**Лабораторная работа №5**

РЕГРЕССИЯ

По дисциплине «Машинное обучение»

Выполнил студент

группы 3530904/80102: Шерман М.Л.

Преподаватель: Селин И.А.

Оглавление

[Задачи 3](#_Toc71972612)

[Пункт 1 4](#_Toc71972613)

[Пункт 2 5](#_Toc71972614)

[Пункт 3 6](#_Toc71972615)

[Пункт 4 6](#_Toc71972616)

[Пункт 5 8](#_Toc71972617)

[Пункт 6 9](#_Toc71972618)

[Пункт 7 10](#_Toc71972619)

[Пункт 8 11](#_Toc71972620)

[Пункт 9 12](#_Toc71972621)

[Вывод 13](#_Toc71972622)

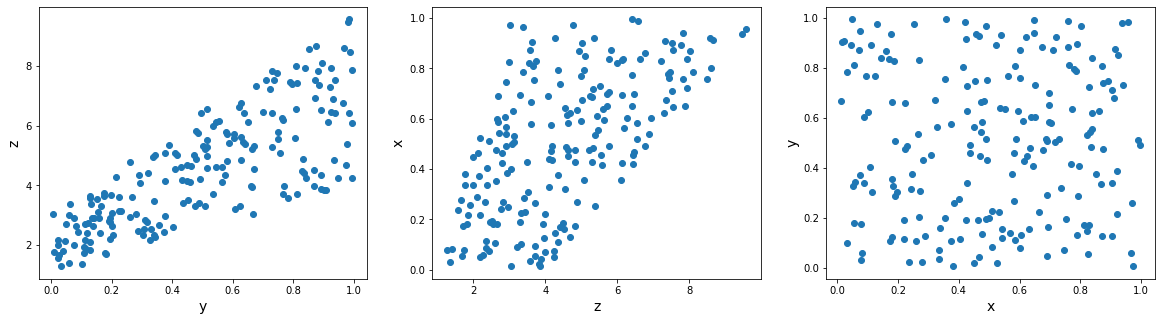
[Приложение 14](#_Toc71972623)

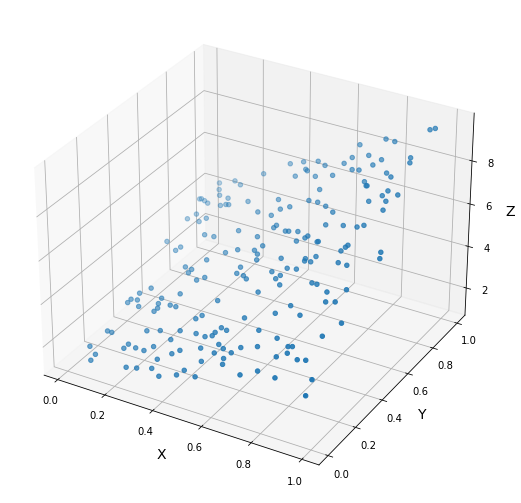
# Задачи

1. Загрузите данные из файла reglab1.txt. Постройте по набору данных регрессии, используя модели с различными зависимыми переменными. Выберите наиболее подходящую модель.
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.
3. Загрузите данные из файла cygage.txt. Постройте регрессию, выражающую зависимость возраста исследуемых отложений от глубины залегания, используя веса наблюдений. Оцените качество построенной модели.
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.
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.
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.
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.
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
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.

# Пункт 1

Рассмотрим данные. У нас есть три признака – **x**, **y**, **z**, рассмотрим их попарное расположение на плоскости:



А также полное представление в трёхмерном пространстве:

Из графиков видно, что данные образуют некое подобие плоскости, расположенную под углом к осям 0X, 0Y, 0Z. Двумерное представление позволяет рассмотреть попарную зависимость признаков. Так, визуально лучшая зависимость между переменными **z** и **y**, немного хуже между **x** и **z** и почти никакая между **x** и **y**. Основываясь на этом, можно сказать, что скорее всего лучшая модель будет с зависимой переменной **z**. Действительно:

|  |  |
| --- | --- |
| Зависимая переменная | Значение коэффициента детерминации |
| x | 0.9187 |
| y | 0.9505 |
| z | 0.9686 |

