**ПРОСТОРОВИЙ АНАЛІЗ ТА СТАТИСТИЧНЕ МОДЕЛЮВАННЯ В ГЕОЛОГІЇ**

***Практичне робота 5 (2023)***

***Завдання 1 :*** Дано набір даних і метою є класифікація нової точки (5, 7) за допомогою k-nearest neighbor класифікацію.

*pointsBlue*

|  |  |
| --- | --- |
| X | Y |
| 8 | 2 |
| 7 | 2 |
| 6 | 5 |
| 7 | 2 |
| 5 | 6 |
| 9 | 4 |
| *... (14 rows omitted)* | |

*pointsYellow*

|  |  |
| --- | --- |
| X | Y |
| 6 | 5 |
| 4 | 3 |
| 1 | 6 |
| 1 | 6 |
| 3 | 4 |
| 5 | 5 |
| *... (14 rows omitted)* | |

Ось діаграма розсіювання даних. Спробуємо класифікувати точку (5, 7).



Який колір точки (5, 7), якщо Ви використовуєте класифікатор з **3** найближчими сусідами?

*2 сині, 1 жовта*





Який колір точки (5, 7), якщо Ви використовуєте класифікатор з **5** найближчими сусідами?

*3 сині, 2 жовті*





Який колір точки (5, 7), якщо Ви використовуєте класифікатор з **7** найближчими сусідами?

*5 сині, 6 жовті – вибір залежить від того, яку точку додати сьомою раніше*





***Завдання 2 :*** Оберіть Вашу відповідь для наступних тверджень:

У класифікації методом k-найближчих сусідів збільшення значення k завжди підвищить точність класифікації для тестової вибірки?





У класифікації методом k-найближчих сусідів збільшення значення k ніколи не підвищує точність класифікації для тестової вибірки?

*Збільшення k зробить класифікацію* ***більш загальною*** *(biased), а зменшення призведе до перенавчання і росту помилок (variance), тому вибір правильного k – це пошук компромісу.*





У класифікації методом k-найближчих сусідів слід вибирати тестову вибірку, яка включає деякі точки даних із навчальної вибірки та деякі точки даних, яких немає в навчальній вибірці.

*Тестова вибірка повинна складатися виключно з точок даних, яких нема в навчальній*





***Завдання 3 :*** Ми хочемо спрогнозувати, чи є об’єкт аномалією (клас 0) чи ні (клас 1), на основі двох атрибутів: (а) ***форми*** та (б) ***градієнту***. У нас є навчальна вибірка із 9 об’єктів і тестову вибірку із 2 об’єктів, як показано на діаграмі нижче. Немає двох об’єктів однакової форми і градієнтами.



Перший об’єкт із тестової вибірки має градієнт 90 (1/см2) і форму 40 (безрозмірна величина). Який клас передбачить класифікатор найближчого сусіда з 1 найближчим сусідом для цього об’єкту?





Який клас передбачить класифікатор з 3 найближчими сусідами для цього об’єкту?





Який клас передбачить класифікатор з 5 найближчими сусідами для цього об’єкту?





***Завдання 4 :*** Припустимо, у нас є ще ***два об’єкти***: один формою 41 (безрозмірна величина) і градієнтом 100 (1/см2), а інший формою 52 (безрозмірна величина) і градієнтом 100 (1/см2). Використовуючи визначення відстані, надане в лекції, знайдіть відстань між цими двома об’єктами для використання методу k-найближчих сусідів, якщо обрами форму та градієнт як єдині дві ознаки.

Який клас передбачить класифікатор з 3 найближчими сусідами для об’єкту формою 41 (безрозмірна величина) і градієнтом 100 (1/см2)?





Який клас передбачить класифікатор з 3 найближчими сусідами для об’єкту формою 52 (безрозмірна величина) і градієнтом 100 (1/см2)?

*рішення залежить від того, яку точку додати третьою раніше*





***Завдання 5:***

Виконайте інструкції наведені в файлі *pr5\_5.ipynb* з директорії pr5\_5 (архіву pr5\_5.zip). Приєднайте до вашого проекту JUPYTER NOTEBOOK систему контролю версій Git та розмістить його на GitHub надавши посилання нижче:

[*Вставити посилання на Вашу Практичну робота 5 (2023) завдання 5\_5 на GitHub*](https://github.com/mashaven/sheva-stats/blob/7d449e0ecba60d6cffa290b022adf25a5622da16/task-5/pr5_5/pr5_5.ipynb)