

МАТЕМАТИЧНА ОСНОВА КЛАСИФІКАЦІЇ KNN В ЕКОНОМІЦІ

Егор Сенюк, студент ФММ

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна, Київ, Україна

Мельничук Вікторія Едуардівна, асистент

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна, Київ, Україна

yehor146@gmail.com

У повсякденному житті часто потрібно віднести об'єкт до певної категорії за його ознаками - це класифікація. Прикладом в економіці є сегментація ринку, тобто розподіл споживачів за інтересами, віком чи доходом для точнішого націлення реклами та послуг.

Постановка задачі. Дана вибірка з генеральної сукупності оформлена у вигляді таблиці, що містить геточок даних (записів) (див. табл. 1). Кожна точка даних складається з n значень f ознак F_i (feature) та мітки c класу C (class) до якого вона належить. На вхід поступає нова точка даних з певними значеннями ознак $(f_i)_{i=1}^n$, визначити найбільш ймовірну мітку.

Таблиця 1. Вибірka.

Вибірka					
№	F1	F2	...	F _n	C
1	f ₁₁	f ₂₁	...	f _{n1}	c ₁
2	f ₁₂	f ₂₂	...	f _{n2}	c ₂
...	...	...	...	...	...
re	f _{1re}	f _{2re}	...	f _{nre}	c _{re}

Векторизація. Для того щоб алгоритми штучного інтелекту могли працювати з даними різного типу уніфіковано (незважаючи на природу цих даних) необхідно розробити єдиний формат подання даних. Одним з таких форматів є вектор чисел. Кожна координата такого вектора містить відцифроване представлення значення відповідної ознаки. Процес представлення початкових даних у вигляді числового вектора - векторизація.

Дискримінантний аналіз. Розділ математики, який вивчає методи дискримінації (класифікації) об'єктів на основі їхніх ознак. Ціль дискримінантного аналізу: на основі ознак об'єктів провести їх дискримінацію, тобто віднести їх до одного з відомих класів. Робиться припущення, що об'єкти одного класу мають схожі ознаки [1].

Дискримінантна функція $g_c(x)$ - функція за допомогою якої на основі ознак оцінюють схожість об'єкта на об'єкти класу c . Об'єкту присвоюється мітка класу з найбільшою оцінкою дискримінантної функції (найбільшою схожістю), тобто:

$$y_x = \arg \max_x g_c(x) \quad (1)$$

Постає питання: яким чином можна чисельно оцінити схожість двох об'єктів? Відповідь - за допомогою задання метричного простору.

Метричний простір - пара (M, d) , де M - множина, $d: M \times M \rightarrow \mathbb{R}$ - функція відстані.

1. $d(x, y) = 0 \Leftrightarrow x = y$ (аксіома тотожності)
2. $d(x, y) \geq 0$ (аксіома невід'ємності)
3. $d(x, y) = d(y, x)$ (аксіома симетричності)
4. $d(x, y) + d(y, z) \geq d(x, z)$ (аксіома трикутника)

Аксіоми 1, 2, 3 інтуїтивно відповідають людському уявленню про відстань, аксіома 4 потребує додаткових пояснень. Вона постулює, що відстань є найкоротший шлях з усіх можливих. Візьмемо 3 точки: A, B , та C (елементи множини M). Розглянемо наступні 3 шляхи:

1. $d(A, C)$ (прямий шлях)
2. $d(A, B) + d(B, C)$ (частина 1 транзитного шляху через B)
3. $d(C, B) + d(B, A)$ (частина 2 транзитного шляху через B)

Очевидно, що існує 2 варіанти як дістатися від А до Б: прямий шлях та транзитний шлях. Людське уявлення про відстань каже нам, що найкоротший варіант - прямий. В той же час, якщо транзитна точка (В) лежить на прямому шляху, то прямий шлях по довжині дорівнює транзитному. Якщо ж транзитна точка не лежить на прямому шляху, то довжина транзитного шляху буде більшою за прямий. Саме це людське уявлення і описує 4-а аксіома.

Приклади поширених функцій відстані.

Евклідова відстань.

$$d(x,y)=||x-y||=(\sum_{i=1}^n(x_i-y_i)^2)^{1/2} \quad (2)$$

Косинус подібності.

$$d(x,y)=\cos(\angle(x,y))=xy/|x||y| \quad (3)$$

Нормалізація. На практиці часто доводиться мати справу з двома і більше ознаками. Ці ознаки можуть мати різні діапазони можливих значень. Це в свою чергу може призвести до проблеми, коли одна ознака має значно менший діапазон можливих значень ніж інша, і через це при підрахунку відстані ознака з меншим діапазоном майже не враховується, хоча вона може бути важливою ознакою для дискримінації. Для усунення цієї проблеми використовується нормалізація - приведення усіх ознак до одного діапазону можливих значень. Таким чином, кожна ознака враховується незалежно від свого діапазону можливих значень [2].

Приклади поширених нормалізацій.

Мінімакс-нормалізація.

$$f'=f-\min[F]\max[F] - \min[F] \quad (4)$$

Z-нормалізація.

$$f'=(f-M[F])/[F] \quad (5)$$

KNN з незваженим голосуванням. На вхід подається вектор x . Обирається k найближчих до x векторів - сусідів. Вектору z присвоюється мітка класу більшості сусідів, тобто:

$$g_c(x)=x_c \text{ clc}(x) \quad (6)$$

$I_c(x)$ -індикаторна функція, $I_c(x)=1$ if $x=C$; $I_c(x)=0$ if $x \neq C$

Такий підхід має суттєвий недолік: він робить припущення, що всі сусіди мають однакову вагу голосу, незалежно від того наскільки близько вони розміщені до x . Проблему дозволяє вирішити введення вагової функції.

Вагова функція $w:AR^+$ - функція, яка використовується для того, щоб при сумі або інтегрованні зробити внесок елементів множини A в результат неоднаковим. Прикладом є обернена відстань.

Обернена відстань.

$$w(x,y)=1/d(x,y)+ \quad (7)$$

- мала додатня константа, що дозволяє уникнути ділення на 0 коли $d(x,y)=0$.

KNN із зваженим голосуванням. На вхід подається вектор x . Обирається k найближчих до x векторів - сусідів. Обраховується внесок кожного з сусідів. x присвоюється мітка сусідів одного класу, зробивших найбільший внесок.

$$g_c(x)=x_c \text{ clc}(x)w(x,x_c) \quad (8)$$

Висновок. Таким чином, представлено метод KNN як один із варіантів розв'язання задачі класифікації. Питання визначення оптимального значення параметра k залишилось відкритим.

Література:

1. GeeksforGeeks. K Nearest Neighbors (KNN) Algorithm [Електронний ресурс]. – Режим доступу: <https://www.geeksforgeeks.org/ml-k-nearest-neighbors-algorithm/> (дата звернення: 28.02.2025).
2. Scikit-learn. Nearest Neighbors [Електронний ресурс]. – Режим доступу: <https://scikit-learn.org/stable/modules/neighbors.html> (дата звернення: 28.02.2025).