Федеральное государственное образовательное бюджетное учреждение высшего образования

«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

(Финансовый университет)

Факультет информационных технологий и анализа больших данных

Кафедра «Прикладная математика и информатика»

Домашнее задание № 4

«Регрессия»

Студенты группы ПМ19-2:

Коротенко В.Р

Пономаренко А.П

Васильева А.Н

Морозов М.

Жигулина Ю.А

Брашич И

Аракелян Р.

Руководитель:

Аксенов Дмитрий Андреевич

Москва 2022

Оглавление.

1. Постановка задачи (физическая модель)

Спрогнозировать какой из критериев будет больше влиять на стоимость автомобиля (год выпуска автомобиля, пробег(в км), тип кузова (седан, хетчбэк, универсал и т.д.), коробка передач (автомат, механика), объем двигателя(в л), тип двигателя (бензиновый, дизельный, гибридный), привод (передний, задний, полный), руль (левый, правый), цвет, состояние (не битый, битый)).

2. Математическая модель

В разделе описывается формульные зависимости в общем виде необходимые для решения класса подобных задач.

3. Алгоритмы

3.1. Алгоритм 1

Линейная регрессия

3.1.1. Описание входных данных

3.1.1.1. Массив предсказываемых данных $y = (y_1, \dots, y_n)$

3.1.1.2. Массив предикатов X размерностью $n \times m$

3.1.1.3. reg – параметр, отвечающий за вид регуляризации (по умолчанию $None$ (без регуляризации), может принимать значения $L_1, L_2, norm$)

3.1.1.3.1. Если $reg = L_1$ или L_2, то вводим коэффициент регуляризации λ ($\lambda \geq 0$; чем больше, тем сильнее регуляризация)

3.1.1.3.2. Если $reg = norm$, то вводим предполагаемое стандартное отклонение остатков σ ($\sigma \geq 0$; чем больше, тем слабее регуляризация)

3.1.2. Описание алгоритма решения

3.1.2.1. Добавить к матрице X колонку единиц слева

3.1.2.2. Убедиться, что $\operatorname{rank}(X) = m + 1$ (если нет, то введены некорректные данные – имеются линейно зависимые предикаты – ошибка, выход из алгоритма)

3.1.2.3. Составляем функцию потерь $f(\omega_0, \dots, \omega_m)$

3.1.2.3.1. Если $reg = None$, то $f(\omega_0, \dots, \omega_m) = \frac{1}{2n} * \sum_{i=1}^{n}(y_i - \vec{\omega}^T \vec{x}_i)^2$, где $y_i -$

реальное значение предсказываемой переменной на $i-$ ом наблюдении; $\vec{\omega} = (\omega_0, \dots, \omega_m) -$ вектор переменных функции f (вектор параметров модели); $i-$ ая строка матрицы X (значения предикатов на $i-$ ом наблюдении)

3.1.2.3.2. Если $reg = L_1$, то $f(\omega_0, \dots, \omega_m) = \frac{1}{2n} * \sum_{i=1}^{n}(y_i - \vec{\omega}^T \vec{x}_i)^2 + \lambda * \sum_{i=1}^{m}|\omega_i|$

3.1.2.3.3. Если $reg = L_2$, то $f(\omega_0, \dots, \omega_m) = \frac{1}{2n} * \sum_{i=1}^{n}(y_i - \vec{\omega}^T \vec{x}_i)^2 + \lambda * \sum_{i=1}^{m} \omega_i^2$

3.1.2.3.4. Если $reg = norm$, то $f(\omega_0, \dots, \omega_m) = \frac{1}{2n} * \sum_{i=1}^{n}(y_i - \vec{\omega}^T \vec{x}_i)^2 + \frac{1}{2\sigma} * \sum_{i=1}^{m} \omega_i^2$

3.1.2.4. С помощью метода сопряжённых градиентов находим минимум $\vec{\omega}^*$ функции $f(\omega_0, \dots, \omega_m)$ ($\omega^0 = (0, \dots, 0)$ – начальная точка)

3.1.2.5. Вычисляем вектор модельных предсказанных данных $\hat{y} = X\vec{\omega}^*$

3.1.3. Описание выходных данных

.На выходе получаем вектор весов.

