Министерство образования Республики Беларусь

Учреждение образования

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет компьютерных систем и сетей

Кафедра информатики

Дисциплина: Машинное обучение (МО)

Отчет о лабораторной работе №3

«Переобучение и регуляризация»

|  |
| --- |
| **Выполнил:** |
| **Студент гр. 858341** |
| **Немкович А. В.** |

|  |
| --- |
| **Проверил:** |
| **Стержанов М. В.** |

**Минск 2019**

Содержание

[Постановка задачи 3](#_Toc25275643)

[Ход выполнения 4](#_Toc25275644)

[Выводы 13](#_Toc25275645)

# Постановка задачи

Набор данных **ex3data1.mat** представляет собой файл формата \*.mat (т.е. сохраненного из Matlab). Набор содержит две переменные X (изменения уровня воды) и y (объем воды, вытекающий из дамбы). По переменной X необходимо предсказать y. Данные разделены на три выборки: обучающая выборка (X, y), по которой определяются параметры модели; валидационная выборка (Xval, yval), на которой настраивается коэффициент регуляризации; контрольная выборка (Xtest, ytest), на которой оценивается качество построенной модели.

**Задание.**

1. Загрузите данные **ex3data1.mat** из файла.
2. Постройте график, где по осям откладываются X и y из обучающей выборки.
3. Реализуйте функцию стоимости потерь для линейной регрессии с L2-регуляризацией.
4. Реализуйте функцию градиентного спуска для линейной регрессии с L2-регуляризацией.
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?
6. 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?
7. Реализуйте функцию добавления p - 1 новых признаков в обучающую выборку (X2, X3, X4, …, Xp).
8. Поскольку в данной задаче будет использован полином высокой степени, то необходимо перед обучением произвести нормализацию признаков.
9. Обучите модель с коэффициентом регуляризации 0 и p = 8.
10. Постройте график модели, совмещенный с обучающей выборкой, а также график процесса обучения. Какой вывод можно сделать в данном случае?
11. Постройте графики из пункта 10 для моделей с коэффициентами регуляризации 1 и 100. Какие выводы можно сделать?
12. С помощью валидационной выборки подберите коэффиент регуляризации, который позволяет достичь наименьшей ошибки. Процесс подбора отразите с помощью графика (графиков).
13. Вычислите ошибку (потерю) на контрольной выборке.
14. Ответы на вопросы представьте в виде отчета.

# Ход выполнения

**Задания 1-2**: Загрузка данных и построение графика, где по осям откладываются X и y из обучающей выборки.

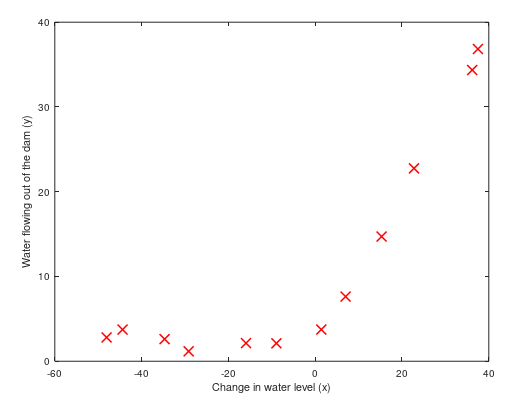
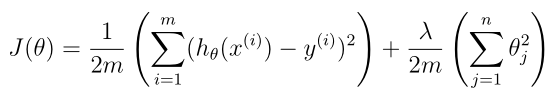


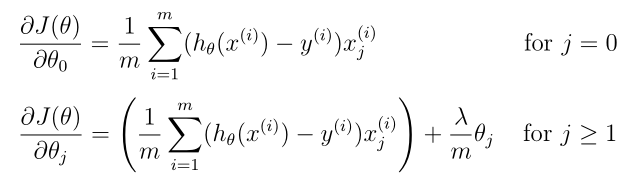
Рис 1. – исходные данные (ex3data1.mat)

**Задание 3-4**: Реализация функций стоимости потерь и градиентного спуска для линейной регрессии с L2-регуляризацией.

Формула L2-регуляризации для функции потерь для линейной регрессии:



Формула L2-регуляризации для частных производных в градиентом спуске для линейной регрессии:



