Учреждение образования

«Белорусский государственный университет

информатики и радиоэлектроники»

Кафедра информатики

Отчет по лабораторной работе:

**Лабораторная работа №3 “Переобучение и регуляризация”**

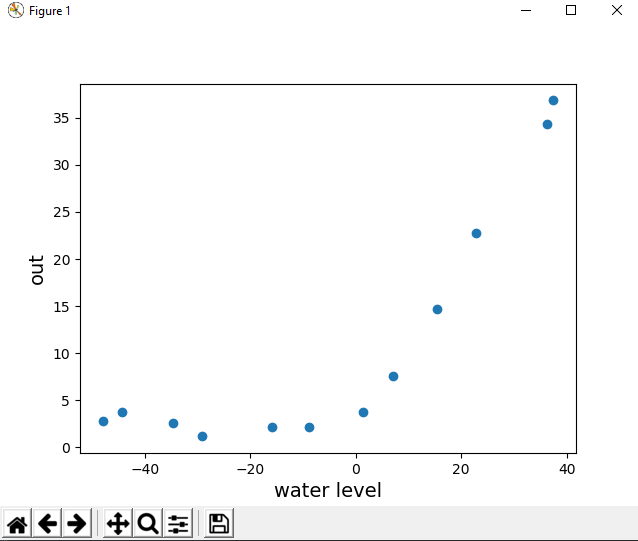
Выполнил: Карп Александр Игоревич

магистрант кафедры информатики

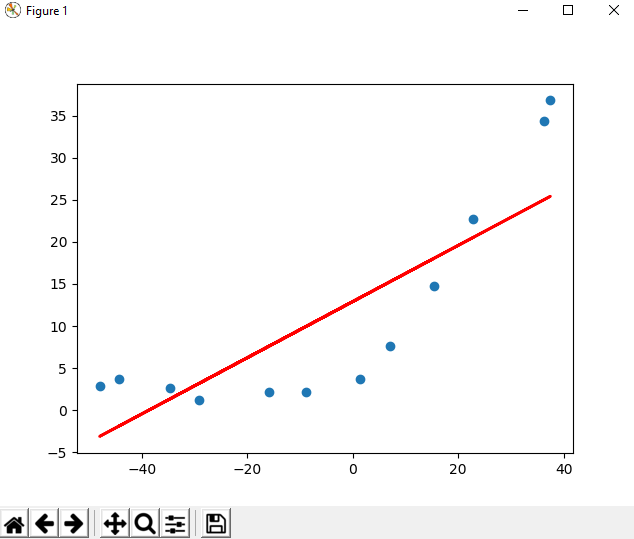
группа №858641

Минск 2019

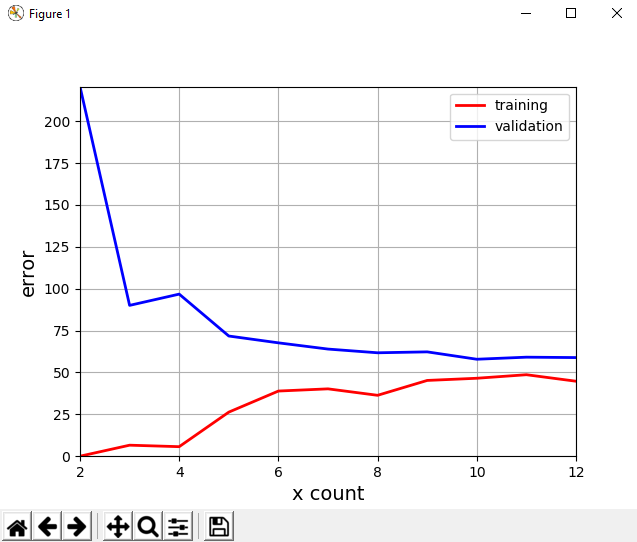
1. Загрузите данные **ex3data1.mat** из файла.  
   <https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L62>
2. Постройте график, где по осям откладываются X и y из обучающей выборки.  
   <https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L75>



3) Реализуйте функцию стоимости потерь для линейной регрессии с L2-регуляризацией.  
<https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L88>  
  
4) Реализуйте функцию градиентного спуска для линейной регрессии с L2-регуляризацией.  
<https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L91>  
  
5) Постройте модель линейной регрессии с коэффициентом регуляризации 0 и постройте график полученной функции совместно с графиком из пункта 2. Почему регуляризация в данном случае не сработает?  
<https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L93>

