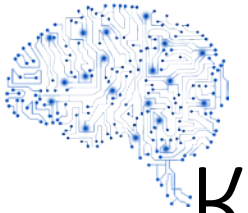
Метричні методи класифікації та регресії. Алгоритм к найближчих сусідів

Професор, д.е.н. Ставицький А.В.

Регресія

- Регресія — розпізнавання чисельної (скалярної або векторної) характеристики об'єкта
- Строге математичне визначення регресії — це умовне математичне сподівання однієї випадкової величини відносно іншої
- В багатьох прикладних задачах (розпізнавальна) регресія є регресією математичною



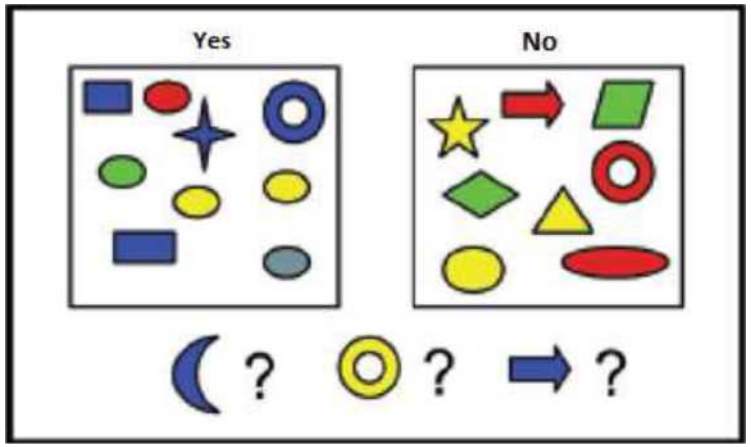
Класифікація

- Класифікація — розпізнавання якісної (дискретної) характеристики об'єкта
- Клас — це множина об'єктів, для яких ця характеристика приймає певне (визначене) значення
- Відповіддю класифікатора для кожного об'єкта краще вважати не номер класу, до якого класифікатор відносить об'єкт, а вектор «впевненості» (confidence в приналежності об'єкта кожному з класів. Тим самим, класифікація перетворюється в спеціальний випадок регресії

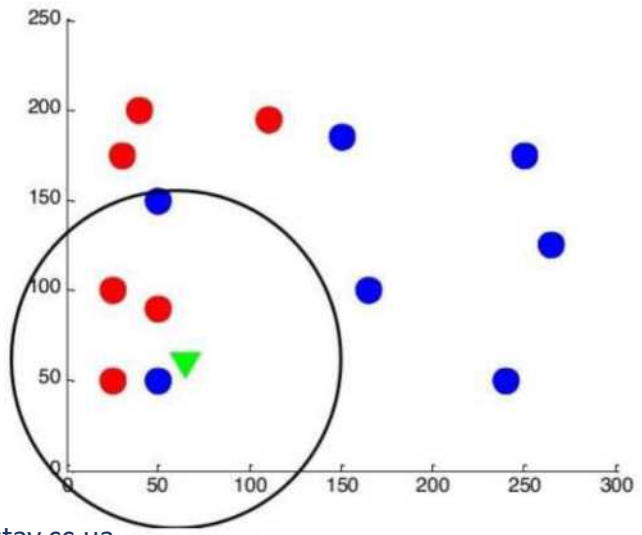
Distance-based vs similarity-based

- Класифікація проводиться шляхом відношення невідомого до відомого за деякою функцією відстані / подібності

similarity-based
(e.g. decision trees)



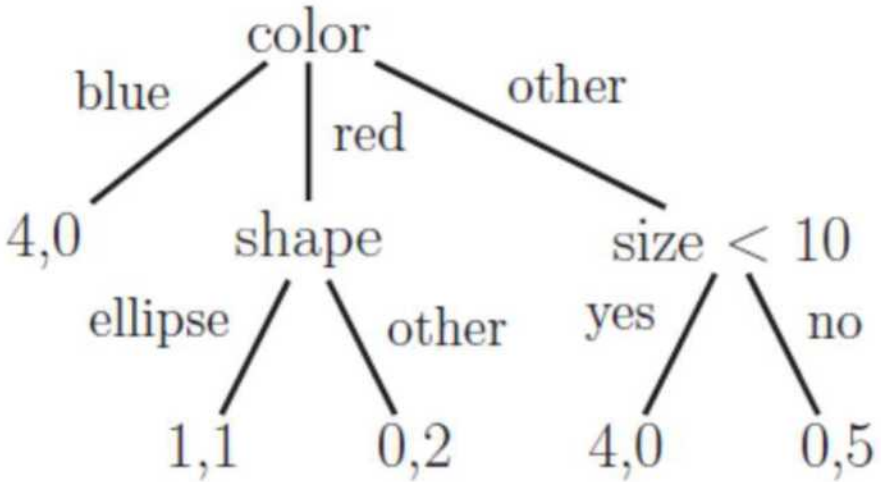
distance-based
(e.g. kNN)