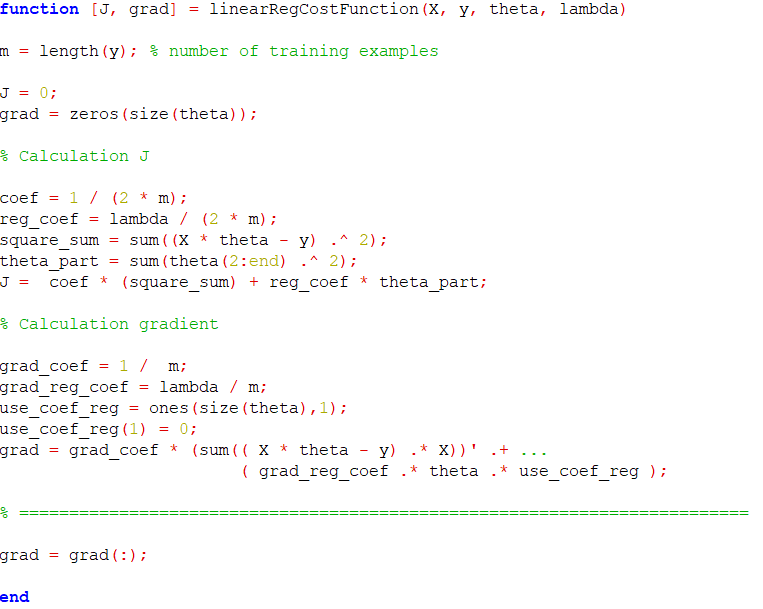


Рис 2. – Код функции потерь и градиентного спуска

**Задание 5**: Построение модели линейной регрессии с коэффициентом регуляризации 0 и построение графика полученной функции совместно с графиком из пункта 2.

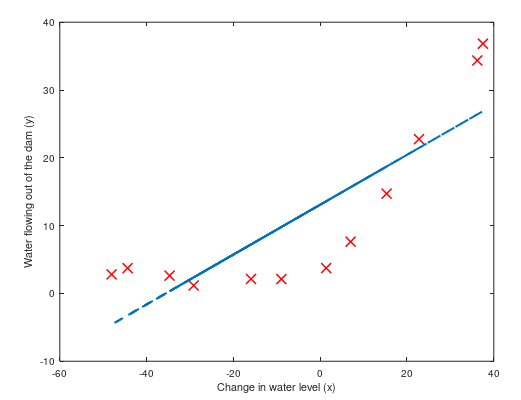


Рис 3. – График полученной модели совместно с исходными данными

**Задание 6**: Построение графика 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).

Для получения данных для данного графика использовалась функция learningCurve.

