Методи класифікації даних:

* Інтервал вручну.
* Метод k-найближчих сусідів
* Рівний інтервал.
* Квантиль.
* Природні перерви (Дженкс)
* Геометричний інтервал.
* Стандартне відхилення.

Ме́тод *k*-найбли́жчих сусі́дів (англ. *k-nearest neighbor method*) — простий непараметричний класифікаційний метод, де для класифікації об'єктів у рамках простору властивостей використовуються відстані (зазвичай евклідові), пораховані до усіх інших об'єктів. Вибираються об'єкти, до яких відстань найменша, і вони виділяються в окремий клас. Метод k-найближчих сусідів — метричний алгоритм для автоматичної класифікації об'єктів. Основним принципом методу найближчих сусідів є те, що об'єкт присвоюється тому класу, який є найбільш поширеним серед сусідів даного елемента. Сусіди беруться, виходячи з множини об'єктів, класи яких уже відомі, і, виходячи з ключового для даного методу значення k, вираховується, який клас є найчисленнішим серед них. Кожен об'єкт має кінцеву кількість атрибутів (розмірностей). Передбачається, що існує певний набір об'єктів з уже наявною класифікацією.

Природні перерви

Природні перерви - це свого роду «оптимальна» схема класифікації, яка знаходить розриви класів, які (для певної кількості класів) мінімізують розбіжності між класами та максимізують відмінності між класами. Одним недоліком цього підходу є те, що кожен набір даних створює унікальне рішення щодо класифікації, і якщо вам потрібно зробити порівняння між картами, наприклад, в атласі чи серії (наприклад, по одній карті на 1980, 1990, 2000), можливо, ви захочете використовувати єдину схему, яку можна застосувати на всіх картах.

Квантиль

Квантильна класифікація добре підходить для лінійної розподілених даних. Кожен клас у такій класифікації містить однакову кількість ознак. Проте подібні ознаки можуть бути в різних класах,а об’єкти з великою різницею можуть потрапити в один клас. Це спотворення можна мінімізувати, збільшивши кількість класів.

Добре для:

• Підкреслення відносного положення, наприклад які округи знаходяться у топ-20% за доходом

Недоліки:

• Функції з однаковими значеннями можуть потрапляти в різні класи, перебільшуючи їхні відмінності

• Також може статися зворотне: широкий діапазон значень може потрапити до одного класу, мінімізуючи відмінності

Метод рівного інтервалу

Ділить дані на класи одного розміру (наприклад, 0-10, 10-20, 20-30 тощо) і найкраще працює на даних, які зазвичай розподіляються по всьому діапазону. Уникайте рівного інтервалу, якщо ваші дані перекошені в один кінець або якщо у вас є одне чи два справді великі значення, що не відповідають вимогам. У цьому випадку викиди, швидше за все, даватимуть порожні класи, витрачаючи цілком хороші класи без жодних спостережень.

Середньоквадратичне відхилення

Середньоквадратичне відхилення визначається як квадратний корінь з дисперсії випадкової величини: вимірюється в одиницях виміру самої випадкової величини і використовується при розрахунку стандартної помилки середнього арифметичного, при побудові довірчих інтервалів, при статистичній перевірці гіпотез, при вимірюванні лінійного взаємозв'язку між випадковими величинами. На практиці, коли замість точного розподілу випадкової величини в розпорядженні є лише вибірка, стандартне відхилення, як і математичне очікування, оцінюють, і робити це можна різними способами. Терміни «стандартне відхилення» і «середньоквадратичне відхилення» зазвичай застосовують до квадратному кореню з дисперсії випадкової величини, але іноді і до різних варіантів оцінки цієї величини на підставі вибірки.

Література:

* <https://www.axismaps.com/guide/data-classification>
* <https://www.deeplearningbook.org/>
* <https://ru.wikipedia.org/wiki/%D0%A1%D1%80%D0%B5%D0%B4%D0%BD%D0%B5%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5_%D0%BE%D1%82%D0%BA%D0%BB%D0%BE%D0%BD%D0%B5%D0%BD%D0%B8%D0%B5>
* https://uk.wikipedia.org/wiki/%D0%9C%D0%B5%D1%82%D0%BE%D0%B4\_k-%D0%BD%D0%B0%D0%B9%D0%B1%D0%BB%D0%B8%D0%B6%D1%87%D0%B8%D1%85\_%D1%81%D1%83%D1%81%D1%96%D0%B4%D1%96%D0%B2#:~:text=k%2Dnearest%20neighbor%20method)%20%E2%80%94,%D0%B4%D0%BE%20%D1%83%D1%81%D1%96%D1%85%20%D1%96%D0%BD%D1%88%D0%B8%D1%85%20%D0%BE%D0%B1'%D1%94%D0%BA%D1%82%D1%96%D0%B2.