

МАТЕМАТИЧНА ОСНОВА КЛАСИФІКАЦІЇ НБК В ЕКОНОМІЦІ

У житті часто зустрічаються задачі віднесення об'єкту до певної категорії спираючись на ряд ознак – класифікацію. Застосування класифікації НБК (Наївного Баєсівського класифікатора) в економіці та менеджменті можна дослідити через практичну частину.

Така вибірка з генеральної сукупності оформлена у вигляді таблиці, що містить r_e точок даних (записів) (див. табл. 1). Кожна точка даних складається з n значень f дискретних 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 [1].

Таблиця 1 – Вибірка

Вибірка					
№	F_1	F_2	...	F_n	C
1	f_{11}	f_{21}	...	f_{n1}	c_1
2	f_{12}	f_{22}	...	f_{n2}	c_2
...	...	...	...	...	...
r_e	f_{1r_e}	f_{2r_e}	...	f_{nr_e}	c_{r_e}

Розглянемо одновимірну дискретну випадкову величину Y – мітка c_i класу C (див. табл. 2). Зауважимо, статистичне визначення безумовної ймовірності мітки – відношення кількості появ мітки даного класу у виборці m_c до загальної кількості появ міток у виборці r_e , тобто:

$$P(Y = y_i) = \frac{m_c}{r_e}. \quad (1)$$

Таблиця 2 – Закон розподілу дискретної одновимірної випадкової величини міток класу – Y

Y	y_1	y_2	...	y_l
p	$P(Y = y_1)$	$P(Y = y_2)$	...	$P(Y = y_l)$

Розглянемо одновимірну дискретну випадкову величину X_i - значення f_i дискретної ознаки F_i (див. табл. 3). Зауважимо, статистичне визначення безумовної ймовірності значення ознаки – відношення кількості появ даного значення f_i ознаки F_i , тобто m_{f_i} до загальної кількості значень ознаки F_i , інакше кажучи r_e . Маємо:

$$P(X_i = x_{ij}) = \frac{m_{f_i}}{r_e}. \quad (2)$$

Таблиця 3 – Закон розподілу дискретної одновимірної випадкової величини значень ознаки F_i – X_i

X_i	x_{i1}	x_{i2}	...	x_{ik_i}
p_i	p_{i1}	p_{i2}	...	p_{ik_i}

Розглянемо двовимірні випадкові величини $(X_i, Y)_{i=1}^n$ (див. табл. 4). По суті це всі можливі значення усіх можливих ознак та ймовірності їх появи разом із певною міткою класу.

Таблиця 4 – Закон розподілу двовимірної дискретної випадкової величини (X_i, Y)

Y	X_i			
	x_{i1}	x_{i2}	...	x_{ik_i}
y_1	$P(X = x_{i1}, Y = y_1)$	$P(X = x_{i2}, Y = y_1)$	...	$P(X = x_{ik_i}, Y = y_1)$
y_2	$P(X = x_{i1}, Y = y_2)$	$P(X = x_{i2}, Y = y_2)$	...	$P(X = x_{ik_i}, Y = y_2)$
...	...	...	...	...
y_l	$P(X = x_{i1}, Y = y_l)$	$P(X = x_{i2}, Y = y_l)$	...	$P(X = x_{ik_i}, Y = y_l)$

Розглянемо двовимірну дискретну випадкову величину $Z = [(X_1, Y), (X_2, Y), \dots, (X_n, Y)]^T$. По суті це всі можливі комбінації значення ознак та ймовірності їх появи разом із певною міткою класу.

Тепер дамо відповідь чому саме «наївний». Справа в тому, що з метою зменшення кількості обчислень, ми робимо припущення про взаємну незалежність дискретних випадкових величин значень ознак [2]. Інакше кажучи, ми припускаємо, що ознаки є незалежним одна від одної:

$$P(Z = z|Y = y) = P(X_1 = x_{1j_1}, X_2 = x_{2j_2}, \dots, X_n = x_{nj_n}|Y = y) = \prod_{i=1}^n P(X_i = x_{ij_i}|Y = y). \quad (3)$$

Зауважимо, що НБК опосередковано спирається на теорему Бернуллі (ЗВЧ), тобто робиться припущення, що $(X)_{i=1}^n$ задовольняють її умови, а тому:

$$W(X_i = x_{ij_i}|Y = y) \approx P(X_i = x_{ij_i}|Y = y). \quad (4)$$

Це припущення дозволяє робити висновки про ймовірність на генеральній сукупності спираючись лише на відносну частоту отриману з обмеженої вибірки.

