Министерство образования и науки Российской Федерации

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

ПОЛИТЕХ
Санкт-Петербургский
политехнический университет
Петра Великого

Отчет по Лабораторной работе № 5.
по дисциплине “Машинное обучение”

Выполнила
студентка гр. 3530202/00201 Козлова Е. А.

Руководитель Селин И. А.

Санкт-Петербург
2023

Оглавление

Задание 1

Загрузите данные из файла reglab1.txt. Постройте по набору данных регрессии, используя модели с различными зависимыми переменными. Выберите наиболее подходящую модель.

В качестве регрессора используется класс LinearRegression из библиотеки Sklearn. Для оценки качества из выборки выделяется 30% тестовой части, а в качестве метрики используется коэффициент детерминации (R2_score в реализации Sklearn).

```
Score X(Y, Z): 0.92
Score Y(X, Z): 0.95
Score Z(X, Y): 0.97
Score X(Y): 0.0002
Score Y(Z): 0.61
Score Z(X): 0.37
```

Вывод: лучше всего рассматривать зависимость Z(X, Y)

Задание 2

Реализуйте следующий алгоритм для уменьшения количества признаков, используемых для построения регрессии: для каждого выбрать подмножество признаков мощности , минимизирующее остаточную сумму квадратов RSS. Используя полученный алгоритм, выберите оптимальное подмножество признаков для данных из файла reglab.txt. Объясните свой выбор.

Результаты с различными вариациями:

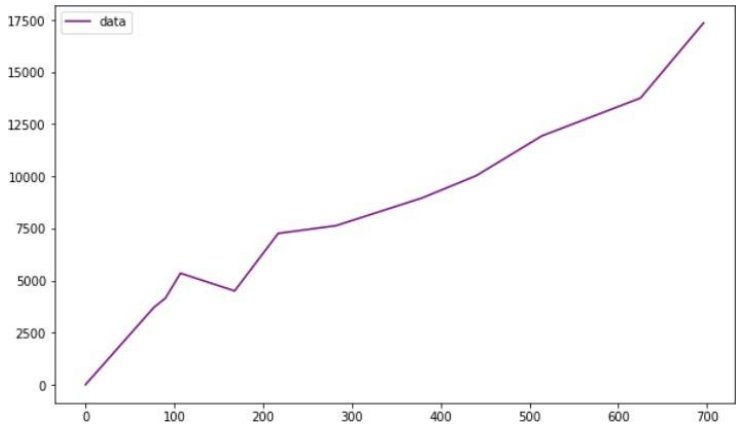
var	RSS
x1, x2, x3	0.0015903048755141662
x1, x2, x4	0.0017417756867110768
x1, x2	0.003552867528399772
x1, x3, x4	0.6380642206862528
x1	0.7035771662135488
x1, x4	0.7166138947728119
x1, x3	0.7424918700173698
x2, x3	1.3839657168716932
x2, x4	1.4353750331404591
x2	1.6008225891761723
x3	1.8688242519700318
x3, x4	1.9624139447171114
x4	2.1538862944736703

Как можно видеть, самым оптимальным является подмножество (x1,x2,x3). Также можно заметить, что при отсутствии x1 и x2 мы значительно теряем точность, в то время как отсутствие x3 и x4 практически не влияет на точность.

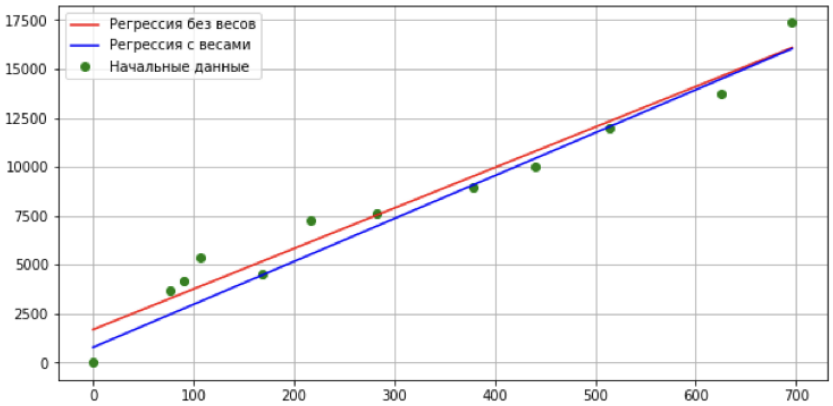
Задание 3

Загрузите данные из файла cygage.txt. Постройте регрессию, выражающую зависимость возраста исследуемых отложений от глубины залегания, используя веса наблюдений. Оцените качество построенной модели.