3.2. Алгоритм 2

Полиномиальная регрессия.

3.2.1. Описание входных данных

3.2.1.1. Массив предсказываемых данных $y = (y_1, \dots, y_n)$

3.2.1.2. Массив предикатов X размерностью $n \times m$

3.2.1.3. $deg-$ степень полинома (чем больше, тем лучше аппроксимация, но при высокой степени будет переобучение)

3.2.1.4. reg – параметр, отвечающий за вид регуляризации (по умолчанию $None$ (без регуляризации), может принимать значения $L_1, L_2, norm$)

3.2.1.4.1. Если $reg = L_1$ или L_2, то вводим коэффициент регуляризации λ ($\lambda \geq 0$; чем больше, тем сильнее регуляризация)

3.2.1.4.2. Если $reg = norm$, то вводим предполагаемое стандартное отклонение остатков σ ($\sigma \geq 0$; чем больше, тем слабее регуляризация)

3.2.2. Описание алгоритма решения

3.2.2.1. Обновить матрицу X с помощью Python:

```
from sklearn.preprocessing import PolynomialFeatures
X = PolynomialFeatures(deg).fit_transform(X)
```

3.2.2.2. Составляем функцию потерь $f(\omega_0, \dots, \omega_m)$, где $m =$ len(X[0])-1

3.2.2.2.1. Если $reg = None$, то $f(\omega_0, \dots, \omega_m) = \frac{1}{2n} * \sum_{i=1}^{n}(y_i - \vec{\omega}^T \vec{x}_i)^2$, где $y_i -$ реальное значение предсказываемой переменной на $i -$ ом наблюдении; $\vec{\omega} = (\omega_0, \dots, \omega_m) -$ вектор переменных функции f (вектор параметров модели); $i -$ ая строка матрицы X (значения предикатов на $i -$ ом наблюдении)

3.2.2.2.2. Если $reg = L_1$, то $f(\omega_0, \dots, \omega_m) = \frac{1}{2n} * \sum_{i=1}^{n}(y_i - \vec{\omega}^T \vec{x}_i)^2 + \lambda * \sum_{i=1}^{m}|\omega_i|$

3.2.2.2.3. Если $reg = L_2$, то $f(\omega_0, \dots, \omega_m) = \frac{1}{2n} * \sum_{i=1}^{n}(y_i - \vec{\omega}^T \vec{x}_i)^2 + \lambda * \sum_{i=1}^{m} \omega_i^2$

3.2.2.2.4. Если $reg = norm$, то $f(\omega_0, \dots, \omega_m) = \frac{1}{2n} * \sum_{i=1}^{n}(y_i - \vec{\omega}^T \vec{x}_i)^2 + \frac{1}{2\sigma} * \sum_{i=1}^{m} \omega_i^2$

3.2.2.3. С помощью метода сопряжённых градиентов находим минимум $\vec{\omega}^*$ функции $f(\omega_0, \dots, \omega_m)$ ($\omega^0 = (0, \dots, 0)$ – начальная точка)

3.2.2.4. Вычисляем вектор модельных предсказанных данных $\hat{y} = X\vec{\omega}^*$

3.2.3. Описание выходных данных

На выходе получаем вектор весов.

3.3. Алгоритм 3

Экспоненциальная регрессия.

3.3.1. Описание входных данных

3.3.1.1. Массив предсказываемых данных $y = (y_1, \dots, y_n)$

3.3.1.2. Массив предикатов X размерностью $n \times m$

3.3.1.3. reg – параметр, отвечающий за вид регуляризации (по умолчанию $None$ (без регуляризации), может принимать значения $L_1, L_2, norm$)

3.3.1.3.1. Если $reg = L_1$ или L_2, то вводим коэффициент регуляризации λ ($\lambda \geq 0$; чем больше, тем сильнее регуляризация)

3.3.1.3.2. Если $reg = norm$, то вводим предполагаемое стандартное отклонение остатков σ ($\sigma \geq 0$; чем больше, тем слабее регуляризация)