Коэффициент детерминации (**R^2**) — это доля [дисперсии](http://www.machinelearning.ru/wiki/index.php?title=%D0%94%D0%B8%D1%81%D0%BF%D0%B5%D1%80%D1%81%D0%B8%D1%8F_%D1%81%D0%BB%D1%83%D1%87%D0%B0%D0%B9%D0%BD%D0%BE%D0%B9_%D0%B2%D0%B5%D0%BB%D0%B8%D1%87%D0%B8%D0%BD%D1%8B&action=edit) зависимой переменной, объясняемая рассматриваемой моделью. Чем ближе данное значение к 1, тем сильнее зависимость между признаками и предсказываемой переменной. Таким образом, оптимальным выбором будет сделать **z** зависимой переменной.

# Пункт 2

Реализуем алгоритм уменьшения количества признаков, используемых для построения регрессии, чтобы найти подмножество признаков, минимизирующих остаточную сумму квадратов (RSS).

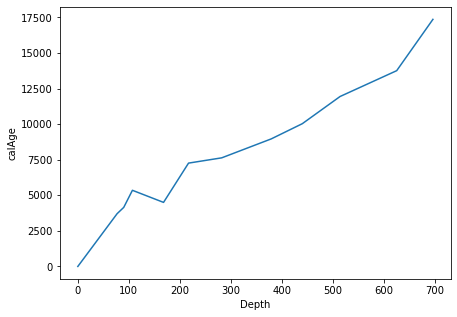
В результате выполнения алгоритма получили следующие данные:

|  |  |
| --- | --- |
| Рассматриваемые признаки | RSS |
| X1 X2 X3 X4 | 0.0302 |
| X1 X2 X3 | 0.0594 |
| X1 X2 X4 | 0.0705 |
| X1 X3 X4 | 26.3869 |
| X2 X3 X4 | 55.2964 |
| X1 X2 | 0.095 |
| X1 X3 | 26.2587 |
| X1 X4 | 26.5816 |
| X2 X3 | 55.483 |
| X2 X4 | 55.4525 |
| X3 X4 | 69.4893 |
| X1 | 26.3517 |
| X2 | 55.6964 |
| X3 | 69.5022 |
| X4 | 69.8646 |

В результате наименьшее значение остаточной суммы квадратов удалось получить, используя все 4 признака. Если есть строгое ограничение на количество используемых признаков, то можно исключить X4, либо X3, либо одновременно X4 и X3. Остаточная сумма при этом увеличится, но не критично, как при исключении X1 или X2.

# Пункт 3

Построим график зависимости возраста исследуемых отложений от глубины залегания:

Данные почти линейно зависимы. Воспользуемся линейной регрессией с заданными весами и посмотрим, насколько качественной будет полученная модель.

Коэффициент детерминации полученной модели составил 0.9737, что говорит о сильной зависимости между признаками и предсказываемой переменной.

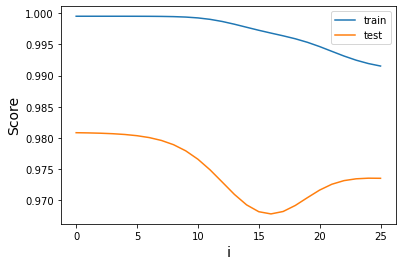
# Пункт 4

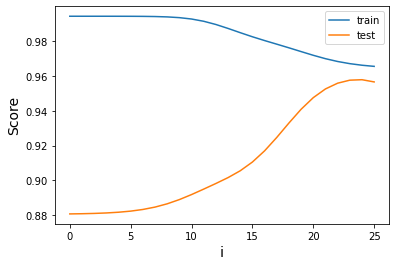
Для предоставленных данных необходимо построить линейную и гребневую регрессии и оценить результаты предсказания. В результате получаем следующие показатели (использовались модели с параметрами по умолчанию):

|  |  |  |
| --- | --- | --- |
| Выборка | Линейная регрессия | Гребневая регрессия |
| Train | 0.9995 | 0.9972 |
| Test | 0.9809 | 0.9682 |

Обе модели показали отличные результаты, однако линейная регрессия предсказывала немного лучше гребневой.