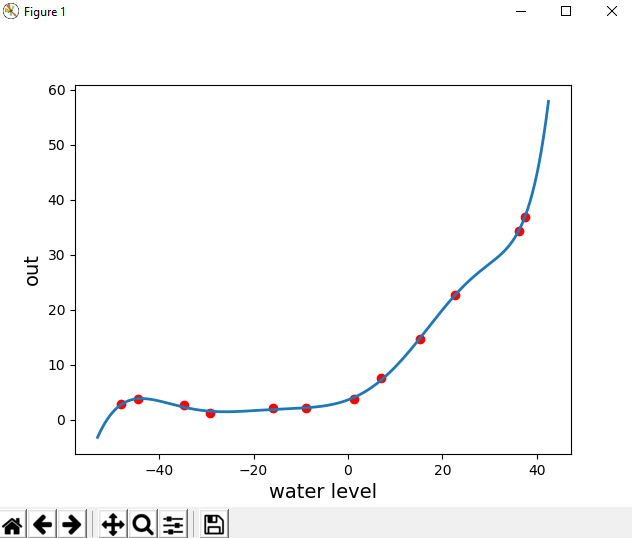


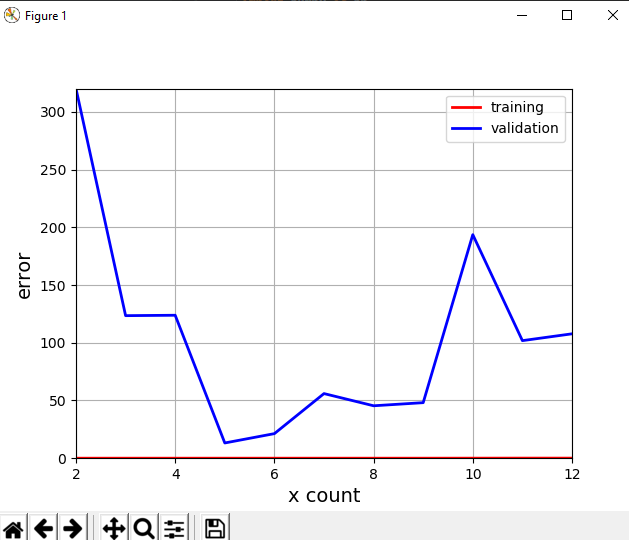
Регуляризация в данном случае не сработает в первую очереди из-за того, что наша функция линейна. Во-вторых, будь она нелинейна, регуляризация бы не сработала, так как регуляризационный коэффициент равен нулю.

1. Постройте график процесса обучения (learning curves) для обучающей и валидационной выборки. По оси абсцисс откладывается число элементов из обучающей выборки, а по оси ординат - ошибка (значение функции потерь) для обучающей выборки (первая кривая) и валидационной выборки (вторая кривая). Какой вывод можно сделать по построенному графику?  
     
   <https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L109>  
   

С ростом количества данных в обучающей выборке ошибка на тренировочном датасете растет, а на валидационном наоборот, уменьшается. Со временем обе ошибка стабилизируются и остаются на одном уровне. Исходные данные, судя по графику из п.2, не могут быть точно апроксимированы линейной функцией, из-за этого мы видим достаточно высокую ошибку.

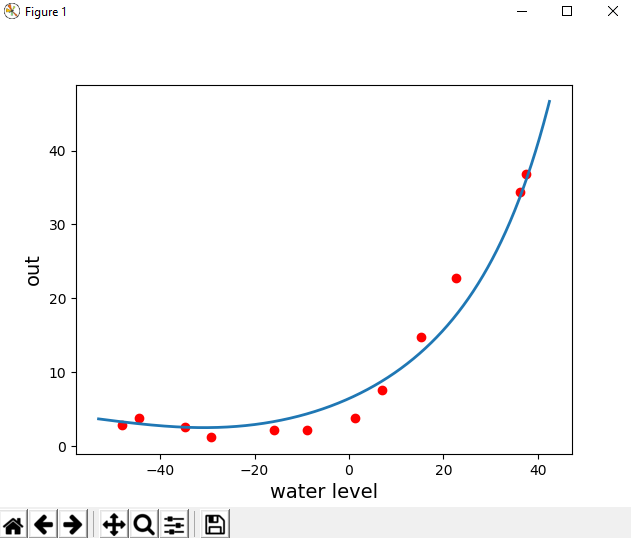
1. Реализуйте функцию добавления p - 1 новых признаков в обучающую выборку (X2, X3, X4, …, Xp).  
   <https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L114>
2. Поскольку в данной задаче будет использован полином высокой степени, то необходимо перед обучением произвести нормализацию признаков.  
   <https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L114>
3. Обучите модель с коэффициентом регуляризации 0 и p = 8.  
   <https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L123>
4. Постройте график модели, совмещенный с обучающей выборкой, а также график процесса обучения. Какой вывод можно сделать в данном случае?  
   <https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L136>