3.3.2. Описание алгоритма решения

3.3.2.1. Вычислим массив y_l: $y_l = \ln(y)$

3.3.2.2. Составляем функцию потерь $f(\omega_0, \dots, \omega_m)$

3.3.2.2.1. Если $reg = None$, то $f(\omega_0, \dots, \omega_m) = \frac{1}{2n} * \sum_{i=1}^{n}(y_{li} - \vec{\omega}^T \vec{x}_i)^2$, где y_{li} – логарифм реального значения предсказываемой переменной ом наблюдении; $\vec{\omega} = (\omega_0, \dots, \omega_m)$ – вектор переменных функции f (вектор логарифмов парамет i – ая строка матрицы X (значения предикатов на i – ом наблюдении)

3.3.2.2.2. Если $reg = L_1$, то $f(\omega_0, \dots, \omega_m) = \frac{1}{2n} * \sum_{i=1}^{n}(y_{li} - \vec{\omega}^T \vec{x}_i)^2 + \lambda * \sum_{i=1}^{m} |\omega_i|$

3.3.2.2.3. Если $reg = L_2$, то $f(\omega_0, \dots, \omega_m) = \frac{1}{2n} * \sum_{i=1}^{n}(y_{li} - \vec{\omega}^T \vec{x}_i)^2 + \lambda * \sum_{i=1}^{m} \omega_i^2$

3.3.2.2.4. Если $reg = norm$, то $f(\omega_0, \dots, \omega_m) = \frac{1}{2n} * \sum_{i=1}^{n}(y_{li} - \vec{\omega}^T \vec{x}_i)^2 + \frac{1}{2\sigma} * \sum_{i=1}^{m} \omega_i^2$

3.3.2.3. С помощью метода сопряжённых градиентов находим минимум $\vec{\omega}_l^*$ функции $f(\omega_0, \dots, \omega_m)$ ($\omega^0 = (0, \dots, 0)$ – начальная точка)

3.3.2.4. Вычисляем вектор параметров $\vec{\omega}^*$: $\vec{\omega}_i^* = e^{\vec{\omega}_{li}^*}$

3.3.2.5. Вычисляем вектор модельных предсказанных данных $\hat{y} = \omega_0 * \omega_1^{x_1} * \omega_2^{x_2} * \dots * \omega_m^{x_m}$

3.3.3. Описание выходных данных

На выходе получаем вектор весов.

4. Варианты использования системы

В разделе перечисляются названия предусматриваемых вариантов использования системы пользователем.

4.1. ВИ 1

В разделе указывается полный алгоритм взаимодействия пользователя с разрабатываемым ПО. Возможно добавление скриншотов интерфейса с указанием функционала кнопок.

4.2. ВИ 2

В разделе указывается полный алгоритм взаимодействия пользователя с разрабатываемым ПО. Возможно добавление скриншотов интерфейса с указанием функционала кнопок.

5. Архитектура решения

В разделе описываются создаваемые для решения задачи методы (функции), разделенные по 3-м принципиальным блокам.

5.1. Функции считывания информации

В разделе описываются все методы(функции) отвечающие за получение программой информации, будь то считывание из файла, ввод пользователем посредствам программного интерфейса или получение данных из сторонних источников.

Для каждого метода (функции) необходимо указать следующую информацию:

- Название метода (функции)
- Входные параметры (сначала обязательные, затем необязательные или параметры по умолчанию)
- Выходные параметры
- Затрагиваемые в ходе работы переменные

5.2. Функции обработки информации

В разделе описываются все методы(функции) за обработку введенной пользователем информации и получение решения.

Для каждого метода (функции) необходимо указать следующую информацию:

- Название метода (функции)
- Входные параметры (сначала обязательные, затем необязательные или параметры по умолчанию)
- Выходные параметры
- Затрагиваемые в ходе работы переменные

5.3. Функции вывода информации

В разделе описываются все методы(функции) отвечающие за вывод информации пользователю, будь то построение графика или распечатка результатов работы функций из п.5.2..

