1. Постановка задачі.

Умовні позначення:

Нехай – множина респондентів (користувачів), а – множина точок, які були обрані респондентом. Кожний респондент має можливість обрати задану кількість точок з цілочисленими координатами на квадратній площині.

Введемо матрицю

де елемент – координати точки під номером респондента . Загалом, частина точок відома (задана), а частина – невідома.

Отже, задача заключається у оцінці потенціальних значень невідомих елементів за вже відомими (заданими) елементами.

1. Колаборативна фільтрація.

Колаборативна фільтрація – метод, за якого невідомі елементи множини оцінюються лише за відомими елементами тієї ж множини без викроистання додаткової інфромації.

Ідея колаборативної фільрації заключається у тому, що «схожі» люди діють подібним чином.

Найбільш використовуваний метод для данного способу – к найближчих сусідів (*K Nearest Neighbors*).

1. Метод к найближчих сусідів.

Мірою подібності респондентів може слугувати будь-яка метрика на просторі m-мірних векторів, наприклад, індукована евклідова метрика, коефіцієнт кореляції Пірсона (показник лінійної залежності між центрованими векторами), косинуса подібності (показник лінійної залежності між векторами) і т.д. Розрізняють два варіанти реалізації методу к найближчих сусідів: орієнтований на користувачів та орієнтований на об’єкти.

Виходячи з формулювання досліджуваної задачі, розглядатимемо реалізацію, орієнтовану на користувачів.

Ідея методу: знаючи m-вимірні вектори послідовного вибору точок на площині для кожного респондента, можна встановити між ними міру подібності та представити невідоме значення точки кожного користувача через лінійну комбінацію (зважену суму) його сусідів.

Коефіцієнт кореляції Пірсона між користувачами розраховується з урахуванням дисперсії:

Коефіцієнт кореляції Спірмена між двома користувачами розраховується за формулою:

де

Виходячи зі специфіки задачі, оцінювати координати невідомих точок будемо окремо для координати x та для координати y.

З коефіцієнтів подібності складається матриця подібності розміром , де – кількість користувачів. Кожен коефіцієнт матриці – кількісна міра подібності між користувачами та .

Будемо обирати для кожного користувача найбільших коефіцієнтів з матриці , які відповідають найбільш схожим на нього користувачам. Коефіцієнт обирається довільно.

Невідомі значення елементів де – значення точки для найближчих сусідів користувача .

1. Оцінка точності алгоритму.

Найпоширенішим підходом для оцінювання точності алгоритму є розрахунок середньоквадратичної похибки алгоритму.

, де – множина відомих точок користувачів, а – множина наближених оцінок відповідно відомих точок.

1. Використовуючи коефіцієнти Пірсона і Спірмена, було отримані наступні значення середньоквадратичної похибки:

|  |  |  |
| --- | --- | --- |
| Експеримент | КК Пірсона | КК Спірмена |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |

Для порівняння двух реалізацій методу порівняємо значення середньоквадратичних помилок цих реалізацій. Для цього використаємо критерій Стьюдента про рівність середніх при невідомих дисперсіях. Право на використання критерія Стьюдента нам надає центальна гранична теорема, котра говорить, що середнє значення випадковохої величини, отриманої з будь-якого розподілу, має нормальний розподіл.