

Цель работы

Проверить освоение студентами базовых приемов работы с простыми моделями регрессии, применить их при анализе реальных данных.

Методические указания

Контрольная работа выполняется в аудитории. В начале занятия преподаватель распределяет по студентам датасеты (задания). Распределение должно проходить случайным образом.

После получения задания студенты выполняют работу на лабораторных компьютерах. Работа должна быть выполнена студентом самостоятельно. Работа выполняется в виде ноутбука формата .ipynb

После завершения выполнения работы студент должен сдать работу преподавателю. Работы сдаются в электронном виде с корпоративной почты студента на корпоративную почту преподавателя. Работа должна быть сдана не позже окончания аудиторного занятия.

Контрольная работа выполняется на лабораторный компьютеры.

Во время выполнения работы можно пользоваться всем (документацией, прошлыми работами, Интернетом), кроме помощи других студентов, работа выполняется самостоятельно.

Критерии оценки

1. Код должен выполняться без ошибок. Программные ошибки - сразу 0 баллов.
2. Код должен содержать пояснения в виде ячеек с текстом. В пояснениях нужно писать значимые выводы по ходу выполнения работы.
3. Все графики должны быть правильно подписаны и понятны.
4. Текстовый вывод должен быть наглядным и отформатирован. Численные значения - округлены до разумного количества значащих цифр.

Содержание работы

1. (1 балл) Загрузите предложенный вам датасет с помощью функции ``sklearn.datasets.fetch_openml``. Выведите текстовое описание загруженного датасета. Обозначьте целевую переменную за ``y``, а остальные данные за ``X``.

2. (1 балл) Выведите основную статистическую информацию о данных. Сделайте количественное описание датасета: число строк (объектов), число столбцов (признаков), статистику по признакам.

3. (1 балл) Выведите тип данных каждого признака и целевой переменной. Убедитесь, что в данных все признаки являются числовыми. В случае, если это не так, удалите нечисловые колонки.

4. (1 балл) Убедитесь, что нет пропущенных значений в данных и у целевой переменной. В случае, если это не так, заполните пропуски медианными значениями.

5. (1 балл) Постройте гистограмму распределения целевой переменной. Сделайте вывод по графику. Предположите, какому виду распределения она принадлежит.

6. (4 балл) Обучите модель линейной регрессии на рассматриваемых данных, написанную своими руками. Проиллюстрируйте работу модели графиком обучения и распределения целевой переменной. Выведите уравнение полученной гиперплоскости.

7. (2 балл) Обучите модель `LinearRegression()` линейной регрессии на рассматриваемых данных из библиотеки `sklearn`. Выведите уравнение полученной гиперплоскости. Убедитесь, что уравнения гиперплоскостей, полученных с помощью модели написанной своими руками и библиотечной, будут примерно одинаковыми.

8. (1 балл) Оцените работу моделей, рассчитав для каждой из них метрики: коэффициент детерминации и ошибку MSE. Сделайте выводы по качеству работы моделей.

Варианты заданий

1. LEV
2. ERA
3. debutanizer
4. kin8nm
5. delta_elevators
6. quake
7. puma8NH
8. bank8FM
9. fri_c0_1000_5
10. analcatdata_supreme
11. liver-disorders
12. qsbr_y2
13. ESL
14. vineyard
15. mbagrade