Eager learners vs Lazy learners

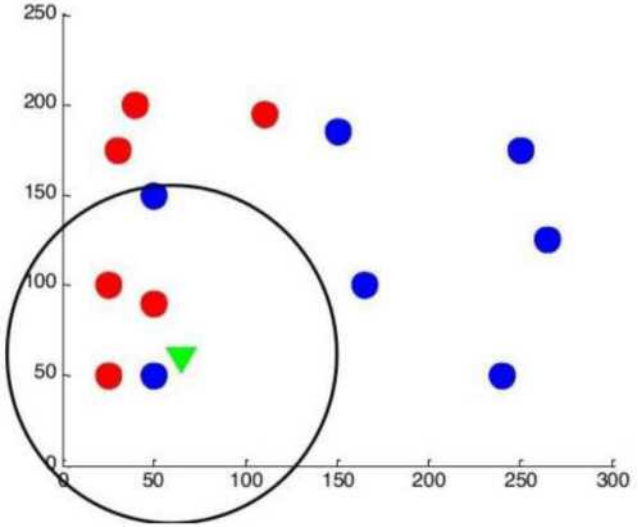
Eager learners

(e.g. decision trees, Bayesian classifiers, Neural networks)



Lazy learners

(e.g. *A*-nearest neighbor, *A*-means clustering)

