Karathanasis Dimitris

CS-543

March 2019

Assignment 2

Exercise 1

1.2

BaseRdd size = 4680

Top 5 =

1.

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

2.

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

3.

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

4.

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

5.

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. First element’s label: 1969.0

4. First element’s features:

[43.071036,-4.035391,23.572293,12.923576,-2.545036,5.052395,9.238151,-4.345975,5.224104,2.935664,-2.752638,1.729396]

5. Length of the features of the first element : 12

6. Min: 1926.0 Max: 2010.0

1.4

2. Min: 0.0 Max: 84.0

1.5

3. shiftedPointsRdd count = 4680

trainData count = 3745

valData count = 459

testData count = 476

They add up to shiftedPointsRdd count so we are good

Exercise 2

2.1

1. Average (shifted) song year : 71

2.3

2. RMSE of predsNLabelsTrain = 11.669560325232723

RMSE of predsNLabelsVal = 12.137914814121014

RMSE of predsNLabelsTest = 11.384328753882267

Exercise 3

3.3

3.3.2 When we print the per-iteration RMSE we osberve that at the end RMSE becomes infinity.

3.3.3 Gradient Descent does not converge to a definite limit which is what we want.

3.3.4 We observe that by changing the number of iterations nothing happens. When we change the alpha though we can see that the number that GD converges changes. For this reason we have to find the optimal value for learning rate(alpha) that gives us the ideal convergion.

3.3.5

```
Alpha =                     pow(2,-10)

Number Of Iterations  =     50

Weights = DenseVector(1.4789646276184545, 0.09523429960097464, -0.4730839904736139)

Error Train = List(84.13161494880869, 65.82997848016262, 32.14112973135499,

13.55331642190436, 10.640358334879565, 10.333696896931302, 10.107303402634326,

9.900302829972455, 9.709195959890879, 9.531448687949132, 9.365124759992248,

9.208705927701246, 9.060976137805275, 8.920943896430227, 8.787788485254561,

8.660821651243086, 8.539459659976298, 8.42320248322493, 8.311618016381475,

8.20432991725318, 8.101008101016035, 8.001361215926977, 7.905130618259209,

7.812085497324054, 7.722018893559156, 7.634744417842119, 7.550093527020649,

7.467913244775071, 7.388064242115843, 7.310419210631326, 7.234861475806386,

7.1612838085691966, 7.08958740156834, 7.019680983165549, 6.951480047207325,

6.884906180647425, 6.819886474278776, 6.756353004385174, 6.694242375177801,

6.633495313548199, 6.574056309026722, 6.515873292949029, 6.458897351749566,

6.403082470060479, 6.348385299925266, 6.2947649529639165, 6.242182812768793,

6.190602365182642, 6.13998904442545, 6.090310093303891)
```

3.4

RMSE on the validation set : 11.57678596563906

Exercise 4

4.1

1. Coefficient:

[0.5185478769737881,-0.02355890134343917,-0.06455305041162726,0.04488673027312045,0.028313372625239055,-0.13488020628395153,-0.0037337124455512232,-0.06924545105233404,-0.12193580923157364,0.1608523869863647,-0.1693450438559648,-0.027253143493580705]

Intercept: 48.651880748062226

2. RMSE on the validation set: 11.349038490129756

3. First 10 predictions:

```
+------------------+-----+--------------------+
|        prediction|label|            features|
+------------------+-----+--------------------+
|  68.7446694125127| 56.0|[45.800256,41.148...|
| 67.49803651908191| 58.0|[40.643545,6.2819...|
|63.007029226025836| 38.0|[44.763875,-17.45...|
| 72.04278408592101| 74.0|[43.174602,-1.595...|
| 68.15738234120599| 45.0|[44.789073,17.241...|
|  72.3372070959877| 83.0|[47.925393,-44.79...|
| 67.06241015658601| 75.0|[40.17081,0.65690...|
| 72.48498738599056| 83.0|[49.011671,17.117...|
| 76.87871149137803| 83.0|[43.366126,44.248...|
| 77.53439630719768| 83.0|[52.450455,59.964...|
+------------------+-----+--------------------+
```

4.2

1. RMSE :

RegParam = 1	:	11.349038490129756	
RegParam = 1e-5	:	11.314980374303529	
RegParam = 1e-10	:	11.314980241852886	***BEST***

2. The Regularization parameter that achieves that is the 1e-10. The lowest RegParam the better. At some point though there is no need to lower the RegParam more because the difference is not that significant.

Exercise 5

5.3

RMSE of the new model : 11.032470479816805

5.4

1.

RMSE of the baseline model : 12.137914814121014

RMSE of the new model : 11.032470479816805

2. First 50 predictions:

```
+-----------------+
|       prediction|
+-----------------+
|73.90750022208775| | 75.4743998332533|  | 69.5688071977952|  | 71.1632111291855|
|71.46352698688945| |73.72481558207596| |69.36451437183086| |66.50192771840317|
|66.25198532807923| |69.39802209318535| |73.21129760302085| | 70.9639326888076|
|73.31405370394323| | 74.3036811874539|  |66.02819621038681| |73.16035609425168|
|67.30092053763306| |75.60582137918132| |74.78620499164528| |66.80164921817921|
|62.00308189364782| |69.21776207629159| |66.88110798638577| | 70.8585936015567|
| 69.9909612693296|  |77.72993415973092| |72.88105166317642| |  76.018429399707|
| 75.8217635781519|  |74.65823291045058| | 69.0155225063675|  | 75.6192867999151|
|73.97660986146316| |72.78257496262844| |76.96380892313488| |72.83641655006004|
|66.77558120285161| | 74.5633134137164|  |73.60223623550091| |73.65943599886766|
|67.08360184531722| |74.39232189775535| |77.94501542998307| |76.74777022513643|
| 72.1075532855313|  |74.54728506624203| |72.63071330404841| |69.98932429691669|
|65.27400208783584| | 65.7200083941045|
+-----------------+
```

5.5

RMSE of the test set: 10.488599404012978