Министерство образования Республики Беларусь

Учреждение образования

БелорусскиЙ государственный университет

информатики и радиоэлектроники

Факультет компьютерных систем и сетей

Кафедра информатики

ОТЧЕТ

по лабораторной работе №10

**Градиентный бустинг**

|  |  |  |
| --- | --- | --- |
| Студент |  | И.Ю. Фалько |
| Преподаватель |  | М. В. Стержанов |

Минск 2019

СОДЕРЖАНИЕ

[Условие лабораторной работы 3](#_Toc25786647)

[Ход выполнения 3](#_Toc25786648)

[Заключение 8](#_Toc25786649)

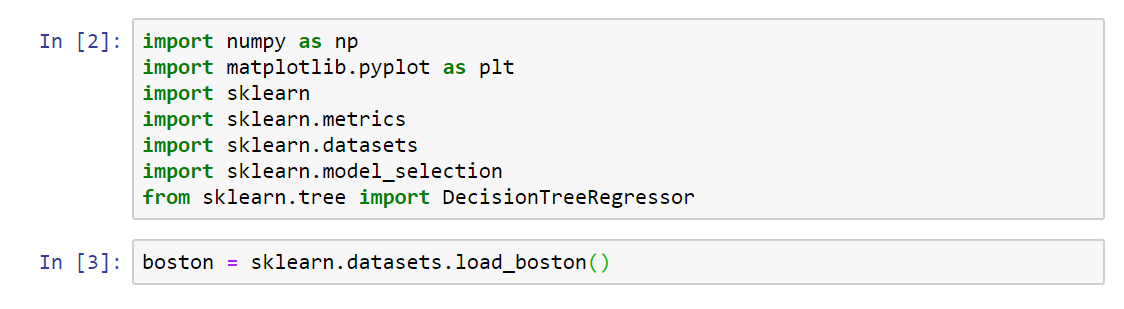
# Условие лабораторной работы

Для выполнения задания используйте набор данных boston из библиотеки sklearn

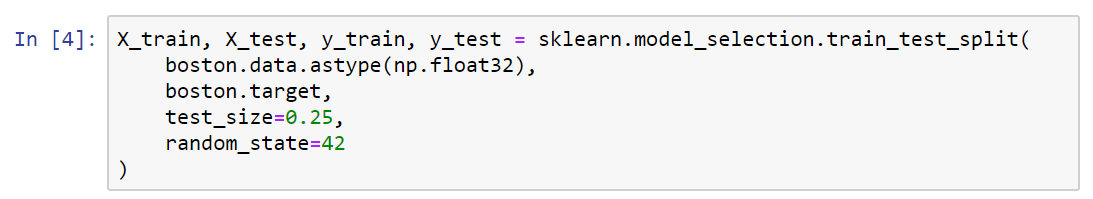
https://scikit-learn.org/stable/datasets/index.html#boston-dataset

# Ход выполнения

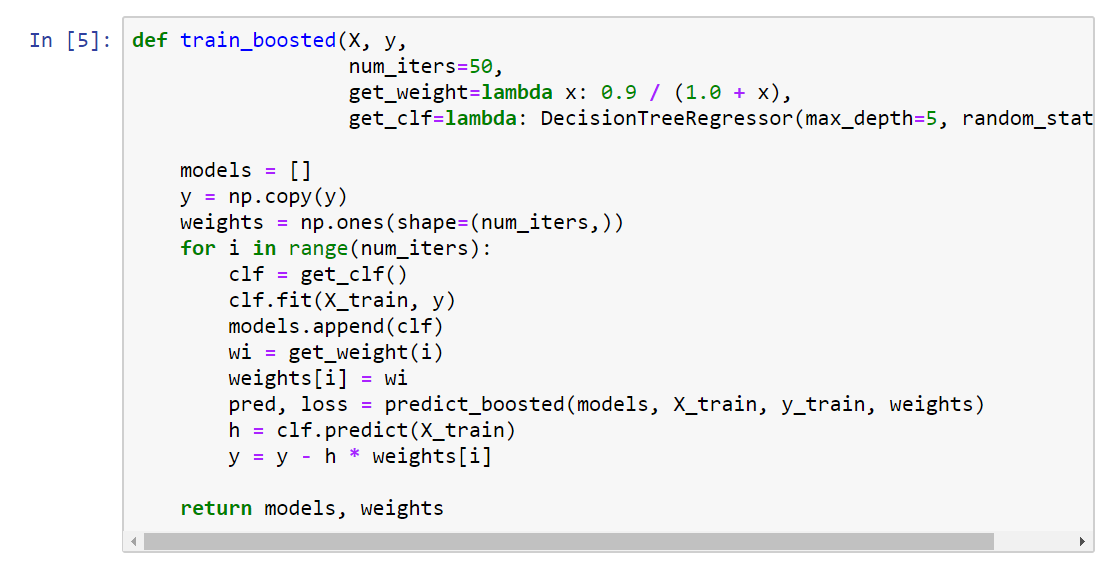
1. Загрузите данные с помощью библиотеки sklearn.



