Лабораторная работа №2

**Многомерная линейная регрессия**

Для выполнения данной работы мы будем использовать набор данных (датасет) с ценами на недвижимость в Бостоне. Датасет прилагается к заданию как в виде отдельного файла, так и может быть загружена непосредственно из scikit-learn.

from sklearn.datasets import load\_boston

**Внимание! В последней версии scikit-learn этот датасет отсутствует (удален по каким-то толерастическим причинам) но если у вас не последняя версия то может и присутствовать.**

В этом наборе имеется 506 выборок и 13 переменных характеристик. Цель состоит в том, чтобы спрогнозировать стоимость дома с использованием заданных характеристик.

**Характеристики набора данных:**

**Количество экземпляров**: 506

**Количество атрибутов**: 13 числовых / категориальных прогнозов. Среднее значение (атрибут 14) обычно является целевым.

**Информация об атрибутах (по порядку)**:

* Уровень преступности на душу населения по городам
* ЗН доля земли под жилую застройку зонирована на участки площадью более 25 000 кв. Футов.
* INDUS доля акров, не относящихся к розничной торговле, на город
* CHAS Фиктивная переменная реки Чарльз (= 1, если участок ограничивает реку; 0 в противном случае)
* Концентрация оксидов азота NOX (частей на 10 миллионов)
* RM среднее количество комнат в доме
* ВОЗРАСТ Доля домов, построенных до 1940 года, занимаемых владельцами
* DIS взвесила расстояния до пяти бостонских центров занятости
* Индекс доступности радиальных автомобильных дорог РАД
* НАЛОГ Полная ставка налога на имущество за 10 000 долларов США
* PTRATIO соотношение учеников и учителей по городам
* B 1000 (Bk — 0,63) ^ 2, где Bk — доля чернокожего населения по городам.
* LSTAT% более низкий статус населения
* MEDV Средняя стоимость частных домов в 1000 долларов

**Отсутствующие значения атрибутов**: Нет

**Создатель**: Харрисон Д. и Рубинфельд Д. Л.

Это копия набора данных о жилищном строительстве UCI ML. <https://archive.ics.uci.edu/ml/machine-learning-databases/housing/>

Этот набор данных был взят из библиотеки StatLib, которая поддерживается в Университете Карнеги-Меллона.

Данные о ценах на жилье в Бостоне, представленные Д. Харрисоном и Рубинфельдом Д. Л. «Гедонические цены и спрос на чистый воздух», J. Environ. Economics & Management, vol.5, 81-102, 1978. Используется в Belsley, Kuh & Welsch, «Regression диагностика…», Wiley, 1980.

Цены на дом указаны переменной MEDV это и будет наша **целевая переменная,** а на основании остальных данных мы будем прогнозировать стоимость дома с помощью линейного регрессора.

**Задание к лабораторной работе.**

1. Исследовать вопрос, способствует ли увеличение количества данных в датасете, улучшению точности модели и каким образом. Вопрос рассмотреть в 2 аспектах:

* Увеличение количества данных (строк в датасете)
* Увеличение количества признаков (столбцов в датасете)

Для исследования по каждому из аспектов построить минимум по 3 модели (например, при исследовании зависимости от количества признаков построить модели с 2, 7, 13 признаками). С количеством данных аналогично. (можно раздробить на более мелкие части и построить соответствующий график)

Для оценки моделей использовать не менее трех различных метрик. **Сделать выводы по результатам исследования (прямо в работе)**.

1. Построить (если в 1 не строилась) модель с 2 признаками. Используя Matplotlib

отрисовать в 3Д режиме данные (как точки в пространстве) и полученную плоскость решения.