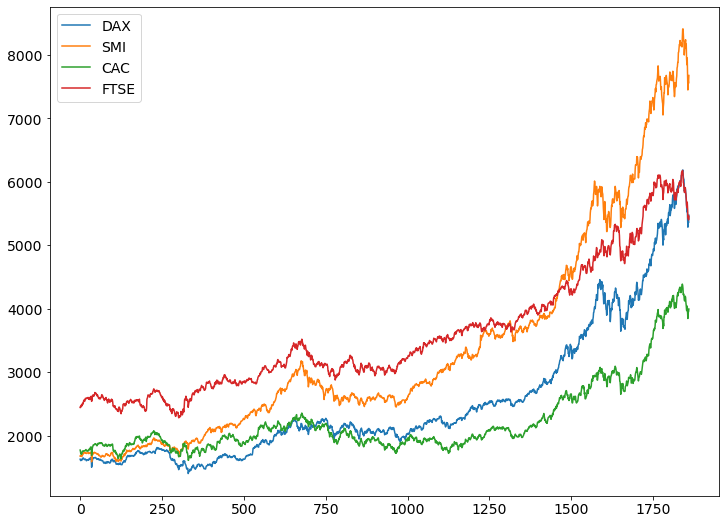
Рассмотрим зависимость точности предсказания гребневой регрессии от параметра , i=0, ..., 25.

Как видно, наилучшая точность достигается при небольшом значении i (примерно в промежутке от 0 до 5). Параметр λ, который мы изменяли, отвечает за «силу» регуляризации. Регуляризация улучшает обусловленность задачи и уменьшает дисперсию оценок. Данный параметр используют, чтобы предотвратить переобучение. В идеале увеличение данного параметра должно улучшать точность, однако из-за случайности разбиения и величины обучающей выборки нам удалось получить достаточно хорошие значения при минимальном значении данного параметра. Взглянем на результаты другого разбиения:



Данный график уже похож на ожидаемое поведение модели при увеличении строгости регуляризации, однако на нём видно, что значение точности хуже, чем при другом разбиении. Для более точного и устойчивого предсказания стоит увеличить размерность выборок (либо хотя бы тренировочной).

# Пункт 5

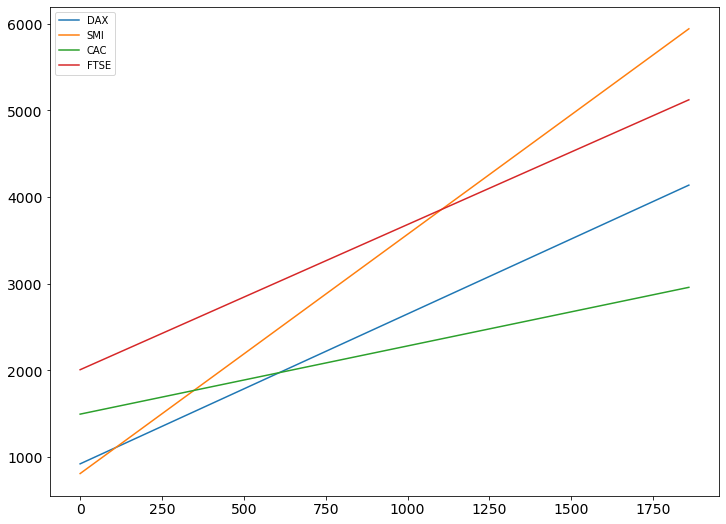
Рассмотрим пять различных котировок пяти различных бирж и выясним, какая из них имеет наибольшую динамику.