Исходные данные:



Регрессии с применением весов и без:



Вывод: регрессия с весами оказалась несколько лучше по значению метрики, чем без них (без весов: 0.9592555, с весами: 0.9736839). В общем случае весами стоит пользоваться, когда есть некоторые предположения о достоверности полученных точек относительно друг друга. В этом случае выбор весовых коэффициентов позволяет лучше подобрать решение для конкретной практической задачи.

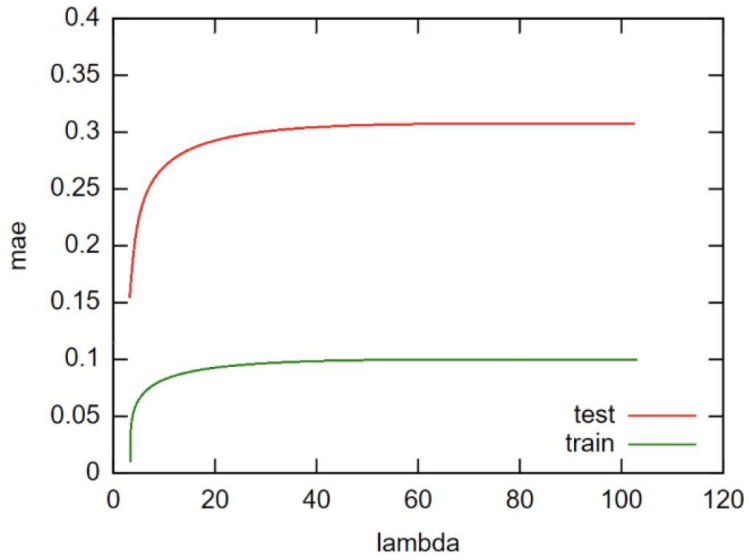
Задание 4

Загрузите данные из файла longley.csv. Данные состоят из 7 экономических переменных, наблюдаемых с 1947 по 1962 годы (n=16). Исключите переменную Population. Разделите данные на тестовую и обучающую выборки равных размеров случайным образом. Постройте линейную регрессию по признаку Employed.
Постройте гребневую регрессию для значений . Подсчитайте ошибку на тестовой и обучающей выборке для линейной регрессии и гребневой регрессии на данных значениях λ, постройте графики. Объясните полученные результаты.

В ходе работы была исключена переменная Population, доля тестовой выборки 0.25. Результаты линейной регрессии по признаку Employed:

```
LinearRegression train: 0.01284337644
LinearRegression test: 0.48727692657
```

Результат гребневой регрессии для заданных значений λ и i:

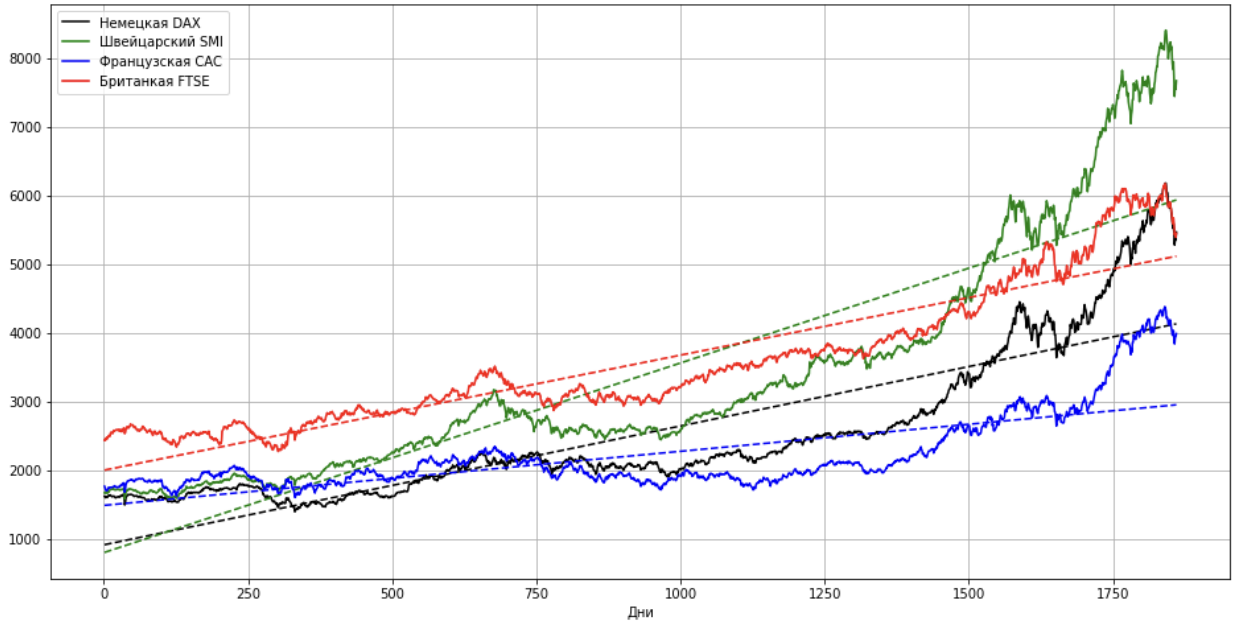


Вывод: по полученному графику мы видим, что лямбда увеличивает ошибку, что плохо характеризует точность предсказаний модели. В остальном, их точность получилась примерно одинаковой.

Задание 5

Загрузите данные из файла eustock.csv. Данные содержат ежедневные котировки на момент закрытия фондовых бирж: Germany DAX (Ibis), Switzerland SMI, France CAC, и UK FTSE. Постройте на одном графике все кривые изменения котировок во времени. Постройте линейную регрессию для каждой модели в отдельности и для всех моделей вместе. Оцените, какая из бирж имеет наибольшую динамику.

Графики изменения котировок во времени:



Метрики построенных регрессий:

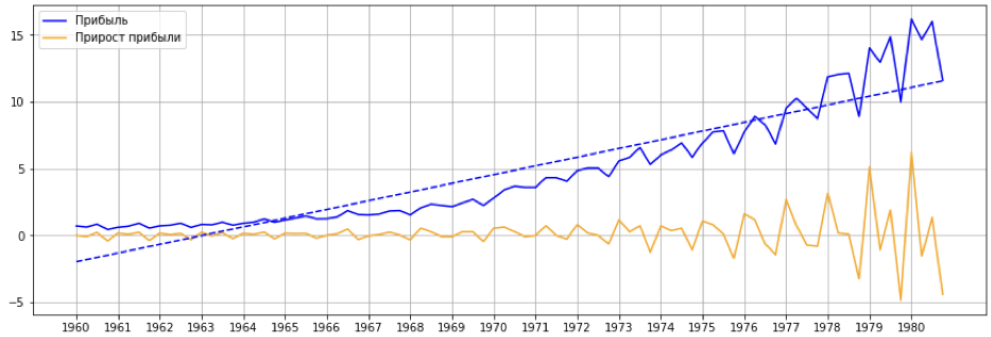
- Биржа DAX: 0.7331
- Биржа SMI: 0.7944
- Биржа CAC: 0.5303
- Биржа FTSE: 0.8482

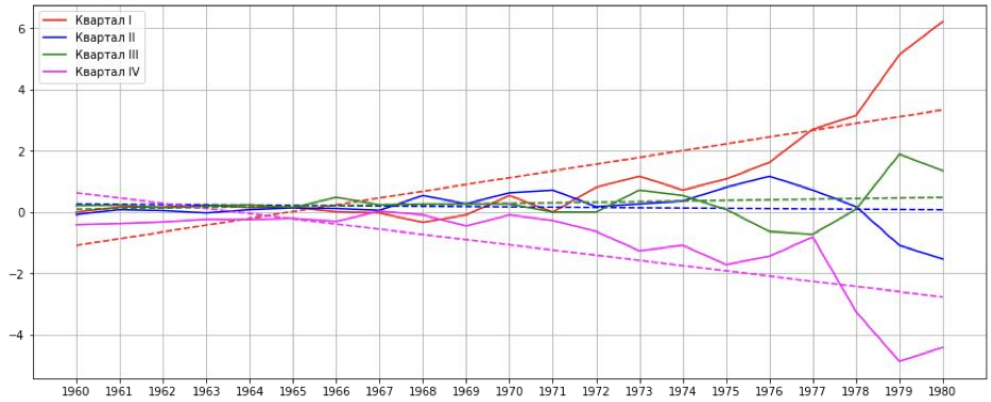
Вывод: из графика видно, что наибольший прирост на данном периоде произошел на швейцарской бирже. Что касается построения регрессора для всех графиков сразу, то это равносильно построению четырёх отдельных регрессоров. В данном случае метрика комбинированного регрессора равна 0.7265.

