Лабораторная работа №4

**Исследование кривых обучения и регуляризации**

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 данный датасет может и присутствовать. Также он прилагается к лабораторной работе №2**

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**

**Задание к лабораторной работе.**

1. Построить и отрисовать кривые обучения для любых 2х моделей - регрессорах.

Сделать вывод – какая из моделей более подходит для предложенного датасета,

Проанализировать наличие недо- или пере- обученности, выдвинуть предложение о значении целевого показателя выбранной метрики.

Отрисовать целевое значение на построенном графике кривых обучения.

Для одной из моделей отрисовать кривую обучения относительно ошибки.

1. Последовательно выполнить обучение моделей линейной регрессии, гребневой регуляризации (Ridge) и Lasso регуляризации.

Сравнить модели используя несколько метрик.

Свести значения весовых коэффициентов в единую таблицу, сделать выводы о коллинеарности данных и о важности имеющихся признаков

1. Опциональное задание.

Попытаться реализовать ElasticNet модель. Сравнить с результатами полученными в п.2.