Частина 1. **Лабораторні роботи до курсу Регресійний аналіз**

1. Загальні відомості

Даний документ містить інформацію щодо лабораторної роботи №1 з алгоритмів машинного навчання. Всього передбачено 12 варіантів. Ваш варіант відповідає Вашому номеру у списку з журналу (якщо Ваш номер більше 12, то відніміть від Вашого номера 12).

Разом з цим документом надіслано архів що містить .csv, файли з даними, які слід аналізувати, кожному варіанту відповідає один файл. У кожному варіанті буде вказана змінна - відгук, а також змінні - регресори. Пояснення до даних можна знайти у файлі DataDescriptions.pdf. Самі дані взяті з сайту http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html, що відповідає книзі: Edward W.: Regression Modeling with Actuarial and Financial Applications.

Для проведення **регресійного** аналізу потрбіно зробити наступне:

1) Побудувати ОНК, зробити висновки, щодо якості моделі та ОНК.

2) Спробувати покращити оцінку, шляхом використання гребеневої регресії.

3) Спробуйте покращити оцінку додавши у модель нелінійність.

4) З'ясуйте чи можна зменшити кількість регресорів без суттєвої шкоди для моделі.

Також пропонується додаткове завдання 5. Завдання 5 є опціональним і виконується за бажанням студента.

5) Напишіть функцію, що примає число N, а також величини $\beta_i,\ i = 0:3$ як параметри. Функція повинна згенерувати три регресора $X_i,\ i = 1,3$ зі стандартним нормальним розподілом, а також похибку з розподілом $\mathcal{N}(0, 0.1)$ розмірності N. Функція повинна повернути дата фрейм з колонкою $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, а також колонками $X_1,\ X_2,\ X_3$. Після цього обчисліть ОНК використовуючи безпосереднью формулу, бібліотечну функцію або за допомогою власної імплементації методу градієнтного спуску (теж оформленого в окрему функцію). Для різних значень параметра N занотуйте час який потрібен для обчислення ОНК кожним із трьох методів. Знайдіть таке N при якому один із методів буде працювати відчутно повільніше. Для чистоти експерименту проведіть його 10 разів для кожного N.

2. Вимоги до виконання, оформлення та здачі

Роботу можна виконувати в R або Python. Вибірку слід розбити на тестову та тренувальну у пропорціх 20% та 80%.

Робота має бути оформлена у вигляді .pdf (можливо .doc/.docx) файлу який містить всю необхідну інформацію. Роботи потрібно здавати на парах або надсилати на email.

Робота повинна містити код, та його інтерпритацію. Обов'язково має бути вказано:

1. ОНК, наявність кореляцій між регресорами, якість ОНК (її дисперсія, якість прогнозування, наведена діаграма залишків, коефіцієнт детермінації та результати тесту Фішера).
2. Підбір параметра λ, покращення (чи відсутність покращення) якості прогнозу, зменшення дисперсії.
3. Висновок щодо ефективності нелінійної моделі.
4. Метод що використовувася для оптимального відбору, та досягнуті результати.

Кожна робота буде розглядатися на відповідність критеріям описаним вище, та на обгрунтованість прийнятих рішень. Кожен студент, повинен виконати свою роботу самостійно. Ідентичні, або майже ідентичні роботи прийматися до уваги не будуть.

3. Варіанти

Варіант 1
Файл з даними: Chicago.csv
Відгук: `theft`
Регресори: Всі окрім `zipcode`

Варіант 2
Файл з даними: CeoCompensation.csv
Відгук: `COMP`
Регресори: `TENURE, EXPER, SALES, VAL, PCNTOWN, PROF`

Варіант 3
Файл з даними: HealthExpend.csv
Відгук: `EXPENDOP`
Регресори: `AGE,famsize,COUNTIP,COUNTOP,EXPENDIP`

Варіант 4
Файл з даними: NAICExpense.csv
Відгук: `EXPENSES`
Регресори: `RBC, STAFFWAGE, AGENTWAGE, LONGLOSS, SHORTLOSS`

Варіант 5
Файл з даними: NAICExpense.csv
Відгук: `EXPENSES`
Регресори: `GPWPERSONAL, GPWCOMM, ASSETS, CASH, LIQUIDRATIO`

Варіант 6
Файл з даними: HospitalCosts.csv
Відгук: `TOTCHG`
Регресори: `AGE, LOS, APRDRG`

Варіант 7
Файл з даними: RiskSurvey.csv
Відгук: `FIRMCOST`
Регресори: `ASSUME, SIZELOG, INDCOST, CENTRAL, SOPH`

Варіант 8
Файл з даними: UNLifeExpectancy.csv
Відгук: `LIFEEXP`
Регресори: `ILLITERATE POP FERTILITY PRIVATEHEALTH HEALTHEXPEND BIRTHATTEND PHYSICIAN GDP`
Коментар: В даному файлі деякі дані пропущені. Запропонуйте варіант розв'язання цієї проблеми.

Варіант 9
Файл з даними: WiscHospCosts.csv
Відгук: `TOT_CHG`

Регресори: NO_DSCHG POPLN NUM_BEDS INCOME CHG_NUM

Варіант 10
Файл з даними: WiscLottery.csv
Відгук: SALES
Регресори: PERPERHH MEDSCHYR MEDHVL PRCRENT PRC55P HHMEDAGE MEDINC POP

Варіант 11
Файл з даними: Medicare.csv
Відгук: COV_CHG
Регресори: TOT_CHG MED_REIB TOT_D NUM_DCHG AVE_T_D

Варіант 12
Файл з даними: MedCPISmooth.csv
Відгук: value
Регресори: PerMEDCPI YEAR MCPISM4 MCPISM8 MCPISMw_2 MCPISMw_8