Розглянемо базове формулювання теореми Байєса.

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(A)}. \quad (5)$$

Зробимо підстановку. Гіпотеза - мітка класу. Подія, що трапилася – поява нової точки даних.

$$A = (Z = z) \mid B_i = (Y = y) \mid P(A|B_i) = P(Z = z|Y = y) \mid P(B_i) = P(Y = y) \mid P(A) = P(Z = z) \\ P(B_i|A) = P(Y = y|Z = z).$$

Перепишемо формулювання теореми Байєса для нашої задачі:

$$P(Y = y|Z = z) = \frac{P(Y = y)P(Z = z|Y = y)}{P(Z = z)} = \frac{P(Y = y) \prod_{i=1}^n P(X_i = x_{ij_i}|Y = y)}{P(Z = z)} \\ = \frac{P(Y = y) \prod_{i=1}^n \frac{P(X_i = x_{ij_i}, Y = y)}{P(Y = y)}}{P(Z = z)}. \quad (6)$$

Знаменник у всіх гіпотез однаковий, тому для порівняння його можна нівелювати.

$$P(Y = y|Z = z) \propto P(Y = y) \prod_{i=1}^n P(X_i = x_{ij_i}|Y = y) \mid P(Y = y|Z = z) \propto P(Y = y) \prod_{i=1}^n \frac{P(X_i = x_{ij_i}, Y = y)}{P(Y = y)}. \quad (7)$$

Знаходимо найбільш ймовірну мітку вхідної точки даних шляхом розрахунку та порівняння усіх наявних гіпотез (міток класів).

$$\hat{y} = \operatorname{argmax}_y P(Y = y)P(Z = z|Y = y). \quad (8)$$

Класифікація НБК у менеджменті та економіці може бути застосована для цілої низки сфер діяльності: класифікація клієнтів (наприклад, для використання для сегментації клієнтів за різними характеристиками, такими як вік, місце проживання або купівельна поведінка та допомагає в таргетуванні маркетингових кампаній або в підвищенні лояльності клієнтів); виявлення шахрайства (для виявлення аномальних транзакцій або підозрілих дій, аналізуючи певні атрибути, наприклад, місце проведення операції, тип товару або суму, і визначати, чи є транзакція підозрілою); оцінка ризиків (для оцінки кредитного ризику, де клієнтів класифікують як «ризиковий» або «безпечний» на основі їх фінансової історії, кредитного рейтингу і інших параметрів); аналіз відгуків клієнтів (класифікувати відгуки або коментарі на основі настроїв (позитивні, негативні, нейтральні), що дозволить компаніям швидко аналізувати відгуки про їхні продукти або послуги та приймати рішення щодо покращення якості); прогнозування попиту (на основі історичних даних про продажі і сезонні тренди, НБК може передбачати попит на певні товари в майбутньому, що дозволяє ефективніше планувати запаси та закупки); персоналізація пропозицій (прогнозування, які продукти або послуги можуть зацікавити конкретного користувача, виходячи з його минулих покупок або переглядів, що допомагає персоналізувати маркетингові пропозиції).

Таким чином, було розглянуто одне з багатьох можливих рішень задачі класифікації, а саме класифікація за допомогою НБК. Зауважимо, в статті не піднята проблема недостатньої репрезентативності вибірки: коли у вибірці немає жодної точки даних, яка б містила значення f_i і мала мітку c_j .

Список використаних джерел

1. GeeksforGeeks. Naive Bayes Classifiers. URL: <https://www.geeksforgeeks.org/naive-bayes-classifiers/> (the date of application: 23.02.2025).
2. Scikit-learn. Naive Bayes. URL: https://scikit-learn.org/stable/modules/naive_bayes.html (the date of application: 23.02.2025).