ЗАВДАННЯ для Лабораторної роботи №3

Метою роботи є пошук моделі оптимальної складності, причому наперед не відомо, яку саме підмножину всіх регресорів/аргументів містить така модель. Тому необхідно сконструювати і перевірити алгоритми для її пошуку.

1. Задано: вибірка *n*, *m*, $X[n\times m]$, $y[n\times 1]$; алгоритм МНКО; критерій C_p.
   Потрібно: запрограмувати алгоритми пошуку моделі оптимальної складності за допомогою МНКО для встановлення того, які саме регресори є зайвими (неінформативними), і вилучити їх з моделі, залишивши в ній тільки інформативні (істинні).
2. Задано число $s_0 < m$ істинних аргументів/регресорів, на яких сформовано істинну модель $y^0 = X^*\theta_0^*$; $X^*[n\times s_0]$ – матриця довільно вибраних (невідомо як розміщених) s_0 стовпців матриці *X*. Решта $m - s_0$ стовпців відповідають зайвим (неінформативним) аргументам.
3. Як і в 2-й роботі, елементи матриці *X* формуються генератором випадкових чисел, вектор $y = y^0 + \xi$, де ξ – випадковий вектор з заданою дисперсією. Для простоти нехай $\sigma^2 = 0.3$ і не змінюється.
4. Один раз обчислюється розширена матриця нормальної системи рівнянь для повної моделі (для всіх *m* регресорів) $W = [X^T X \vdots X^T y]$, $X = [x_1 x_2 \dots x_m]$, яка використовується в усіх алгоритмах. Ця матриця зберігається в початковому вигляді, а всі розрахунки виконуються в робочій/допоміжній матриці $W_{sub}[H \vdots h]$ такої ж розмірності.
   Матриці W та W_{sub} спочатку мають однаковий вигляд:

$x_1^T x_1$	$x_1^T x_2$	…	$x_1^T x_m$	$x_1^T y$
…	…	…	…	…
…	…	…	…	…
$x_m^T x_1$	…	…	$x_m^T x_m$	$x_m^T y$

Запрограмувати 3 алгоритми з використанням МНКО:

5. **A1** (метод *кореляційного включення*):

Крок 1: Порахувати парні кореляції:

$$k(x_j, y), \quad j = \overline{1,m}$$

Крок 2: Впорядкувати кореляції в порядку спадання:

$$k_1 = k_{max}, k_2, \dots, k_m; \;\; k_j > k_{j+1}$$

і запам'ятати відповідний порядок нового розміщення регресорів $j_1, \dots, j_m$, наприклад $[3, 8, 1, \dots, m, \dots, 7]$.

Крок 3: переписати в матрицю W_{sub} рядки і стовпці початкової матриці W в порядку номерів $j_1, \dots, j_m$: в перший рядок і стовпець W_{sub} записуємо з W рядок і стовпець з номером j_1, який відповідає k_1, і т.д.

Крок 4: застосувати МНКО до елементів матриці W_{sub}, тобто знайти параметри послідовності m моделей.

Крок 5: для кожної з цих моделей обчислити вихід моделі $\hat{y}_s,\ s = 1, \dots, m$, та відповідне значення критерію $C_p(s)$.

Крок 6: визначити оптимальну складність s^* з мінімальним значенням критерію, тобто $s^*: Cp(s) \to min$.

Крок 7: запам'ятати s^*, структуру (порядок слідування регресорів) моделі та її параметри θ^* , а також значення критерію як вихід алгоритму.

Це є 1-й варіант *фіксованого* (кореляційного) формування послідовності включення регресорів у модель.

6. **A2** (метод *випадкового включення*):

Крок 1: задати K – число варіантів ланцюжків включення аргументів у модель. Наприклад, K = 20; 50; 100.

Крок 2: K разів випадковим чином генерувати послідовність включення регресорів, щоразу запам'ятовуючи цю послідовність $j_1, \dots, j_m$.

Далі виконуються в циклі K разів ті ж кроки 3 – 7, що і в алгоритмі **A1.** На першому кроці циклу запам'ятовуємо $s^*, j_1, \dots, j_{s^*}, \theta^*, C_p(s^*)$, а на кожному наступному порівнюємо результат із тим, що в пам'яті: якщо отриманий $C_p(s^*)$ виявився меншим, то запам'ятовуємо всю нову інформацію, а в іншому випадку нічого не змінюємо. І таким чином виконуємо усі K ітерацій.

