Завдання на лабораторну роботу №1 з дисципліни «Інтелектуальний аналізу даних та штучний інтелект»

Синтез оптимальної регресійної моделі за експериментальними даними

В лабораторній роботі відновлюється двофакторна регресійна залежність $y = f(x_1, x_2)$ для однієї із задач з репозиторію автоматичного навчання Каліфорнійського університету в Ірвіні – UCI Machine Learning Repository. Лабораторна робота полягає у виконанні таких завдань.

1. Сформулювати змістовну постановку задачі згідно варіанту з табл. 1.
2. Розбити дані на навчальну та тестову вибірки та перевірити їх репрезентативність.
3. Зобразити експериментальні дані у формі однофакторних залежностей.
4. Віднайти коефіцієнти п’яти регресійних залежностей, серед яких 2 моделі слід згенерувати самостійно. Обов’язковими є такі 3 моделі:

$$y = a_0 + a_1 x_1 + a_2 x_2;$$

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1^2 + a_4 x_2^2;$$

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1^2 + a_4 x_2^2 + a_5 x_1 x_2.$$

5. Обрати оптимальну регресійну модель та обґрунтувати свій вибір. Окрім точності врахувати інтерпретабельність та складність моделей. Візуалізувати оптимальну регресійну модель разом з експериментальними даними.
6. Почергово додати у постановку задачі ще одну вхідну зміну, і побудувати усі можливі лінійні регресійні моделі з трьома змінними та порівняти їх моделлю з попереднього завдання.
7. Розробити лінійну регресійну модель з усіма кількісними вхідними змінними та порівняти з іншими моделями.
8. Знайти моделі інших дослідників і порівняти власні результати з результатами конкурентів.

Додаткові завдання

9. Розробити лінійні регресійні моделі для усіх пар вхідних змінних, обрати кращу та порівняти з іншими моделями **(+5 балів)**.
11. Обрати одну із моделей та дослідити як вона зміниться за робастного регресійного аналізу. Ідентифікувати викиди і проаналізувати їх. **(+5 балів)**.
12. Дослідити залежність RMSE від кількості вхідних змінних моделі. Складність моделі нарощувати за жадібним алгоритмом від меншого значення до більшого. Обрати кращу та порівняти з іншими моделями **(+8 балів)**. Додатково реалізувати жадібний алгоритм за зворотною схемою (шляхом вилучення найгіршої змінної), дослідити залежність RMSE від кількості вхідних змінних моделі та порівняти з іншими моделями **(+2 бали)**.

Можливий полегшений варіант лабораторної без завдань 6 та 7. Він оцінюється зі ***знижкою у 5 балів****.*

Звіт оформити як один pdf-файл. Звіт - це покроковий протокол виконання завдання + висновки. Теоретичні відомості вставляти непотрібно. Звіт захищається.

Література

1. Штовба С.Д., Козачко О.М. Machine Learning: стартовий курс. Ел. навчальний посібник. – Вінниця: ВНТУ, 2020. – 81 с.
2. Штовба С.Д., Мазуренко В.В. Інтелектуальні технології ідентифікації залежностей. Лабораторний практикум: ел. навч. пос. – Вінниця: ВНТУ, 2014. – 113 с.

Таблиця 1 – Варіанти завдання

№	Задача	Вхідні змінні
1	Abolone	Diameter Shell weight
2	Abolone	Diameter Viscera weight
3	Abolone	Height Whole weight
4	Abolone	Height Shucked weight
5	Abolone	Height Whole weight
6	Abolone	Length Diameter
7	Abolone	Length Shucked weight
8	Abolone	Length Viscera weight
9	Abolone	Shell weight Shucked weight
10	Abolone	Shell weight Shucked weight
11	Abolone	Viscera weight Shell weight
12	Abolone	Whole weight Length
13	Airfoil Self-Noise	Frequency Suction side displacement thickness
14	Airfoil Self-Noise	Frequency Angle of attack
15	Average Localization Error (ALE) in sensor node localization process in WSNs	anchor_ratio trans_range
16	Average Localization Error (ALE) in sensor node localization process in WSNs	anchor_ratio tnode_density
17	Computer Hardware	MYCT MMAX