Построим 5 линейных регрессий (для каждой биржи по отдельности и всех вместе), оценим полученные коэффициенты детерминации и оценочные коэффициенты для задачи линейной регрессии.

|  |  |  |
| --- | --- | --- |
| Рассматриваемые биржи | Коэффициент детерминации | Оценочный коэффициент |
| DAX | 0.733 | 1.729 |
| SMI | 0.794 | 2.759 |
| CAC | 0.530 | 0.786 |
| FTSE | 0.848 | 1.674 |
| Все | 0.726 | Массив из всех предыдущих коэффициентов |

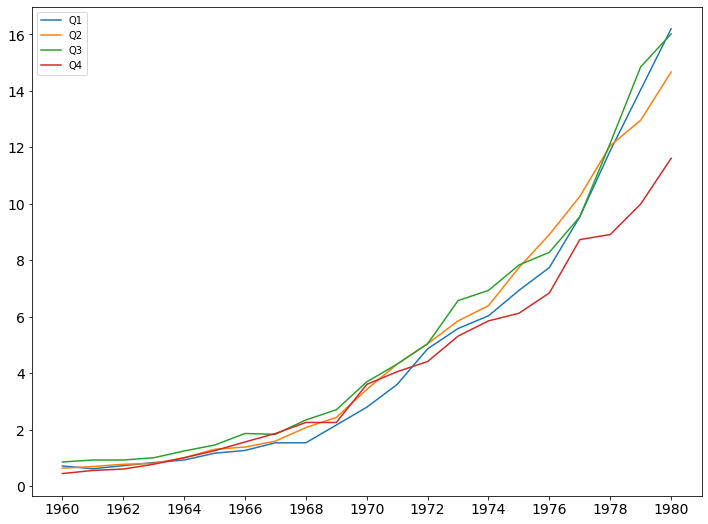
Коэффициент детерминации групповой модели равен среднему значению все коэффициентов моделей с одной значимой переменной.

Когда у нас есть один признак и одна зависимая переменная, линейная регрессия строит прямую y=ax+b. Оценочный коэффициент как раз равен «a» и чем он больше, тем быстрее растёт функция. Следовательно, наибольшую динамику будет иметь котировки биржи SMI. Посмотрим на графике:

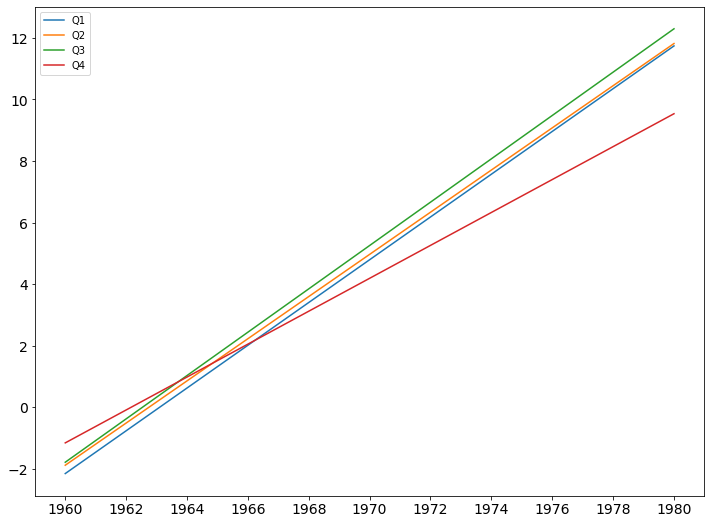


Так и оказалось. Наибольшую динамику имеет биржа SMI.

