^k, \quad y_i \in R^m$ տվյալների զույգեր: y_i ներկայացնում է one-hot վեկտոր: Սահմանենք ֆունկցիա, որը մեզ կվերադարձնի հավանականային վեկտոր:

$$f(x) = \left[\frac{e^{w_1^T x + b_1}}{\sum_{i=1}^{m} e^{w_i^T x + b_i}}, \ \ldots, \ \frac{e^{w_m^T x + b_m}}{\sum_{i=1}^{m} e^{w_i^T x + b_i}}\right]$$

Մեր նպատակն է գտնել այնպիսի $(b_i, w_i)_{i=1}^{m}$ պարամետրեր, որ $f(x_i) \approx y_i, \ i = 1, \ldots n$ և ոչ միայն մեր ունեցած տվյալների համար: Օրինակ՝ $x_i \sim$ նկար, $y_i \sim$ պիտակ (շուն, կատու, փիղ): Հետևյալ խնդիրը լուծելու համար օգտագործենք cross-entropy կորստի ֆունկցիան մի քանի պիտակի դեպքում:

Երկու պիտակի դեպքում այն ուներ հետևյալ տեսքը.

$$L = \frac{1}{n}\sum_{i=1}^{n} -y_i ln(f(x_i)) - (1-y_i)ln(1-f(x_i)) \tag{1}$$
$$y_i \in \{0,1\},\ f(x_i) \in (0;1)$$

Մի քանի պիտակի դեպքում այն ունի հետևյալ տեսքը.

$$L = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m} -y_i^k ln(f(x_i)^k) \tag{2}$$

$y_i \in$ one-hot վեկտոր, $f(x_i) \in$ հավանականային վեկտոր, m-ը պիտակների քանակ

y_i^k-ն ընդունում է 1 կամ 0 արժեք (one-hot վեկտոր)։ 1 արժեք ընդունելու դեպքում մեզ պետք է մինիմիզացնել $-ln(f(x_i)^k)$, որը իր փոքրագույն արժեքը ընդունում է, երբ $f(x_i^k) = 1$ (հավանականային վեկտոր)։ Օրինակ՝ երբ մեր նկարի պիտակը եղել է շուն և մոդելը 100% վստահությամբ ասել ենք շուն, այդ դեպքում մեր կորուստը կլինի 0։
Նկատենք որ (1) կորստի ֆունկցիան նույնն է, ինչ (2) կորստի ֆունկցիան, երբ պիտակների քանակը հավասար է 2։

$$L = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{2} -y_i^k ln(f(x_i)^k) = \frac{1}{n}\sum_{i=1}^{n} -y_i^1 ln(f(x_i)^1) - y_i^2 ln(f(x_i)^2)$$
$$y_i = \begin{bmatrix}1\\0\end{bmatrix} \text{ կամ } \begin{bmatrix}0\\1\end{bmatrix},\ y_i^1 = 1 \text{ կամ } 0,\ y_i^2 = 0 \text{ կամ } 1$$
$$f(x_i)^1 = \frac{e^{w_1^T x + b_1}}{\sum_{i=1}^{2} e^{w_i^T x + b_i}},\ f(x_i)^2 = \frac{e^{w_2^T x + b_2}}{\sum_{i=1}^{2} e^{w_i^T x + b_i}}$$
$$y_i^2 = 1 - y_i^1,\quad f(x_i)^2 = 1 - f(x_i)^1$$

Այս երկու պայմանը հաշվի առնելով և տեղադրելով (2) կորստի ֆունկցիայի մեջ կստանանք (1) կորստի ֆունկցիան։
Իսկ ինչպե՞ս կարող ենք սոֆթմաքս դասակարգիչը ներկայացնել, որպես նեյրոնային ցանց։ f(x) ֆունկցիային նայելիս կարող ենք տեսնել, որ այն իրենից ներկայացնում է m հատ գծային ֆունկցիա։ Այսինքն կունենանք m հատ նեյրոնից բաղկացած նեյրոնային ցանց և ակտիվացիոն ֆունկցիան սոֆթմաքս։