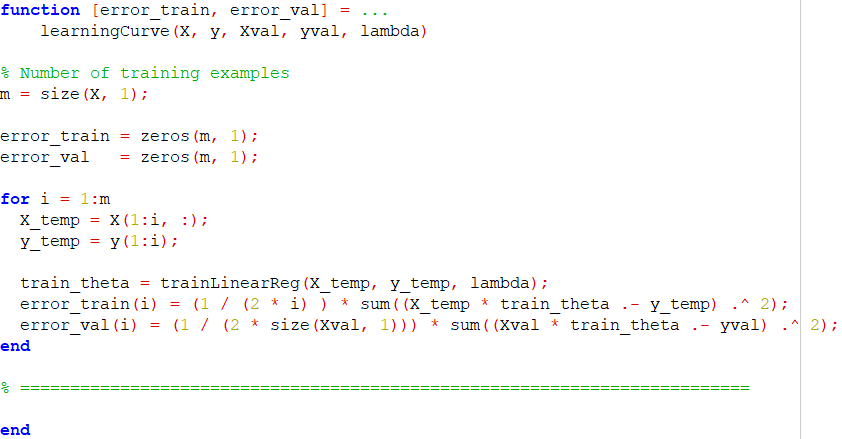


Рис 4. – Код функции learningCurve

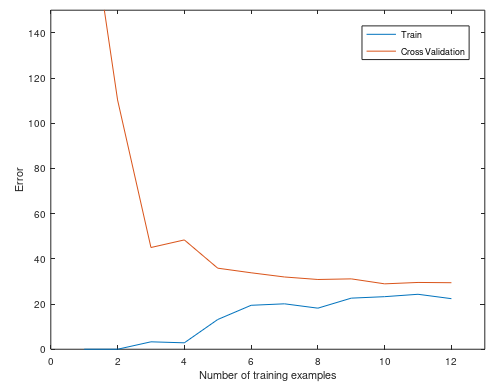


Рис 5. – график процесса обучения

**Задание 7**: Реализация функции добавления p - 1 новых признаков в обучающую выборку (X2, X3, X4, …, Xp).

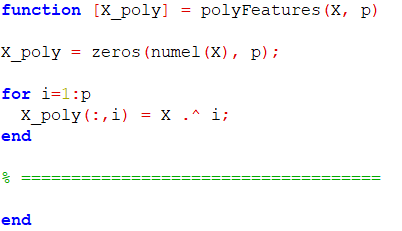


Рис 6. – Код функции polyFeatures

**Задание 8**: Нормализация признаков.

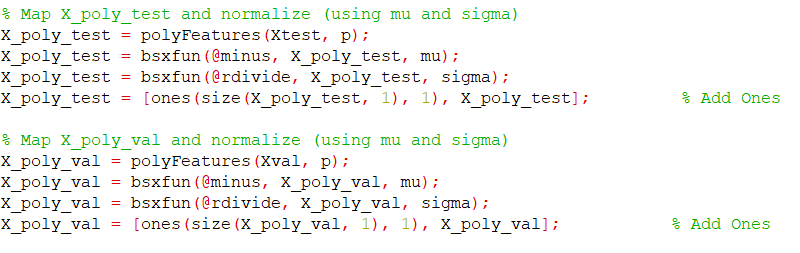


Рис 7. – Код нормализации признаков

**Задания 9-10**: Обучение модели с коэффициентом регуляризации 0 и p = 8. Построение графика модели, совмещенного с обучающей выборкой, а также графика процесса обучения.

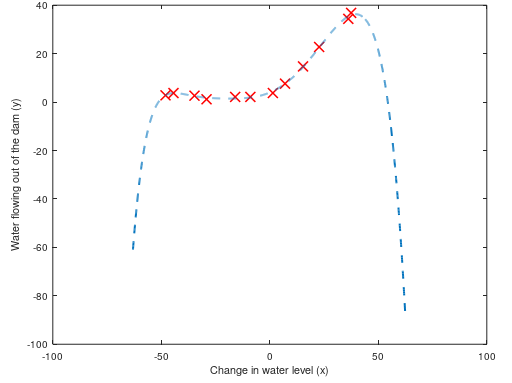


Рис 8. – График модели, совмещенный с обучающей выборкой (lambda = 0)

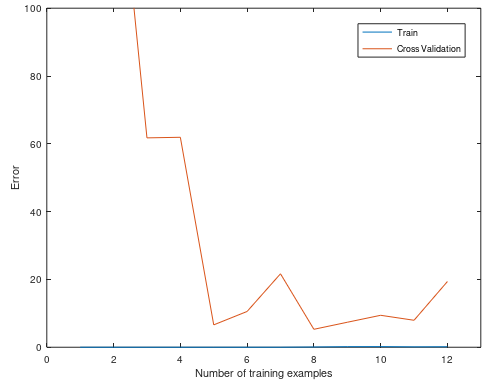


Рис 9. – График процесса обучения (lambda = 0)

**Задание 11**: Построение графика из пункта 10 для моделей с коэффициентами регуляризации 1 и 100.

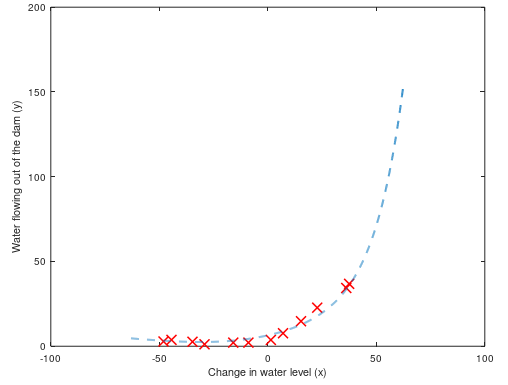


Рис 10. – График модели, совмещенный с обучающей выборкой (lambda = 1)

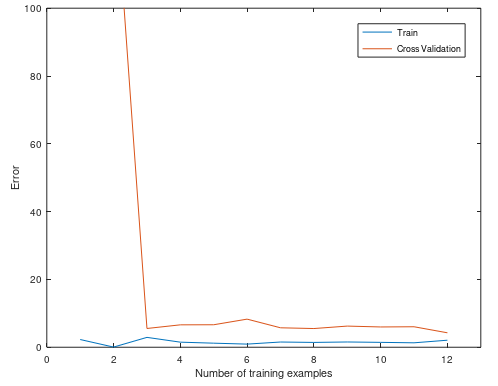


Рис 11. – График процесса обучения (lambda = 1)