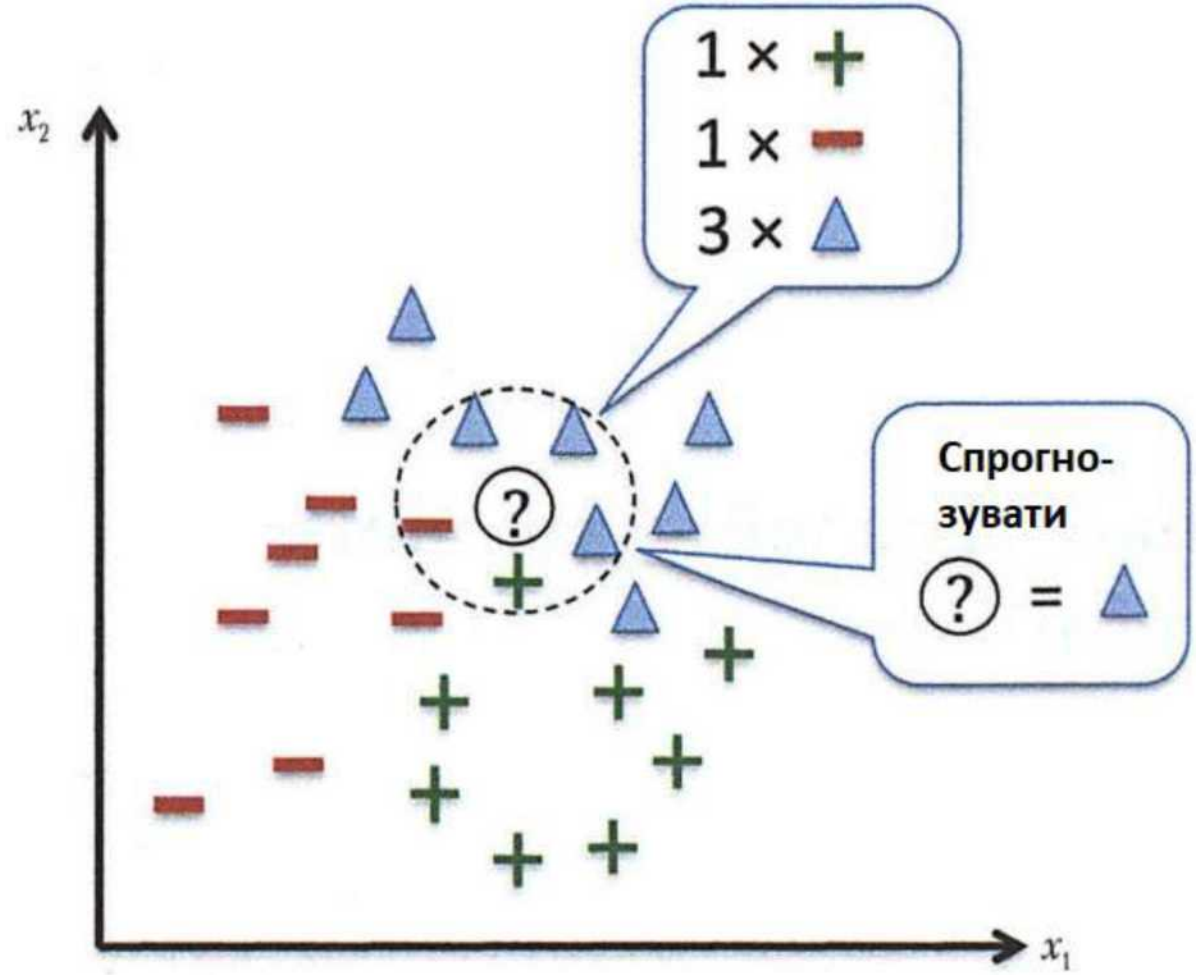


Проста аналогія

- Розкажіть мені про своїх друзів (хто ваші сусіди), і я скажу вам, хто ви

Алгоритм к найближчих сусідів

- Вибрати число k і метрику відстані
- Знайти k найближчих сусідів зразка, який ми хочемо класифікувати
- Присвоїти мітку класу мажоритарним голосуванням



Алгоритм к найближчих сусідів

Customer	Age	Income	No. credit cards	Class
George	35	35K	3	No
Rachel	22	50K	2	Yes
Steve	63	200K	1	No
Tom	59	170K	1	No
Anne	25	40K	4	Yes
John	37	50K	2	YES

Distance from John

sqrt $[(35-37)^2+(35-50)^2+(3-2)^2]=15.16$

sqrt $[(22-37)^2+(50-50)^2+(2-2)^2]=15$

sqrt $[(63-37)^2+(200-50)^2+(1-2)^2]=152.23$

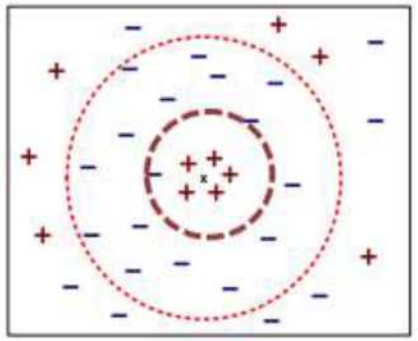
sqrt $[(59-37)^2+(170-50)^2+(1-2)^2]=122$

sqrt $[(25-37)^2+(40-50)^2+(4-2)^2]=15.74$

- k = 1, може вплинути шум (15 і 15,16)
- k = 5, викривлення, бо «захоплюється» багато «сусідів»

Як вибрати k?

- Якщо k занадто мале, то алгоритм чутливий до шумових точок
- Більше k працює добре, але занадто велике k може включати більшість точок з інших класів
- На практиці хороших результатів для більшості даних наборів даних із невеликою кількістю розмірів можна отримати значення, що знаходяться між 5 і 10
- Значення для k можна також вибрати через перехресну перевірку
- Розпізнавання методом k = 1 найближчого сусіда не робить жодної помилки на пред'явленому йому вхідному наборі даних (на навчальній вибірці), але може помилятися на невідомих йому векторах ознак
- Розпізнавання методом k > 1 найближчих сусідів не обов'язково безпомилково розпізнає точки навчальної вибірки, зате при невеликих k, зазвичай, менше помиляється на невідомих йому векторах



«Ліниве» машинне навчання

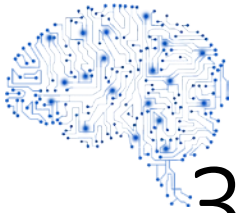
- Навчання класифікатора (чи розпізнавача) при методі k найближчих сусідів є тривіальним і зводиться до запам'ятовування навчальної вибірки
- Розпізнавання теж тривіальне, але є дуже складним, і трудомісткість зростає пропорційно обсягу навчальної вибірки
- Погано: на навчальній виборці — швидко, на тестовій виборці — повільно!