Для каждого метода (функции) необходимо указать следующую информацию:

- Название метода (функции)
- Входные параметры (сначала обязательные, затем необязательные или параметры по умолчанию)

- Выходные параметры
- Затрагиваемые в ходе работы переменные

6. Тестирование

В разделе приводится тестирование работы программы. Оптимальный способ представления результатов тестирования – это следующая таблица:

Таблица 1. Результаты тестирования программы

Входные данные:

Массив предсказанных данных y = (1, 2, 3, 2, 1)

Массив предикатов X размерностью $n \times m$: $\begin{pmatrix} 0 \\ -1 \\ 1 \\ 2 \\ 5 \end{pmatrix}$

Параметр отвечающий за вид регуляции – None(без регуляции)

Для полиномиальной регрессии степень полинома 1

Параметр	**Линейная регрессия**	**Полиномиальная регрессия**	**Экспоненциальная регрессия**
Полученное решение	Вектор модельных предсказанных данных: [1.97167800099167, 2.09431330914668, 1.84904269283666, 1.72640738468165, 1.35850146021662] Массив коэффициентов регрессии: [-0.122635308155011] Свободный член: 1.97167800099167 Функция в аналитическом виде: y^=1.97167800099167 + -0.122635308155011 * x1	Вектор модельных предсказанных данных: [1.97167800099167, 2.09431330914668, 1.84904269283666, 1.72640738468165, 1.35850146021662] Массив коэффициентов регрессии: [-0.122635308155011] Свободный член: 1.97167800099167 Функция в аналитическом виде: y^=1.97167800099167 + -0.122635308155011 * x1	Вектор модельных предсказанных данных: [7.18271899144544, 8.11986321953359, 6.35373413507259, 5.62042556687334, 3.89035907363203] Массив коэффициентов регрессии: [0.884586205118123] Свободный член: 7.18271899144544 Функция в аналитическом виде: y^=7.18271899144544 * 0.884586205118123**x1

Время исполнения (в секундах)	3.6043612957000732	3.7390010356903076	3.3779664039611816

Таблица 2. Результаты тестирования программы

Входные данные:

Массив предсказанных данных y = (1, 2, 3, 2, 1)

Массив предикатов X размерностью $n \times m$: $\begin{pmatrix} 0\,1 \\ -1\ -1 \\ 1\,2 \\ 2\,3 \\ 5\,4 \end{pmatrix}$

Параметр отвечающий за вид регуляции – L1. Лямбда = 0.95

Для полиномиальной регрессии степень полинома 1

Параметр	**Линейная регрессия**	**Полиномиальная регрессия**	**Экспоненциальная регрессия**
Полученное решение	Вектор модельных предсказанных данных: [1.06705899373755, 0.476305266704051, 1.28228026851944, 1.49750154330134, 1.39210046314384] Массив коэффициентов регрессии: [-0.160311177469701, 0.375532452251598] Свободный член: 0.691526541485948 Функция в аналитическом виде: y^=0.691526541485948 + -0.160311177469701 * x1 + 0.375532452251598 * x2	Вектор модельных предсказанных данных: [1.06705899373755, 0.476305266704051, 1.28228026851944, 1.49750154330134, 1.39210046314384] Массив коэффициентов регрессии: [-0.160311177469701, 0.375532452251598] Свободный член: 0.691526541485948 Функция в аналитическом виде: y^=0.691526541485948 + -0.160311177469701 * x1 + 0.375532452251598 * x2	Вектор модельных предсказанных данных: [2.90681794657233, 1.61011445527750, 3.60485038761067, 4.47050574060897, 4.02329196041247] Массив коэффициентов регрессии: [0.851878662271006, 1.45576633435854] Свободный член: 1.99676134690473 Функция в аналитическом виде: y^=1.99676134690473 * 0.851878662271006**x1 * 1.45576633435854**x2

Время исполнения (в секундах)	1.199791669845581	1.2217321395874023	1.1918120384216309

Таблица 3. Результаты тестирования программы

Входные данные:

Массив предсказанных данных y = (1, 2, 3, 2, 1)

Массив предикатов X размерностью $n \times m$: $\begin{pmatrix} 0\ 1 \\ -1\ -1 \\ 1\ 2 \\ 2\ 3 \\ 5\ 4 \end{pmatrix}$

