Національний технічний університет України «КПІ ім. Ігоря Сікорського»

Факультет Інформатики та Обчислювальної Техніки

Кафедра інформаційних систем та технологій

Лабораторна робота №3

з дисципліни «Прикладні задачі машинного навчання»

на тему

«Класифікація, регресія і кластеризація з використанням бібліотеки scikit-learn»

15 варіант

Виконав:

студент групи ІС-21 Костюк А. С.

Викладач:

Нестерук А.О.

Київ – 2024

**Зміст**

[**1. Постановка задачі** 3](#_Toc163144401)

[**2. Виконання** 4](#_Toc163144402)

[**3. Висновок** 15](#_Toc163144403)

**1. Постановка задачі**

1. Повторити дії описані в пункті «Часові ряди і проста лінійна регресія частина 2» даної лабораторної роботи та порівняти з результатом попередньої лабораторної роботи.

2. Аналогічно з прикладом з лекції згенеруйте набір даних та класифікуйте його використавши класифікатор SVC (слайд).

3. Порівняти декілька класифікаційних оцінювачів наприклад KNeighborsClassifier, SVC та GaussianNB для вбудованого в scikit-learn одного набору даних (вибрати довільний за бажанням).

4. Зробити звіт про роботу.

**2. Виконання**

1) **Повторити дії описані в пункті «Часові ряди і проста лінійна регресія частина 2» даної лабораторної роботи та порівняти з результатом попередньої лабораторної роботи.**

Зображення, що містить текст, знімок екрана, програмне забезпечення

Автоматично згенерований опис

При виборі одного стовпчика в двовимірному DataFrame результат являє собою одновимірну колекцію Series. Однак оцінювачі scikit-learn вимагають, щоб в якості навчальних і тестових даних використовувалися двовимірні масиви (або двовимірні структури, подібні до масивами, наприклад списки списків або колекції pandas DataFrame). Щоб використовувати одномірні масиви з оцінювачем, необхідно перетворити їх з одновимірного масиву з n елементами в двовимірний масив c n рядками і одним стовпцем. Дані розбиваємо на навчальний і тестовий набори. Ключовий аргумент random\_state використовується для забезпечення відтворюваності результатів:

Зображення, що містить текст, знімок екрана, Шрифт, програмне забезпечення

Автоматично згенерований опис

Для перевірки пропорції навчальних тестових даних (75% до 25%) задамо розміри X\_train і X\_test (Так як у нас дані до 2023, маємо 129 значень):

Зображення, що містить текст, знімок екрана, Шрифт, число

Автоматично згенерований опис

У scikit-learn немає окремого класу для простої лінійної регресії, тому що проста лінійна регресія є окремим випадком множинної лінійної регресії, тому ми скористаємося оцінювачем LinearRegression:

Зображення, що містить текст, знімок екрана, Шрифт, ряд

Автоматично згенерований опис

Значення кута нахилу і точки перетину з віссю, що використовуються у формулі y = mx + b, можуть використовуватися для прогнозування. Кут нахилу зберігається в атрибуті coeff\_ оцінювача (m у формулі), а точка перетину - в атрибуті intercept\_ (b у формулі):

Зображення, що містить текст, знімок екрана, Шрифт

Автоматично згенерований опис

Протестуємо модель за даними з X\_test і перевіримо прогнози по набору даних, виводячи прогнозовані і очікувані значення для кожного п'ятого елементу:

Зображення, що містить текст, знімок екрана, Шрифт, програмне забезпечення

Автоматично згенерований опис

Скористаємося отриманими значеннями кута нахилу і точки перетину для прогнозування середньої температури в березні 2024 року, а також оцінки середньої температури в березні 1890 року:

Зображення, що містить текст, знімок екрана, програмне забезпечення, монітор

Автоматично згенерований опис

Тепер побудуємо діаграму розкиду даних за допомогою функції scatterplot бібліотеки Seaborn і функції plot бібліотеки Matplotlib. Для виведення точок даних скористаємося методом scatterplot з колекцією DataFrame з ім'ям avrMarchTemp.

Зображення, що містить текст, знімок екрана, Шрифт, ряд

Автоматично згенерований опис

Встановимо масштаб осі y від 20 до 60.