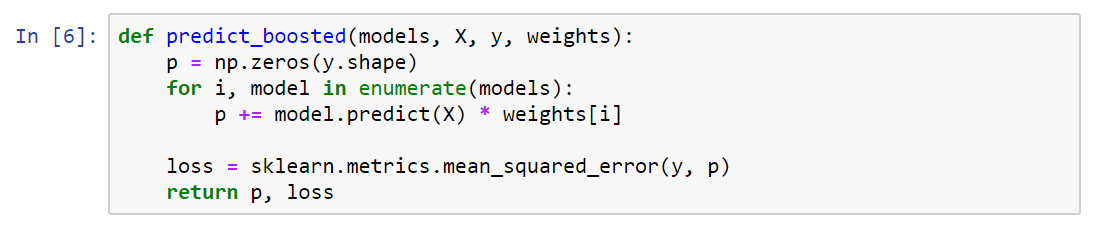
1. Разделите выборку на обучающую (75%) и контрольную (25%).



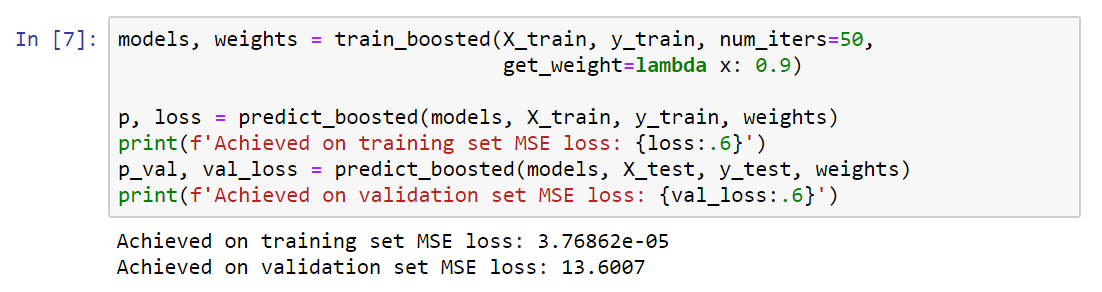
1. Заведите массив для объектов DecisionTreeRegressor (они будут использоваться в качестве базовых алгоритмов) и для вещественных чисел (коэффициенты перед базовыми алгоритмами).
2. В цикле обучите последовательно 50 решающих деревьев с параметрами max\_depth=5 и random\_state=42 (остальные параметры - по умолчанию). Каждое дерево должно обучаться на одном и том же множестве объектов, но ответы, которые учится прогнозировать дерево, будут меняться в соответствие с отклонением истинных значений от предсказанных.
3. Попробуйте всегда брать коэффициент равным 0.9. Обычно оправдано выбирать коэффициент значительно меньшим - порядка 0.05 или 0.1, но на стандартном наборе данных будет всего 50 деревьев, возьмите для начала шаг побольше.



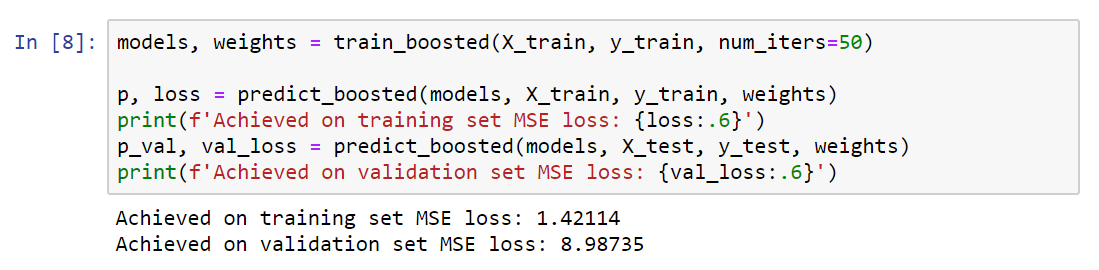
1. В процессе реализации обучения вам потребуется функция, которая будет вычислять прогноз построенной на данный момент композиции деревьев на выборке X. Реализуйте ее. Эта же функция поможет вам получить прогноз на контрольной выборке и оценить качество работы вашего алгоритма с помощью mean\_squared\_error в sklearn.metrics.



