Регресія

Мета роботи

Ознайомитись з основами машинного навчання та аналізу даних для розв'язання задачі регресії, реалізувати методи, що базуються на алгоритмі XGBoost. Застосування регуляризації (Lasso, Elastic Net тощо).

Рекомендована література

Про роботу з лінійними регресіями в Python:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

https://scikit-learn.org/stable/modules/linear_model.html

https://realpython.com/linear-regression-in-python/

XGBoost:

https://drive.google.com/file/d/1kuyu5L_1Hwb7QQFsyIpPPCTEwaO9Vj4X/view?usp=sharing

PyTorch/Tensorflow/Keras/SciKit Learn:

https://drive.google.com/file/d/1faM-aL5Bsea5DcXpFVr6ApOL5qWNWved/view?usp=sharing

https://drive.google.com/file/d/1tAD_-4hZ1nqe5Fqfyic92hmyrQlZnzPe/view?usp=sharing

https://drive.google.com/file/d/1D5m901531N9-v-UZQC77YhtRCTdbhAG4/view?usp=sharing

https://drive.google.com/file/d/10-cRqjUrSp_qOquyjKLuzqO8KPA9yA3r/view?usp=sharing

Ансамблінг у SciKit learn:

https://scikit-learn.org/stable/modules/ensemble.html#voting-regressor

Хід роботи

Дані

- Для отримання **максимального** балу: Використовувати один з публічних наборів даних з https://www.kaggle.com/datasets
- Для отримання балу **не вище 75%**: можна використовувати типові ідеальні набори даних (типу ірисів Фішера і т.п.)

Підготовчий етап (для всіх рівнів)

1. Провести аналіз вибраного набору даних,
2. визначити вхідні та вихідні параметри,
3. візуалізувати залежності входів та виходу,
4. детектувати аномалії, неповні зразки у даних, тощо,
5. провести кореляційний аналіз входів та виходів набору даних, виявити взаємозалежні фактори.
6. Провести підготовку даних до подальшого використання.

7. Обраний та відфільтрований набір даних розбити на навчальну та тестову частину (70% на навчання, 30% на тест).

Реалізація моделі

Написати код у Python / R, який реалізований з використанням **двох** (для отримання балу не вище 75%) чи **трьох** (на максимальний бал) підходів на вибір:

1) SciKit learn
2) XGBoost
3) Vanilla Python (Numpy/Pandas)

Для отримання **максимального балу** за лабораторну роботу заборонено використовувати бібліотеки з вже реалізованими аналогічними алгоритмами (потрібно реалізовувати алгоритми самостійно). Вбудовані алгоритми використовувати лише для порівняння власно-запрограмованого алгоритму.

Аналіз результатів

1. Вибір оптимальних параметрів регресій, їх обгрунтування
2. Оцінка помилок на начальній та тестовій вибірках
3. Порівняння результатів різних підходів (співпали чи ні, причини чому могли не співпасти, runtime тощо)
4. Порівняти результати з аналогічними результатами, які отримані в результаті використання вбудованих функцій
5. Результати оформити протоколом

Студенти, що розраховують на високий бал мають:
1. Прокоментувати код (що кожна строка робить, окрім import / library)
2. Якісно візуалізувати результати та проміжні етапи (за необхідністю)
3. Обрати порівняно складний датасет, який потребує додаткового очищення
4. Реалізовувати алгоритми самостійно
5. Ансамблювання виходів реалізованих методів за прикладом у посиланні з розділу “Рекомендована література” - додатковий бал.

Контрольні питання

1. Що таке ансамблювання та для чого воно використовується?
2. Які моделі базуються на принципі лінійної регресії?
3. Що таке XGBoost?