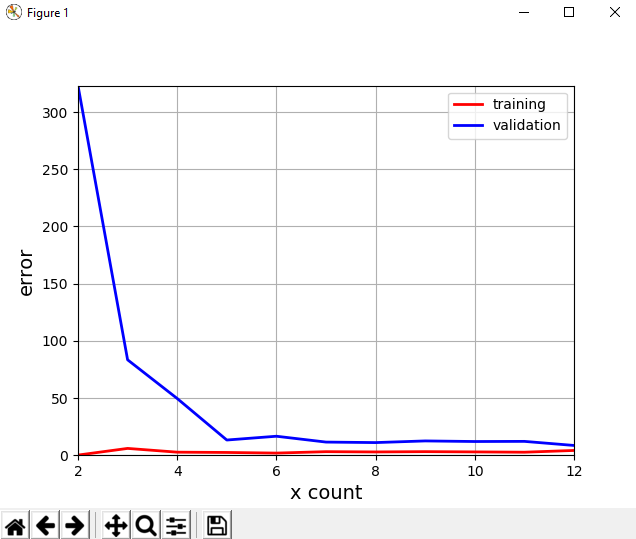




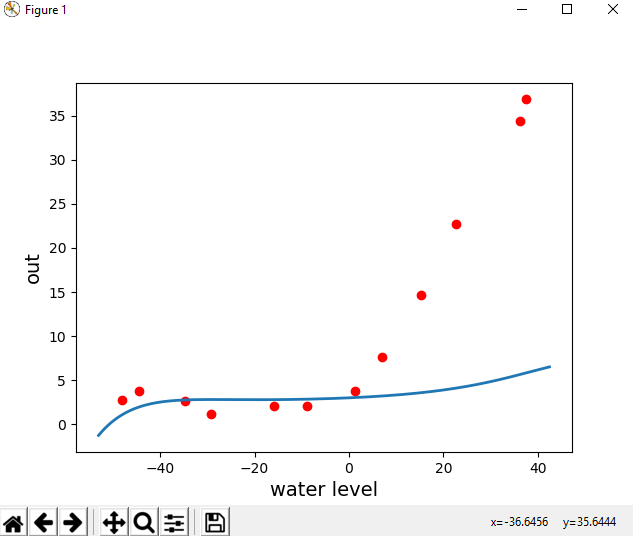
Здесь можно наблюдать, что модель идеально подстроилась под данные из тренировочного сета, но ошибка на валидационном сете непредсказуема, т.к. .

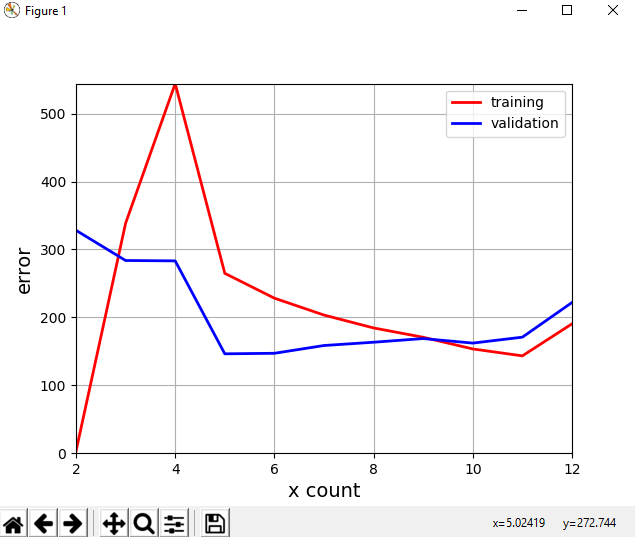
11) Постройте графики из пункта 10 для моделей с коэффициентами регуляризации 1 и 100. Какие выводы можно сделать?  
<https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L155>





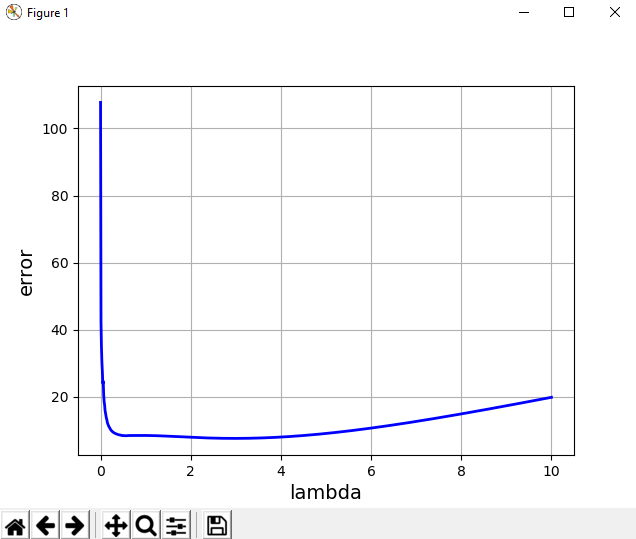
Как видно из первого графика, регуляризация «сгладила» нашу функцию, она теперь не идет строго по точкам из тренировочного датасета. Как видно из второго графика, регуляризация помогла нам избежать переобучения и ошибка как на тренировочном, так и на валидационном сете минимальна.





Как видно из графиков, штраф при таком коэффициенте регуляризации оказывается слишком большим, модель не может точно апроксимировать исходные данные

12) С помощью валидационной выборки подберите коэффиент регуляризации, который позволяет достичь наименьшей ошибки. Процесс подбора отразите с помощью графика (графиков).  
<https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L184>



13) Вычислите ошибку (потерю) на контрольной выборке.  
<https://github.com/Karpengold/machine_learning/blob/master/lab3/index.py#L203>