Задание 6

Загрузите данные из файла JohnsonJohnson.csv. Данные содержат поквартальную прибыль компании Johnson & Johnson с 1960 по 1980 гг. Постройте на одном графике все кривые изменения прибыли во времени. Постройте линейную регрессию для каждого квартала в отдельности и для всех кварталов вместе. Оцените, в каком квартале компания имеет наибольшую и наименьшую динамику доходности. Сделайте прогноз по прибыли в 2016 году во всех кварталах и в среднем по году.

Кривые изменения прибыли во времени и линейная регрессия для кварталов:





Вывод: согласно этим линейным моделям, в 2016 году компания будет иметь следующую прибыль:

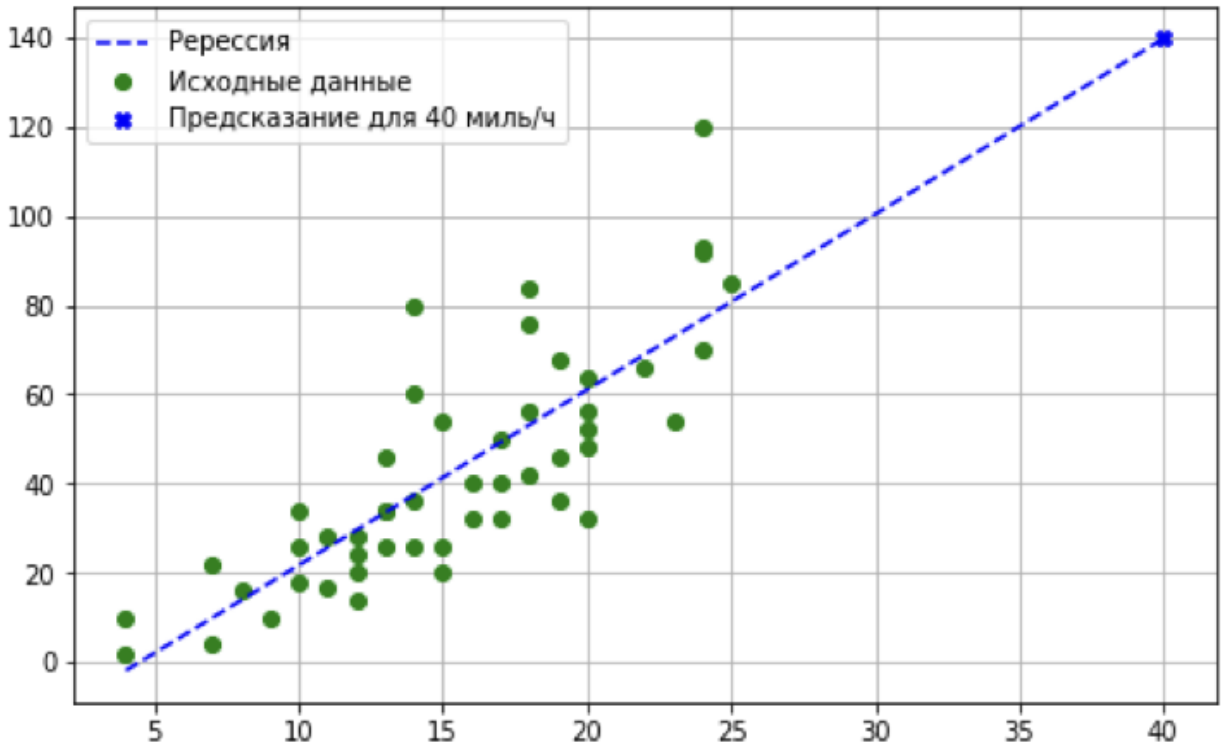
- Квартал I: 11.28
- Квартал II: -0.2702
- Квартал III: 1.165
- Квартал IV: -8.860
- Год: 34.56

Разумеется, говорить о правдоподобности этих предсказаний не приходится. Как минимум потому, что зависимость, как видно из графиков, у показателей нелинейная.

Задание 7

Загрузите данные из файла cars.csv. Данные содержат зависимости тормозного пути автомобиля (футы) от его скорости (мили в час). Данные получены в 1920 г. Постройте регрессионную модель и оцените длину тормозного пути при скорости 40 миль в час.