Алгоритм к найближчих сусідів

- Для класифікації:
  - Вибрати число k та метрику відстані
  - Знайти k найближчих сусідів зразка, який ми хочемо класифікувати
  - Присвоїти мітку класу мажоритарним голосуванням
- Для регресії:
  - Вибрати число k та метрику відстані
  - Знайти k найближчих сусідів зразка, для якого ми хочемо знайти значення
  - Присвоїти їх середньоарифметичне значення

Підбір вагів

- Параметр k зазвичай налаштовується за допомогою крос-валідації.
- У класичному методі до найближчих сусідів усі об'єкти мають одиничні значення вагів. Такий підхід, однак, не є розумним.
- Якщо один з сусідів знаходиться занадто далеко, то він не повинен відображати сильний вплив на відповідь. Ця ідея реалізується за допомогою важелів, оберено пропорційним відстані.



Зашумлені характеристики

- Зашумлені характеристики можуть зробити сильний вплив на метрику
- Виявити такі характеристики можна, видаляючи по черзі всі характеристики і дивлячись на помилку на тестовій вибірці.

Нормалізація характеристик

- Помножимо одну з характеристик (наприклад, першу) на константу C
- Евклідова відстань прийме наступний вид

$$\rho_2(x, y) = \sqrt{C(x_1 - y_1)^2 + \sum_{i=2}^{d}(x_i - y_i)^2}$$

- Таким чином, відмінність за першою характеристикою буде вважатися в C раз більш значущою, ніж відмінності за всіма іншими. При цьому розташування об'єктів один відносно іншого не змінилося — змінився лише масштаб (тобто міряємо у метрах чи кілометрах, наприклад)!

Нормалізація характеристик

Широко застосовуються такі способи:

- нормування на середньо квадратичне відхилення (СКВ):

$$x_{\text{норм. } i} = \frac{\text{x}_i - x_{\text{сер } i}}{\text{СКВ}(\text{x}_\text{i})}$$

- нормування на відрізок [0; 1] шляхом ділення на розмах:

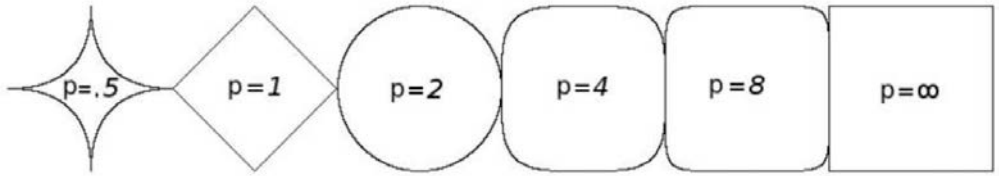
$$x_{\text{норм. } i} = \frac{\text{x}_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

- у цих формулах x_i - це вектор, складений із і-тих ознак усіх об’єктів (тобто це і-та колонка матриці «об’єкти-ознаки»)

Метрика Мінковського

$$\rho_p(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p}$$

- Окремі випадки метрики Мінковського:
  - (p =2) Евклідова метрика, визначає відстань як довжину прямої, що з'єднує дві визначені точки в одно-, два-.....багатовимірному просторі
  - (p =1) Манхеттенська відстань, визначає відстань як мінімальну довжину шляху між визначеними точками за умови, що можна рухатися лише паралельно осям координат
  - (p=<n) Метрика Чебишева, визначає відстань як максимальну із відстаней між координатами цих двох точок: $\rho_\infty(x, y) = \max_{i=1,\ldots,d} |x_i - y_i|.$



- По мірі збільшення параметра *p* метрика слабкіше штрафує невеликі відмінності між векторами і сильніше штрафує значні відмінності

Косинусна міра

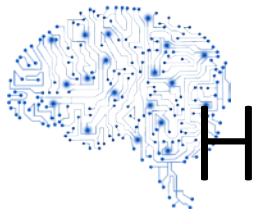
$$\rho_{\cos}(x, y) = \arccos\left(\frac{\langle x, y\rangle}{\|x\|\|y\|}\right) = \arccos\left(\frac{\sum_{i=1}^{d} x_i y_i}{\left(\sum_{i=1}^{d} x_i^2\right)^{1/2}\left(\sum_{i=1}^{d} y_i^2\right)^{1/2}}\right)$$

- Косинусна міра використовується для вимірювання схожості між текстами
- Кожен документ описується вектором, кожна компонента якого відповідає слову зі словника
- Компонента дорівнює одиниці, якщо відповідне слово зустрічається в тексті, і нулю в іншому випадку