Проверим работу алгоритма:

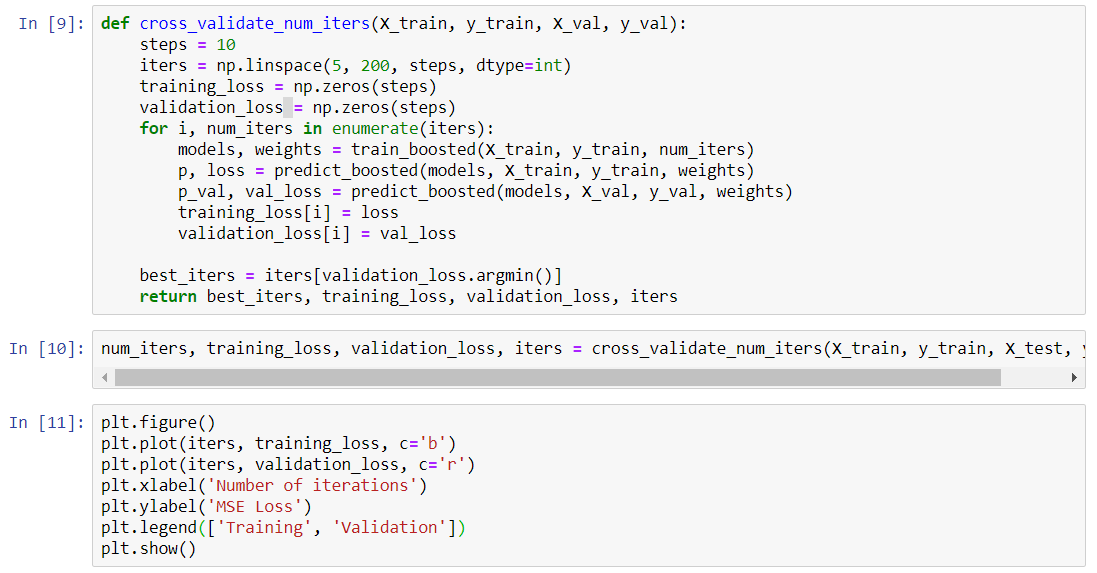


1. Попробуйте уменьшать вес перед каждым алгоритмом с каждой следующей итерацией по формуле 0.9 / (1.0 + i), где i - номер итерации (от 0 до 49). Какое получилось качество на контрольной выборке?

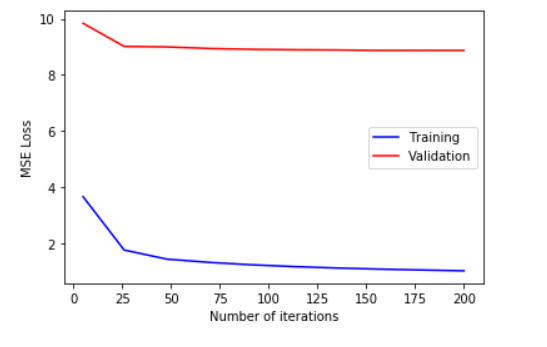


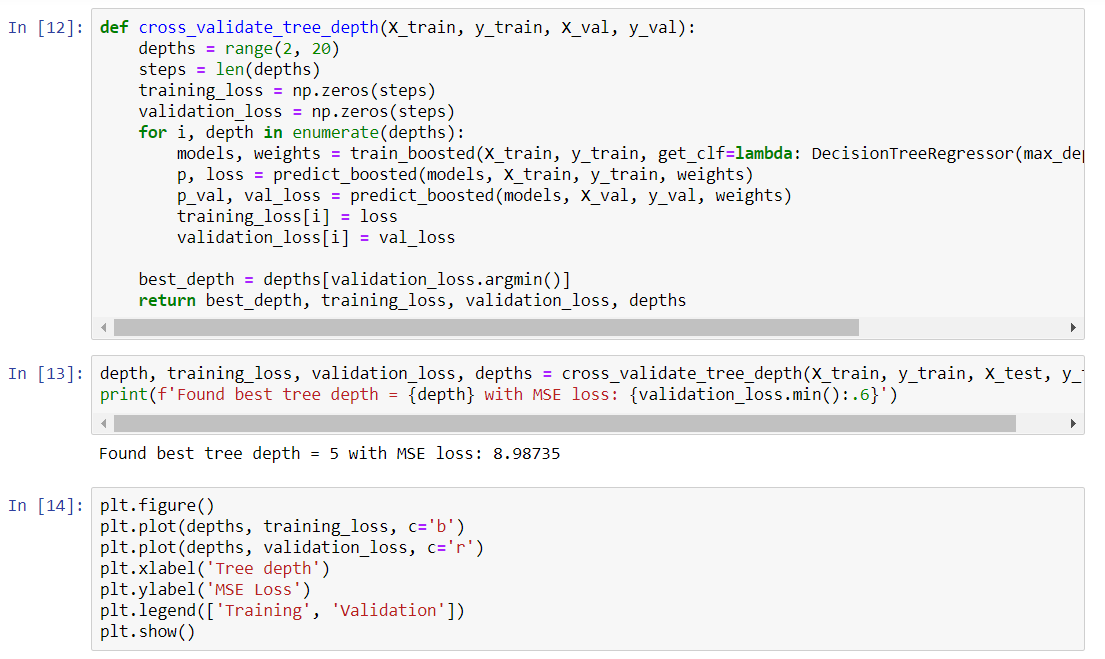
По результатам обучения двух моделей, одна с коэффициентом 0.9, вторая с адаптивным коэффициентом, видно что для коэффицинта 0.9 модель намного лучше обучается на тренировочной выборке, но хуже себя показывает на валидационной выборке.