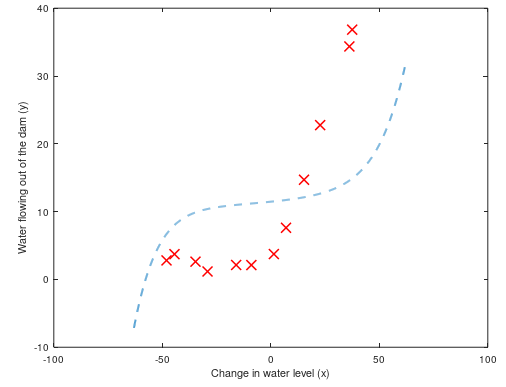


Рис 12. – График модели, совмещенный с обучающей выборкой (lambda = 100)

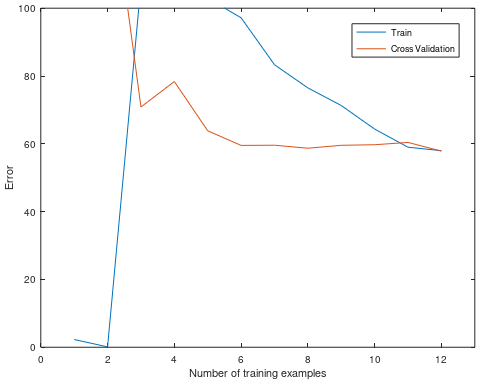


Рис 13. – График процесса обучения (lambda = 100)

**Задание 12**: Подбор коэффициента регуляризации, который позволяет достичь наименьшей ошибки.

Для подбора коэффициента регуляризации использовалась функция validationCurve.

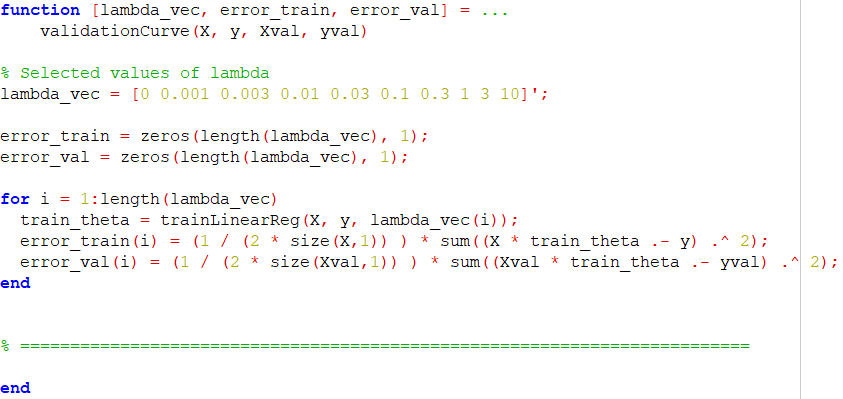


Рис 14. – код функции validationCurve

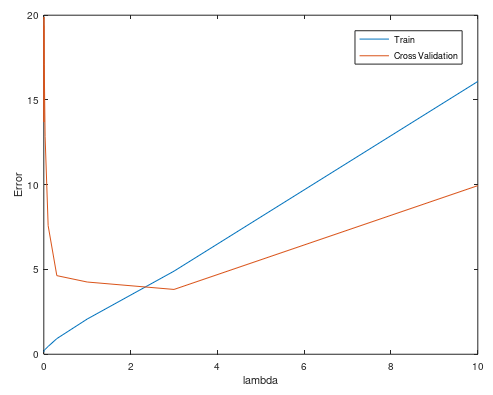


Рис 15. – график подбора коэффициента регуляризации

**Задание 13**: Вычисление ошибки (потери) на контрольной выборке.



Рис 16. – Сравнение ошибки на тренировочный и контрольной выборках

# Выводы

Мы научились определять проблемы недообучения и переобучения модели с помощью графика процесса обучения. Научились на практике устранять проблему переобучения с помощью регуляризации. Также научились добавлять новые признаки на основе текущих для решения проблемы недообученной модели.

Наш процесс принятия решения можно разбить следующим образом:

* Получение большего количества обучающих примеров – исправляет переобученную модель;
* Уменьшение набора признаков – исправляет переобученную модель;
* Добавление признаков – исправляет недообученную модель;
* Добавление полиномиальных признаков – исправляет недообученную модель;
* Уменьшение λ – исправляет недообученную модель;
* Увеличение λ – исправляет переобученную модель.