Регрессионная модель и оценка длины тормозного пути при скорости 40 миль в час:

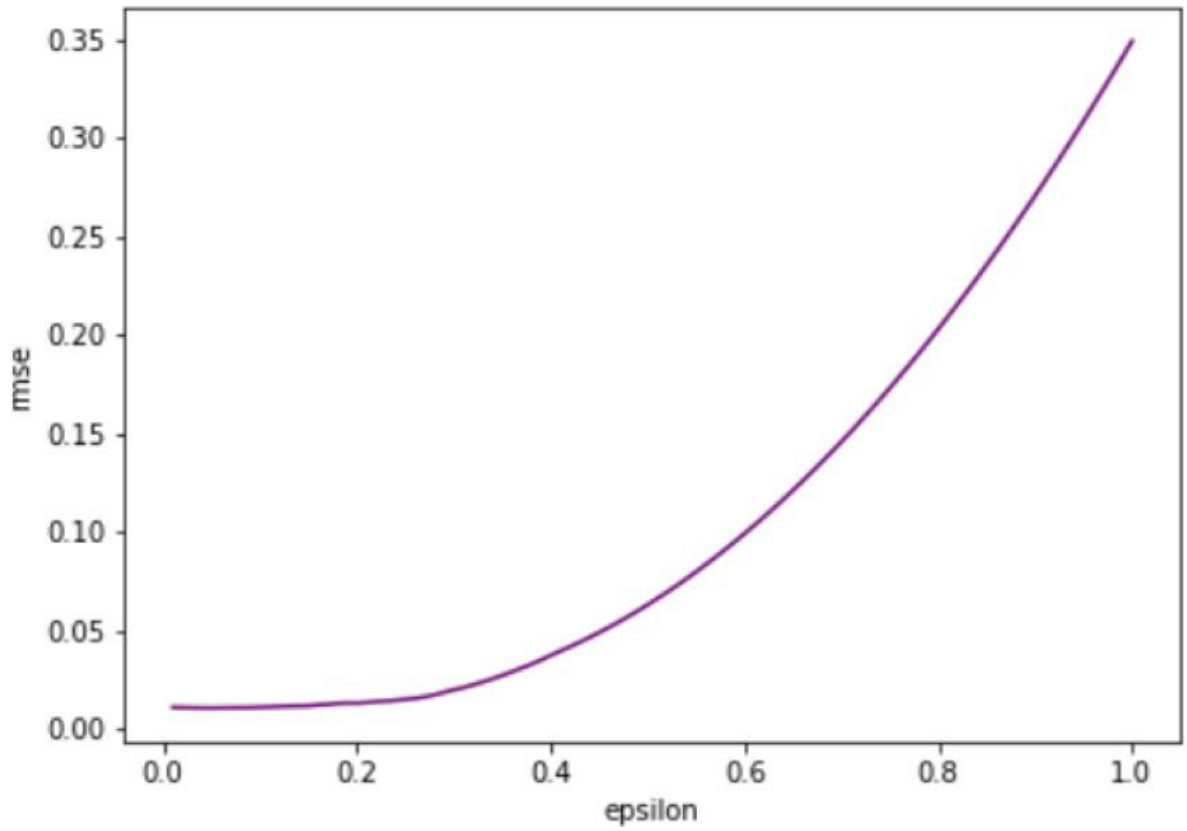


Вывод: регрессионная модель оценивает тормозной путь в 139.7 футов

Задание 8

Загрузите данные из файла svmdata6.txt. Постройте регрессионный алгоритм метода опорных векторов (sklearn.svm.SVR) с параметром C = 1, используя ядро "rbf". Отобразите на графике зависимость среднеквадратичной ошибки на обучающей выборке от значения параметра ε. Прокомментируйте полученный результат

График зависимости среднеквадратичной ошибки на обучающей выборке от значения параметра эпсилон:



Вывод: как мы можем видеть - эпсилон зависит экспоненциально

Задание 9

Загрузите набор данных из файла nsw74psid1.csv. Постройте регрессионное дерево (sklearn.tree.DecisionTreeRegressor) для признака re78. Постройте линейную регрессионную модель и SVM-регрессию для этого набора данных. Сравните качество построенных моделей, выберите оптимальную модель и объясните свой выбор

Было построено регрессионное дерево для признака re78. Была построена линейная регрессионная модель и SVM-регрессия для этого набора данных. Ниже представлены результаты:

```
DecisionTreeRegressor: 0.346435982735
LinearRegression: 0.578923465762
SVR-regression: 0.0456772359
```

Вывод: как мы можем видеть по результатам, все три модели недостаточно хорошо описывают исходные данные, но LinearRegression дает лучший результат.