HY543 | Assignment 2

Papageorgiou Efthymios - csdp1344

Million Song Pipeline

This report answers every question (‘?’ at the end of the sentence) into the assignment document.

1 Data Collection

[1.2.1] The give Dataset consists of **4680** data points.

[1.4.2] In the parsedDataPoints the smallest label is 1926 and the largest is 2010. Since, we sub every value with the minimum (1926) we expect:

- **0** to be the smallest (1926-1926)
- **84** to be the largest (2010-1926)

In any case, the code computes those values.

[1.5.3]

- Number of elements in trainData: **3805**
- Number of elements in valData: **431**
- Number of elements in testData: **444**
- Sum of counts: **4680** -> (3805 + 431 + 444 = 4680)

2 Create a Baseline Model

Nothing to say here 🙂

3 Linear Regression with Gradient Descent

[3.3.3] The Gradient Descent doesn’t converge!

[3.3.4] It appears that the number of iterations is not the determining factor. As illustrated in Figure 1, even with the same number of iterations (50), gradient descent converges for an alpha value of 0.002.

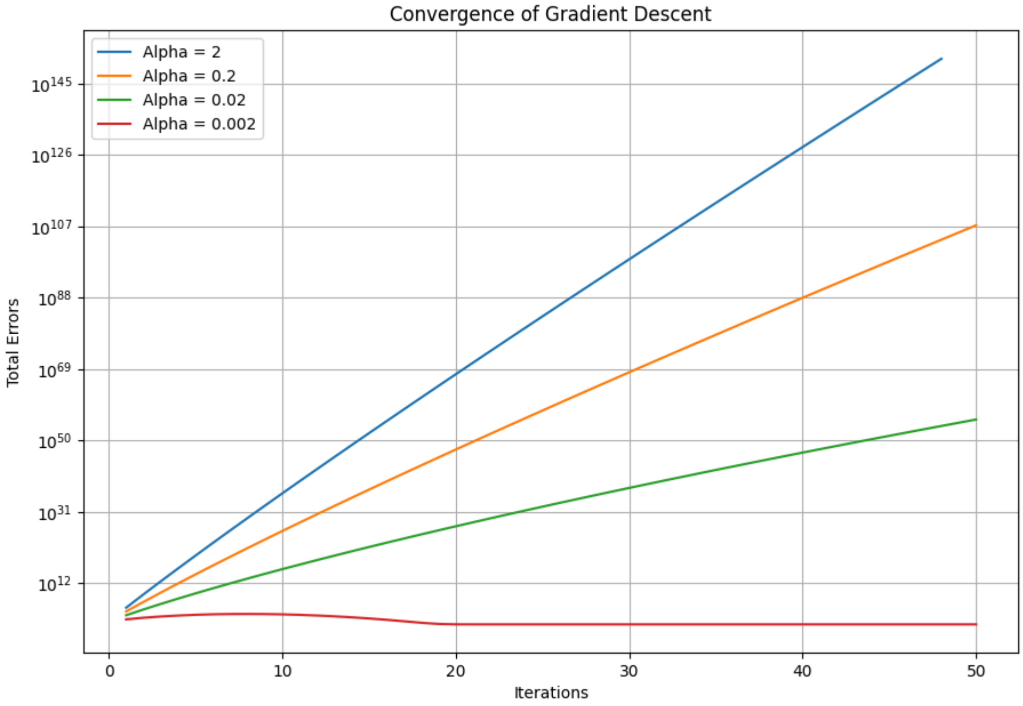


Figure 1: Comparison of Gradient Descent Convergence for Different Alpha Values

4 Train using MLlib and grid search

[4.2.1] The RMSE of the best model is **10.918536892986703**

[4.2.2] The above above RMSE is achieved by the regularizatio parameter **1.0E-5**

5 Add interactions between features

nothing to say here :)

6 How to run the project

Prerequisites: Java 11, sbt version 1.99, scala 2.13.11

Warning: If you want to use Java 17, you have to set a VM flag

the project structure is:
.bsp build.sbt hw2.pdf project src target

Before run the project, edit the file MilionSongPipeline.scala in lline 26 (val dataset = “path_to_dataset”)