1. Исследуйте, переобучается ли градиентный бустинг с ростом числа итераций, а также с ростом глубины деревьев. Постройте графики. Какие выводы можно сделать?

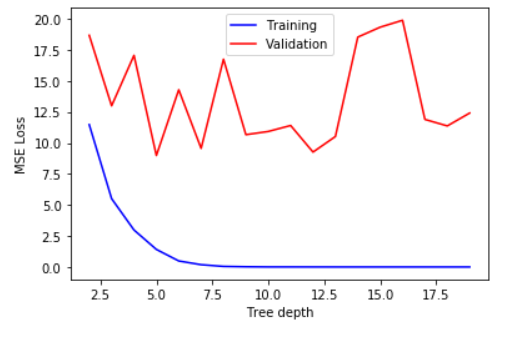


Используя адаптивный шаг для каждого последуещего дерева и кол-во итераций от 5 до 200 по графику видим, что функция потерь убывает для тренировочного набора, так и для валидационного. Т.е. переобчить модель не получилось:

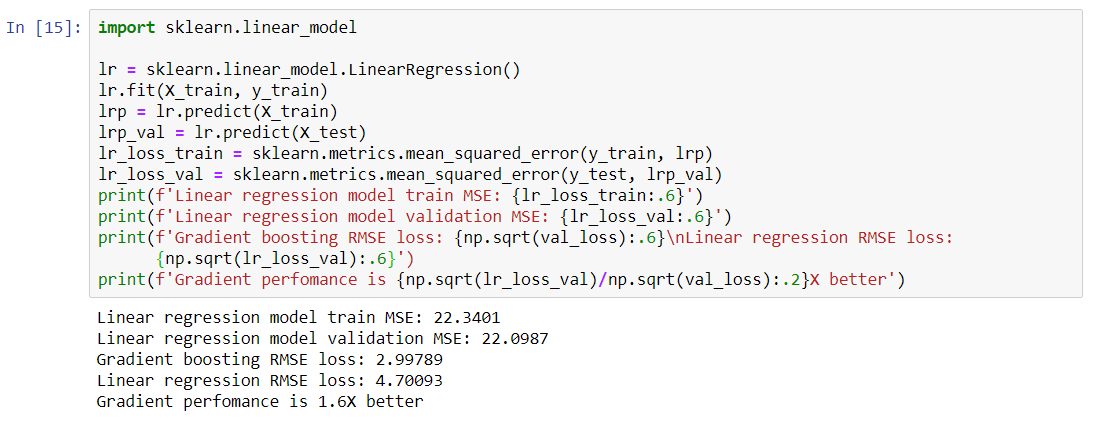




Используя адаптивный шаг для каждого последуещего дерева, кол-во итераций от 50 и глубину деревьев от 2 до 20 по графику видим, что функция потерь для валидационной выборки ведет себя не стабильно в зависимости от грубины дереве, так или иначе в целом она возрастает после 5, а функция потерь на тренировочной выборке всегда убывает, что говорит о переобучении нашей модели:



1. Сравните качество, получаемое с помощью градиентного бустинга с качеством работы линейной регрессии. Для этого обучите LinearRegression из sklearn.linear\_model (с параметрами по умолчанию) на обучающей выборке и оцените для прогнозов полученного алгоритма на тестовой выборке RMSE.



# ЗАКЛЮЧЕНИЕ

В данной работе была построена модель градиентного бустинга используя деревья решений как целевую модель для бустинга, проверили как параметры модели влияют на результаты работы. Сравнили результаты работы модели градиентного бустинга с моделью линейной регрессии.