# Пункт 6

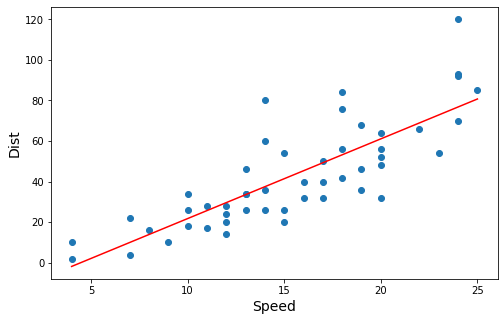
Предоставлены данные о поквартальной прибыли компании Johnson & Johnson. Необходимо построить линейную регрессию для каждого квартала по отдельности и всех кварталов вместе. Будем действовать аналогично предыдущему пункту.

|  |  |  |  |
| --- | --- | --- | --- |
| Рассматриваемый квартал | Коэффициент детерминации | Оценочный коэффициент | Прогноз прибыли в 2016 году |
| 1 | 0.836 | 0.695 | 36.76 |
| 2 | 0.891 | 0.685 | 36.489 |
| 3 | 0.865 | 0.704 | 37.654 |
| 4 | 0.928 | 0.534 | 28.793 |
| Весь год | 0.880 | Массив из всех предыдущих коэффициентов | 34.924 |

Основываясь на оценочных коэффициентах, компания будет иметь наибольшую динамику доходности в третьем квартале и наименьшую в четвёртом.

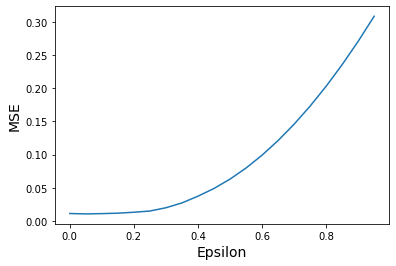
# Пункт 7

Исследуем зависимость длины тормозного пути от скорости автомобиля.

Полученное значение коэффициента детерминации: 0.651

Предсказанная длина тормозного пути при 40 милях в час: 139.717

# Пункт 8

Обучим модель на основе метода опорных векторов на предоставленных данных. Рассмотрим зависимость среднеквадратичной ошибки полученной модели от параметра ɛ:

Параметр ɛ отвечает за предел величины отклонения от опорных векторов, при которых не начисляется штраф. Чем больше данное значение, тем больше ошибок мы допускаем в нашем решении. Поэтому при увеличении данного параметра у нас увеличивается среднеквадратичная ошибка.

# Пункт 9

В данном пункте будем рассматривать 3 типа моделей: регрессионное дерево, линейную регрессию и svm-регрессию. Для оценки моделей будем рассматривать коэффициент детерминации.

Возьмём 5 моделей:

* Линейную регрессию
* SVM-регрессию с параметрами по умолчанию
* SVM-регрессию с ядром ‘poly’ и степенью 2
* Регрессионное дерево с параметрами по умолчанию
* Регрессионное дерево с максимальной глубиной 4

Мы специально взяли по два экземпляра SVM-регрессии и регрессионного дерева, так как с параметрами по умолчанию значение коэффициента детерминации будет слишком малым:

|  |  |  |
| --- | --- | --- |
| Модель | Коэффициент детерминации (тренировочная выборка) | Коэффициент детерминации (тестовая выборка) |
| LinearRegression | 0.5807 | 0.6065 |
| SVM | 0.0216 | 0.0128 |
| SVR(kernel='poly', degree=2) | 0.1813 | 0.2106 |
| DecisionTreeRegressor | 0.997 | 0.1674 |
| DecisionTreeRegressor(max\_depth=4) | 0.5931 | 0.5165 |

SVM-регрессия показывает очень плохой результат, и хоть при использовании другого ядра точность повышается, она всё равно остаётся достаточно малой. Поведение регрессионного дерева без ограничений предсказуемо, так как данный алгоритм склонен к переобучению. При ограничении его глубины удаётся получить неплохой результат. Однако наилучший показатель коэффициента детерминации имеет линейная регрессия.

В итоге, оптимальным выбором будет использование линейной регрессии, либо регрессионного дерева с дополнительными ограничениями.

# Вывод

В результате проделанной работы удалось познакомиться с регрессионными моделями и метриками для их оценивания. Приобретены навыки применения линейной регрессии, svm-регрессии, а также регрессионного дерева. Также улучшены навыки по предобработке данных

# Приложение

Весь код и графики можно найти в следующем репозитории:

<https://github.com/Mark-Sherman-SE/ML-Labs>