Перейдемо до висновку регресійної прямої. Почнемо зі створення масиву, що містить мінімальні і максимальні значення дати з avrMarchTemp.Date. Вони стануть координатами x початкової і кінцевої точок регресійної прямої:

Зображення, що містить знімок екрана, текст, Шрифт, ряд

Автоматично згенерований опис

В результаті передачі predict масиву x буде отримано масив відповідних прогнозованих значень, які будуть використовуватися як координати y:



Нарешті, функція plot бібліотеки Matplotlib малює лінію по масивам x та y, що зображує координати x та y точок відповідно:

Зображення, що містить Шрифт, текст, знімок екрана, Графіка

Автоматично згенерований опис

Результат:

Зображення, що містить знімок екрана, текст, ряд, Графік

Автоматично згенерований опис

Можна побачити, що отримана діаграма розкиду даних практично ідентична тій, що була отримана в 2 лабораторній роботі.

2) **Аналогічно з прикладом з лекції згенеруйте набір даних та класифікуйте його використавши класифікатор SVC (слайд).**

Генерація набору даних:

Зображення, що містить текст, знімок екрана, Шрифт, програмне забезпечення

Автоматично згенерований опис

Результат:

Зображення, що містить знімок екрана, Барвистість, схема, ряд

Автоматично згенерований опис

Класифікуємо дані, використавши класифікатор SVC. Для цього спочатку напишемо функцію призначену для візуалізації границі прийняття рішень класифікатором на основі двох ознак. Функція отримує набір даних X та відповідний набір міток класів y, класифікатор classifier та, якщо використовується тестовий набір даних, індекс тестових даних test\_idx.

Функція генерує рішення класифікатора на сітці точок з розміром кроку resolution, і відображає його на графіку за допомогою методу contourf з використанням переданої колірної мапи cmap. Мітки класів відображаються за допомогою методу scatter, де кожен клас має свій маркер marker та колір color. Якщо використовується тестовий набір даних, то його точки позначаються окремим маркером:

Зображення, що містить текст, знімок екрана, програмне забезпечення, Шрифт

Автоматично згенерований опис

Зображення, що містить текст, знімок екрана, програмне забезпечення

Автоматично згенерований опис

Результат:

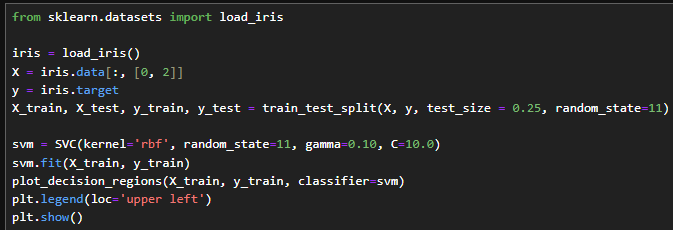
Зображення, що містить текст, знімок екрана, схема

Автоматично згенерований опис

3) **Порівняти декілька класифікаційних оцінювачів наприклад KNeighborsClassifier, SVC та GaussianNB для вбудованого в scikit-learn одного набору даних (вибрати довільний за бажанням).**

Порівняємо класифікаційні оцінювачі SVC, KNeighborsClassifier та GaussianNB за допомогою вбудованого в scikit-learn набору даних iris.

Класифікатор SVC:



Результат:

Зображення, що містить текст, знімок екрана, схема, Графік

Автоматично згенерований опис

Класифікатор KNeighborsClassifier:

Зображення, що містить текст, знімок екрана, Шрифт

Автоматично згенерований опис

Результат:

Зображення, що містить текст, знімок екрана, схема, карта

Автоматично згенерований опис

Класифікатор GaussianNB:

Зображення, що містить текст, знімок екрана, Шрифт, монітор

Автоматично згенерований опис

Результат:

Зображення, що містить текст, знімок екрана, схема

Автоматично згенерований опис

**3. Висновок**

В даній лабораторній роботі були розглянуті класифікація, регресія і кластеризація з використанням бібліотеки scikit-learn, обробка та візуалізація яких була здійснена за допомогою Python та його бібліотек scikit-learn, Matplotlib, NumPy, seaborn та pandas. Отримані знання були використані для обробки даних середньої температури в березні 1895-2023 років в місті Spokane штату Washington, отриманий результат збігається з результатом 2 лабораторної, що підтверджує правильність виконання поставленої задачі. Також була розглянута та проведена кластеризація та класифікація певних наборів даних за допомогою бібліотеки scikit-learn, а саме використовуючи класифікатори SVC, KNeighborsClassifier та GaussianNB. Дані, які оброблялися: випадково згенерований набір даних та вбудований в бібліотеку scikit-learn набір iris. Результати цих дій представлені в лабораторній.