№	Задача	Вхідні змінні
18	Estimation of Obesity Levels Based On Eating Habits and Physical Condition	Age NCP
19	Estimation of Obesity Levels Based On Eating Habits and Physical Condition	Weight Height
20	Facebook Metrics	Lifetime Post Total Reach Lifetime Post Total Impressions
21	Housing	B LSTAT
22	Housing	NOX LSTAT
23	Housing	PTRATIO LSTAT
24	Housing	RM PTRATIO
25	Housing	B NOX
26	Housing	NOX PTRATIO
27	Student performance	Freetime health
28	LT-FS-ID: Intrusion detection in WSNs	Sensing Range Transmission Range
29	Physicochemical Properties of Protein Tertiary Structure	Fractional area of exposed non polar residue Secondary structure penalty.
30	QSAR aquatic toxicity	MLOGP GATS1
31	QSAR fish toxicity	SM1_Dz(Z) MLOGP
32	Residential Building, Actual construction costs (output - V10)	V11 V13
33	Residential Building, Actual construction costs (output - V10)	V12 V15
34	Residential Building, Actual construction costs (output - V10)	V21 V22
35	Residential Building, Actual construction costs (output - V10)	V25 V29
36	Residential Building, Actual construction costs (output - V10)	V4 V5
37	Residential Building, Actual construction costs (output - V10)	V6 V8
38	Residential Building, Actual sales prices (output – V9)	V13 V21
39	Residential Building, Actual sales prices (output – V9)	V22 V25

№	Задача	Вхідні змінні
40	Residential Building, Actual sales prices (output – V9)	V26 V29
41	Residential Building, Actual sales prices (output – V9)	V5 V8
42	Residential Building, Actual sales prices (output – V9)	V5 V13
43	Residential Building, Actual sales prices (output – V9)	V8 V22
44	Seoul Bike Sharing (Seasons = Autumn)	Hour Humidity
45	Seoul Bike Sharing (Seasons = Spring)	Hour Windspeed
46	Seoul Bike Sharing (Seasons = Summer)	Hour Temperature
47	Seoul Bike Sharing (Seasons =Winter)	Hour Snowfall
48	Skil craft	APM NumberOfPACs
49	Wine Quality (Red)	Free sulfur dioxide Alcohol
50	Wine Quality (Red)	pH Alcohol
51	Wine Quality (White)	Density Alcohol
52	Wine Quality (White)	Sulphates Alcohol
53	Infrared Thermography Temperature (Gender=Male), target= aveOralF	T_atm Distance
54	Infrared Thermography Temperature (Gender=Female), target= aveOralF	T_offset1 Max1R13_1
55	Infrared Thermography Temperature (Gender=Male), target= aveOralM	aveAllL13_1 T_RC1
56	Infrared Thermography Temperature (Gender=Female), target= aveOralM	aveAllR13_1 RCC1
57	Average Localization Error (ALE) in sensor node localization process in WSNs	anchor_ratio trans_range
58	Average Localization Error (ALE) in sensor node localization process in WSNs	anchor_ratio node_density

№	Задача	Вхідні змінні
59	Average Localization Error (ALE) in sensor node localization process in WSNs	trans_range node_density
60	Average Localization Error (ALE) in sensor node localization process in WSNs	node_density iterations
61	LT-FS-ID: Intrusion detection in WSNs	Sensing Range Area
62	LT-FS-ID: Intrusion detection in WSNs	Area Transmission Range
63	Auction Verification	process.b1.capacity property.price
64	Auction Verification	process.b2.capacity property.product
65	Auction Verification	process.b3.capacity property.winner
66	Auction Verification	process.b1.capacity process.b4.capacity
67	Seoul Bike Sharing (Seasons = Winter)	Hour Temperature
68	Seoul Bike Sharing (Seasons = Autumn)	Hour Windspeed
69	Seoul Bike Sharing (Seasons = Summer)	Hour Humidity
70	Seoul Bike Sharing (Seasons = Spring)	Hour Temperature