Відстань Жаккарда

$$\rho_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

- Якщо об'єктами є множини (наприклад, кожен об'єкт — це текст, представлений множиною слів), то їх відмінність/схожість можна також вимірювати за допомогою відстані Жаккарда



Наближені методи пошуку найближчих сусідів

- Два способи боротьби з високою складністю пошуку найближчих сусідів при великій кількості ознак:
  1. Запам'ятовувати не всю навчальну вибірку, а лише її представницьку підмножину, вибрану за певною евристикою (наприклад, алгоритм STOLP)
  2. Шукати k найближчих сусідів наближено, тобто дозволяти результату пошуку бути трохи далі від нового об'єкта, ніж k його справжніх сусідів

Ідея алгоритму STOLP

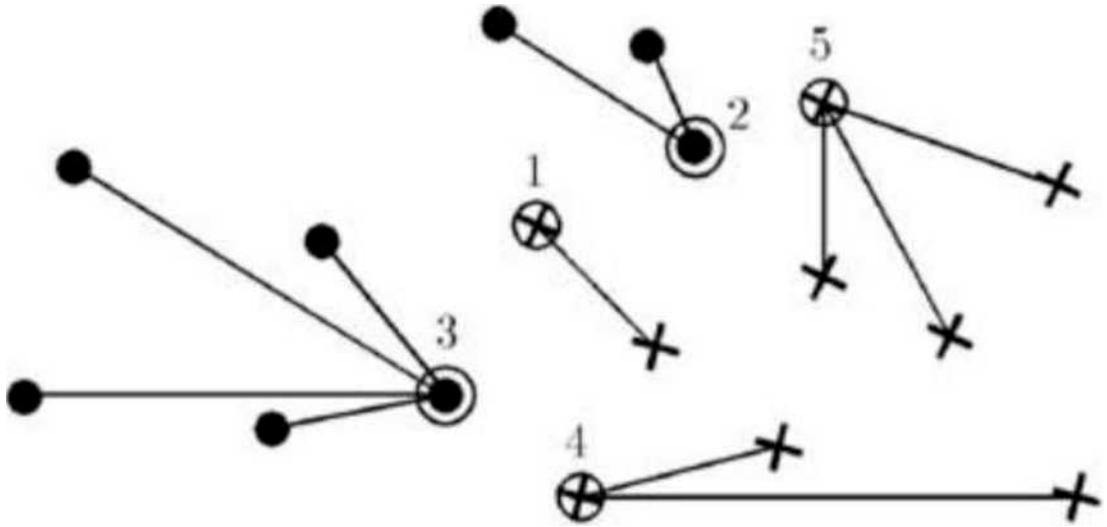
- Залишити із навчальної вибірки тільки прецеденти (=«опорні точки») класу, які забезпечують виконання такої умови:
  - відстань від будь-якої точки навчальної вибірки певного класу до найближчого свого прецеденту менше відстані до найближчого прецеденту іншого класу
- Такий набір прецедентів забезпечить безпомилкове розпізнавання всіх реалізацій навчальної вибірки
- Але пошук прецедентів має комбінаторний характер і в загальному випадку вимагає повного перебору всіх варіантів

Алгоритм STOLP

1. Для кожної точки навчальної вибірки визначаються відстані до найближчої точки свого класу (r_{in}) і найближчої точки інших класів (r_{out}).
2. Спочатку знаходяться «напружені» прикордонні точки, тобто серед точок кожного класу вибирається по одній точці з максимальним значенням величини $W=r_{in}/r_{out}$ ризику для даної точки бути розпізнаної як точка іншого класу. Ці точки заносяться у список прецедентів.
3. Потім робиться пробне розпізнавання всіх точок навчальної вибірки з опорою на прецеденти і з використанням правила найближчого сусіда: точка відноситься до того класу, відстань до прецеденту якого є мінімальною.
4. Серед точок, розпізнаних неправильно, вибирається точка з максимальним значенням ризику W і нею поповнюється список прецедентів, після чого повторюється процедура пробного розпізнавання всіх точок.
5. Так триває до тих пір, поки всі точки навчальної вибірки не стануть розпізнаватися без помилок

Приклад

- У прикладі на рисунку першими прецедентами були обрані точки 1 і 2
- Потім список прецедентів поповнився точками 3, 4 і 5
- Але якщо класів багато, то алгоритм стає дуже ресурсозатратним



Питання?