Тема 10. Простейшие модели линейной регрессии

Практические задания для самостоятельного выполнения

Задания выполняются по вариантам. Формулировки заданий общие для всех вариантов; конкретный набор данных, подлежащий изучению, выбирается в соответствии с номером своего варианта.

Результаты выполнения заданий необходимо представить в виде двух файлов:

1) ноутбук в формате *ipynb*, содержащий программный код, результаты его выполнения, а также все необходимые пояснения, выводы и комментарии (в текстовых ячейках);
2) файл в формате *pdf* (или *html*), полученный путем экспорта (или вывода на печать) ноутбука из п. 1).

Внимание: в названии файлов должна обязательно присутствовать фамилия автора. Безымянные работы проверяться не будут.

Общее оформление работы: программный код для каждого пункта задания должен быть написан в отдельной ячейке. Выше или ниже (как удобно в конкретном случае) в текстовой ячейке должны быть написаны пояснения, рассуждения, обоснования и выводы, полученные по результатам работы этого кода. Для каждой кодовой ячейки – своя текстовая ячейка с рассуждениями и комментариями в соответствии с заданием.

Обратите внимание, что все необходимые для выполнения задания программные конструкции рассмотрены в учебном ноутбуке, размещенном в системе LMS. После изучения этих материалов выполнение задания не потребует больших усилий.

Максимальная оценка за выполнение задания вне аудитории – 1 балл. Дополнительные баллы (от 0 до 3) можно будет получить на следующем практическом занятии по результатам тестирования.

Внимание: самостоятельное и вдумчивое выполнение задания серьезно повышает вероятность успешного прохождения теста.

Задание 1.

Используя инструментарий библиотеки **sklearn**, реализовать вычислительные эксперименты с построением моделей линейной регрессии, проанализировать их результаты и сделать выводы. В процессе работы выполнить следующие действия.

1. Сгенерировать модельный набор данных для задачи линейной регрессии с одной целевой переменной и двумя признаками, из которых информативным является только один. Количество объектов в выборке

положить равным 110. Обеспечить воспроизводимость результатов, установив номер генератора случайных чисел равным номеру своего варианта. Параметр, определяющий степень рассеянности данных, задать в соответствии с номером своего варианта (см. таблицу 1).

2. Вывести сгенерированные веса признаков; определить, какой именно из двух признаков является информативным (значимым), а какой нет. Записать в текстовой ячейке полученное уравнение регрессии и пояснение о значимости признаков.
3. Вывести на одном графике сгенерированное облако точек в координатах «информативный признак – целевая переменная» и прямую регрессии со сгенерированными коэффициентами. На другом графике вывести облако точек в координатах «неинформативный признак – целевая переменная». Записать комментарии к виду полученных графиков.
4. Поэкспериментировать с величиной шума, увеличивая и уменьшая значение, заданное в таблице 1. Вывести три графика (облака вместе с прямой регрессии) в одном ряду, с заголовками, сообщающими об используемом значении параметра шума. Прокомментировать полученные результаты.
5. Выполнить разовое разбиение трех наборов данных с различными значениями шума на обучающую и тестовую выборки в соотношении 70/30.
6. Для каждого из трех наборов создать модель линейной регрессии LinearRegression и обучить ее на обучающей выборке.
7. Для каждой из обученных моделей вывести коэффициенты уравнения регрессии, полученные в результате обучения. Записать в текстовых ячейках соответствующие уравнения регрессии. Сопоставить эти уравнения с уравнением зависимости, сгенерированной в п. 1. Записать свои комментарии.
8. Выполнить визуализацию всех полученных результатов. Записать комментарии (визуальную оценку обобщающей способности полученных моделей).
9. Получить предсказания обученных моделей для объектов обучающих и тестовых выборок. Вывести массивы ответов на тестовых выборках и массивы предсказаний моделей для тестовых данных. Сопоставить эти значения.
10. Вывести отдельно предсказание модели для объекта тестовой выборки с номером, равным номеру своего варианта. Вывести также значение ошибки на этом объекте.
11. Вычислить среднеквадратичную ошибку каждой модели на обучающей и тестовой выборке. Дать оценку полученным результатам (записать в текстовой ячейке).

12. Дополнительное задание (+0,25 балла):
    - для одного из модельных наборов данных, сформированных при выполнении основного задания, найти оптимальные коэффициенты уравнения регрессии аналитическим методом; прокомментировать все выполняемые действия;
    - сопоставить уравнения регрессии: полученное аналитическим методом и с помощью LinearRegression; комментарии записать в текстовой ячейке.

Таблица 1. Значения параметра ***noise***.

Вариант	**Значение**	**Вариант**	**Значение**	**Вариант**	**Значение**
1	16	**11**	11	**21**	4
2	11	**12**	1	**22**	18
3	15	**13**	8	**23**	5
4	12	**14**	12	**24**	17
5	7	**15**	1	**25**	1
6	16	**16**	17	**26**	1
7	12	**17**	15	**27**	12
8	17	**18**	16	**28**	12
9	6	**19**	15	**29**	14
10	15	**20**	14	**30**	15

Задание 2.

В этом задании используется набор данных о росте и весе 25 тыс. подростков в дюймах и фунтах соответственно. Наборы данных по вариантам, сохраненные в csv-файлах, представляют собой фрагменты этого набора. Имена файлов: ВариантN, где N – номер варианта.
Выполнить первичное изучение имеющихся данных и построить линейную модель прогнозирования роста по весу для рассматриваемой возрастной группы. В процессе работы выполнить следующие действия.

1. Импортировать данные из файла и вывести несколько первых записей (для контроля корректности импорта и получения представления о наборе).
2. Выполнить первичный анализ данных:
    - визуализировать данные в виде облака точек;
    - построить гистограммы распределения признаков;
    - проанализировать характер распределений признаков, наличие/отсутствие выбросов;
    - оценить корреляцию признаков.

Напоминание: распределение признаков, близкое к нормальному, при наличии заметной корреляции позволяет надеяться на успешное построение линейной модели взаимосвязи признаков.

Прокомментировать все выполняемые действия, проанализировать полученные результаты, сделать выводы (все комментарии, рассуждения и выводы записать в текстовых ячейках).

3. Выполнить разбиение набора данных на обучающую и тестовую выборки в соотношении 90/10.
4. Обучить на обучающих данных модель линейной регрессии LinearRegression.
5. Вывести коэффициенты уравнения регрессии, полученные в результате обучения. Записать в текстовой ячейке соответствующее уравнение регрессии.
6. Выполнить визуализацию полученных результатов. Дать визуальную оценку обобщающей способности полученной модели (рассуждения записать в текстовой ячейке).
7. Получить предсказания обученных моделей для объектов обучающей и тестовой выборки.
8. Для нескольких объектов тестовой выборки рассчитать значения ошибки модели на этих объектах; вывести предсказания модели, правильные ответы и значения ошибок.
9. Вычислить среднеквадратичную ошибку модели на обучающей и тестовой выборке. Дать оценку полученным результатам (записать в текстовой ячейке).