Мета: встановити, якого K достатньо для побудови правильної моделі.

Це є 2-й варіант *випадкового* формування послідовностей включення регресорів у модель.

7. **A3** (метод *перебірного включення*):

Крок 1: Будуємо всі моделі складності 1 (m штук):

$$s = 1:\ \hat{y}_{1j} = \hat{\vartheta}_{1j} x_j,\ \hat{\vartheta}_{1j} = \frac{x_j^T y}{x_j^T x_j},\ \ j = \overline{1, m}$$

Знаходимо j^* як розв'язок задачі $C_{\text{p}}(j) \to \min\limits_{j=\overline{1,m}}$, і переносимо рядок і стовпчик матриці W під таким номером на перші місця матриці W_{sub}.

Крок 2: Будуємо в циклі всі m – 1 моделей складності s = 2, вважаючи, що на першому місці знаходиться аргумент x_{j^*}, а далі додаємо по одному всі інші:

$$s = 2:\ \ \hat{y}_{2j} = \hat{\vartheta}_{1j^*} x_{j^*} + \hat{\vartheta}_{2l} x_l,\ \ \ l = \overline{2, m}$$

Зрозуміло, що при цьому щоразу для оцінювання параметрів використовуємо алгоритм МНКО, а саме його 2-й крок.

Знаходимо l^* як розв'язок задачі $C_{\text{p}}(j^*, l) \rightarrow \min\limits_{l=\overline{2,m}}$, і переносимо рядок і стовпчик під таким номером на другі місця.

І так далі, до $s = m$, фактично виконуючи в циклі по s від 1 до m операції кроків 3 – 5 з першого алгоритму.

При цьому для кожного s запам'ятовуємо «рекордну» для цієї складності s інформацію $s, j_1, \dots, j_s, \hat{\theta}_s, C_p(s)$, причому на першому кроці запам'ятовуємо те, що стосується $s = 1$, а на кожному наступному порівнюємо результат із тим, що в пам'яті: якщо отриманий $C_p(s)$ виявився меншим, то записуємо всю нову інформацію, а в іншому випадку нічого не змінюємо. В кінці циклу (при $s = m$) фактично отримаємо всю інформацію про модель оптимальної складності: $s^*, j_1, \dots, j_{s^*}, \theta^*, C_p(s^*)$.

Таким чином тут загалом отримуємо перебір $[m + (m-1) + (m-2) + \cdots + 1]$ (знайдіть формулу цієї суми!) варіантів моделей різної структури, що дає порядок складності m^2 замість експоненційної складності 2^m при повному переборі всіх можливих варіантів структур моделей.

Це 3-й варіант *перебірного*, покрокового, спрямованого формування послідовності включення регресорів у модель («найскоріший спуск»).

Зауваження: Замість другого та/або третього алгоритму можна запропонувати свій власний – наприклад, на основі генетичних операторів.

8. Основне Завдання:

1) Виконати Лабораторну роботу №2 для того ж Тесту 1:

$$y^0 = 3x_1 - 2x_2 + x_3 \mid s_0 = 3,\ m = 5$$

але регресори з розмістити у зворотному порядку, тобто послідовність регресорів $[1, 2, 3, 4, 5]$ замінити на $[5, 4, 3, 2, 1]$. Побудувати графіки трьох критеріїв – їх мінімуми повинні бути при s=5, адже значущими регресорами є 3, 2, 1, а 5 і 4 – зайвими. Це означає, що заданий у початковій вибірці порядок аргументів, як правило, не дозволяє вилучити неінформативні аргументи.

2) Запрограмувати вказані 3 алгоритми А1, А2 та А3 для пошуку моделі оптимальної складності (або замість А2 запропонувати власний алгоритм).

3) Подати на вхід кожної програми Тест 1 зі зворотним порядком регресорів, і дослідити, як вони «відкривають» істинну структуру моделі.

4) Дослідити алгоритми на великому тестовому прикладі $m = 15;\ s_0 = 10$:

$y^0 = \vartheta_1^0 x_1 + \cdots + \vartheta_{s^0}^0 x_{s^0}$, де ϑ_0^0 – випадкові цілі числа: $\vartheta_j^0 \in [1, 10]$.

5) Для кожного прикладу розмістити отримані результати так, щоб їх зручно було порівнювати, і зробити належні висновки щодо порівняльної ефективності «відкриття» істинної структури моделі трьома алгоритмами.

Увага: Заохочується розв'язання власної *реальної* задачі: + від 2 до 5 балів!