Параметр отвечающий за вид регуляции – L2. Лямбда = 0.95

Для полиномиальной регрессии степень полинома 1

Параметр	**Линейная регрессия**	**Полиномиальная регрессия**	**Экспоненциальная регрессия**
Полученное решение	Вектор модельных предсказанных данных: [1.56046057072504, -0.0637871558360776, 1.90283759723199, 2.24521462373894, 0.708604303151448] Массив коэффициентов регрессии: [-0.939493673547219, 1.28187070005417] Свободный член: 0.278589870670871 Функция в аналитическом виде: y^=0.278589870670871 + -0.939493673547219 * x1 + 1.28187070005417 * x2	Вектор модельных предсказанных данных: [1.56046057072504, -0.0637871558360776, 1.90283759723199, 2.24521462373894, 0.708604303151448] Массив коэффициентов регрессии: [-0.939493673547219 1.28187070005417] Свободный член: 0.278589870670871 Функция в аналитическом виде: y^=0.278589870670871 + -0.939493673547219 * x1 + 1.28187070005417 * x2	Вектор модельных предсказанных данных: [4.76101352369997, 0.938204669669168, 6.70489326059104, 9.44244191118834, 2.03115440361240] Массив коэффициентов регрессии: [0.390825670645149, 3.60337425687880] Свободный член: 1.32126534306317 Функция в аналитическом виде: y^=1.32126534306317 * 0.390825670645149**x1 * 3.60337425687880**x2

Время исполнения (в секундах)	3.6043612957000732	3.3330869674682617	3.7001049518585205

Таблица 4. Результаты тестирования программы

Входные данные:

Массив предсказанных данных y = (1, 2, 3, 2, 1)

Массив предикатов X размерностью $n \times m$: $\begin{pmatrix} 0\ 1 \\ -1\ -1 \\ 1\ 2 \\ 2\ 3 \\ 5\ 4 \end{pmatrix}$

Параметр отвечающий за вид регуляции – norm. δ = 0.1

Для полиномиальной регрессии степень полинома 1

Параметр	**Линейная регрессия**	**Полиномиальная регрессия**	**Экспоненциальная регрессия**
Полученное решение	Вектор модельных предсказанных данных: [1.48069444353561, -0.339329399871366, 1.88107952216139, 2.28146460078716, 0.643342307102093] Массив коэффициентов регрессии: [-1.01925368615542, 1.41963876478120] Свободный член: 0.0610556787544107 Функция в аналитическом виде: y^=0.0610556787544107 + -1.01925368615542 * x1 + 1.41963876478120 * x2	Вектор модельных предсказанных данных: [1.48069444353561, -0.339329399871366, 1.88107952216139, 2.28146460078716, 0.643342307102093] Массив коэффициентов регрессии: [-1.01925368615542, 1.41963876478120] Свободный член: 0.0610556787544107 Функция в аналитическом виде: y^=0.0610556787544107 + -1.01925368615542 * x1 + 1.41963876478120 * x2	Вектор модельных предсказанных данных: [4.39599739314743, 0.712247796111563, 6.56058333592186, 9.79100983423445, 1.90283010543803] Массив коэффициентов регрессии: [0.360864157617071, 4.13562623653971] Свободный член: 1.06295809672239 Функция в аналитическом виде: y^=1.06295809672239 * 0.360864157617071**x1 * 4.13562623653971**x2

Время исполнения (в секундах)	5.206078290939331	5.011598348617554	4.981677293777466

7. Заключение

В заключении необходимо указать решает ли предпочтительный алгоритм поставленную в п.1. задачу, привести решение для задачи из п.1., а так же указать точность полученного решения.

Так же требуется произвести сравнение выбранных алгоритмов по разным критериям. Оптимальный вид сравнения приведен в следующей таблице:

Таблица 3. Сравнение алгоритмов.

Критерий	Алгоритм 1	Алгоритм 2	Алгоритм 3
Пример: Количество возможных параметров	Не ограничено	Не ограничено	Не более 2
Критерий 1	Значение	значение	значение
Критерий 2	Значение	значение	значение
Критерий 3	Значение	значение	значение
Критерий 4	Значение	значение	значение
Критерий 5	Значение	значение	значение

По результатам таблицы 3 необходимо выбрать оптимальный для заказчика, по Вашему мнению, алгоритм.

Для выбранного алгоритма следует описать возможные перспективы развития данного алгоритма.