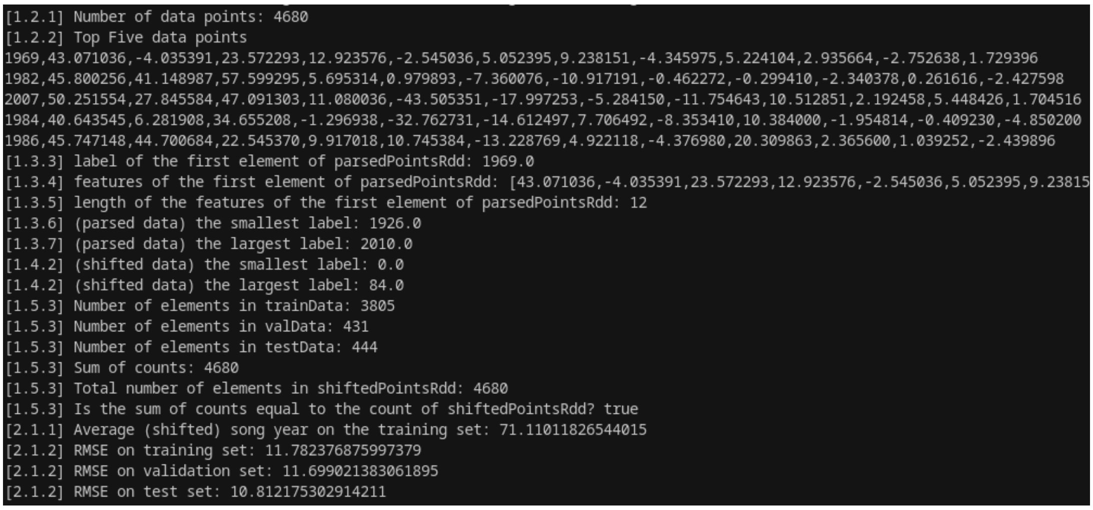
Compile and Run the project:

$sbt run

Clean the project:

$sbt clean

Example Output:



```
[1.2.1] Number of data points: 4680
[1.2.2] Top Five data points
1969,43.071036,-4.035391,23.572293,12.923576,-2.545036,5.052395,9.238151,-4.345975,5.224104,2.935664,-2.752638,1.729396
1982,45.800256,41.148987,57.599295,5.695314,0.979893,-7.360076,-10.917191,-0.462272,-0.299410,-2.340378,0.261616,-2.427598
2007,50.251554,27.845584,47.091303,11.080036,-43.505351,-17.997253,-5.284150,-11.754643,10.512851,2.192458,5.448426,1.704516
1984,40.643545,6.281908,34.655208,-1.296938,-32.762731,-14.612497,7.706492,-8.353410,10.384000,-1.954814,-0.409230,-4.850200
1986,45.747148,44.700684,22.545370,9.917018,10.745384,-13.228769,4.922118,-4.376980,20.309863,2.365600,1.039252,-2.439896
[1.3.3] label of the first element of parsedPointsRdd: 1969.0
[1.3.4] features of the first element of parsedPointsRdd: [43.071036,-4.035391,23.572293,12.923576,-2.545036,5.052395,9.23815
[1.3.5] length of the features of the first element of parsedPointsRdd: 12
[1.3.6] (parsed data) the smallest label: 1926.0
[1.3.7] (parsed data) the largest label: 2010.0
[1.4.2] (shifted data) the smallest label: 0.0
[1.4.2] (shifted data) the largest label: 84.0
[1.5.3] Number of elements in trainData: 3805
[1.5.3] Number of elements in valData: 431
[1.5.3] Number of elements in testData: 444
[1.5.3] Sum of counts: 4680
[1.5.3] Total number of elements in shiftedPointsRdd: 4680
[1.5.3] Is the sum of counts equal to the count of shiftedPointsRdd? true
[2.1.1] Average (shifted) song year on the training set: 71.11011826544015
[2.1.2] RMSE on training set: 11.782376875997379
[2.1.2] RMSE on validation set: 11.699021383061895
[2.1.2] RMSE on test set: 10.812175302914211
```