COMP-551: Applied Machine Learning - Assignment #1

Antonios Valkanas

Student ID: 260672034

Question 1

1.1 Since we can only sample the uniform distribution [0,1] the proposed algorithm takes advantage of the fact that that sum of the probabilities for all the activities is 1 to construct a cumulative probability function. Here is the pseudocode:

```
BEGIN
sample = uniform(0,1)
if sample < 0.2
   then str = 'movies'
if sample < 0.6
   then str = 'COMP-551'
if sample < 0.7
   then str = 'playing'
else
   str = 'studying'
return str
END
```

1.2 By running the implemented pseudocode above multiple times, we sample the distribution for 100 and 1000 days and get the following results.

```
Breakdown of student activities for 100 days.
Movies for 18 days.
COMP-551 for 39 days.
Playing for 15 days.
Studying for 28 days.
```

```
Breakdown of student activities for 1000 days.
Movies for 187 days.
COMP-551 for 379 days.
Playing for 100 days.
Studying for 334 days.
```

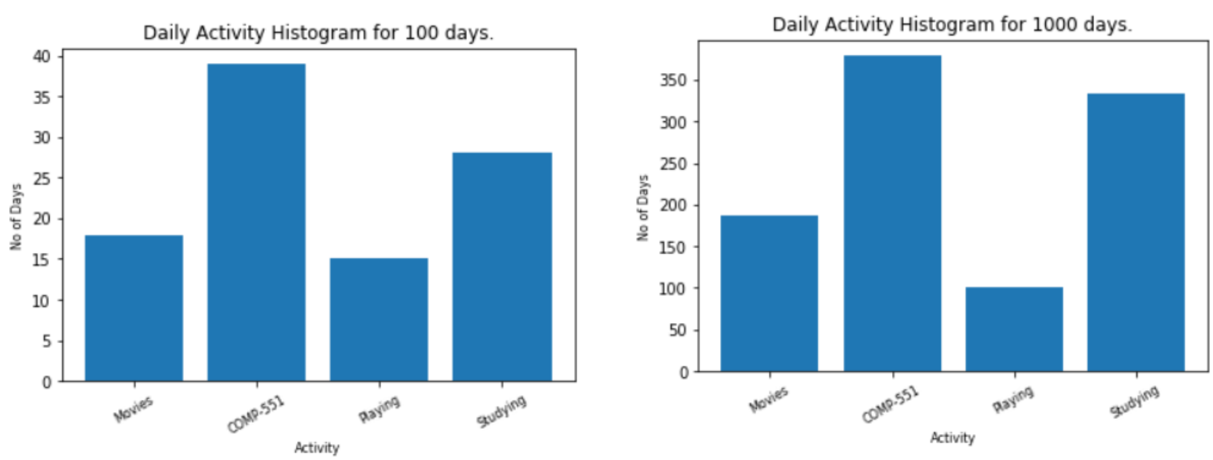


Figure 1: Daily activity sampling for 100 and 1000 days.

We notice that as the sample size increases, as dictated by the law of large numbers, the observed outputs of the sample approach the expectation value of the multinomial distribution.

Table 1: Expectation values and observed outcomes from sampling.

Expected	100 days	1000 days
movies: 0.2 COMP-511: 0.4 playing: 0.1 studying: 0.3	movies: 0.18 COMP-511: 0.39 playing: 0.15 studying: 0.28	movies: 0.187 COMP-511: 0.379 playing: 0.100 studying: 0.334

Question 2

2.1 After training without regularization we get the following results:
Training MSE: 6.474757468571469
Validation MSE: 1422.204539670829

2.1 b) Using matplotlib we can visualize the fit of the validation data. The training data points are also shown in the plot but they are not fitted (hence the trendline ignores some blue points).

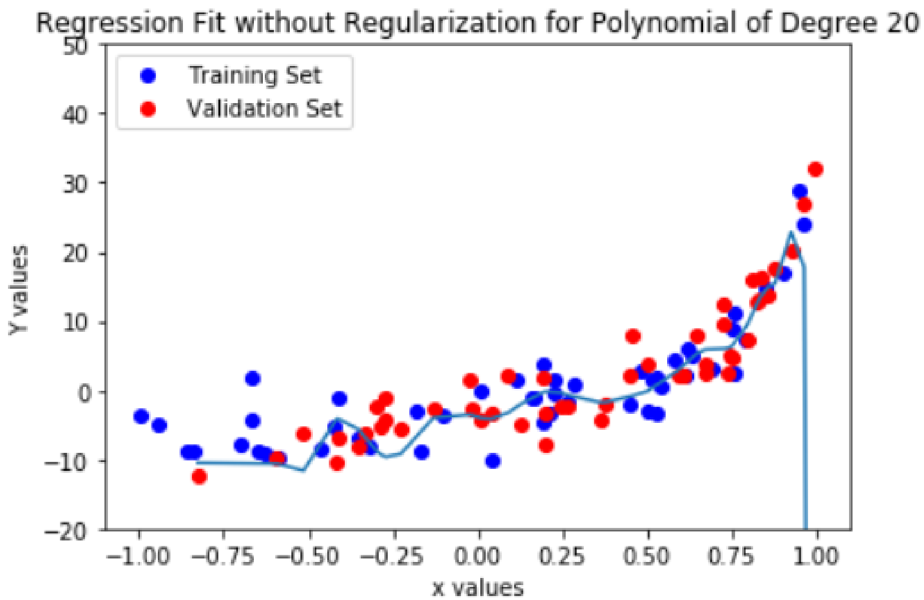


Figure 2: Visualization of the validation data using simple regression.

2.1 c) The model seems to be highly overfitting. This can be verified both visually from the plot, but also from the very large difference between the reported training MSE and the reported validation MSE.

2.2 a) Varying the lambda we can find the best L2 regularization coefficient for the problem.

Figure 3: Validation MSE vs λ.

2.2 b) From figure 3 we can find the minimum validation MSE using the numpy library which turns out to be 9.13 for lambda equal to 0.0197. Using the model for this value and fitting the test set we get the following plot.

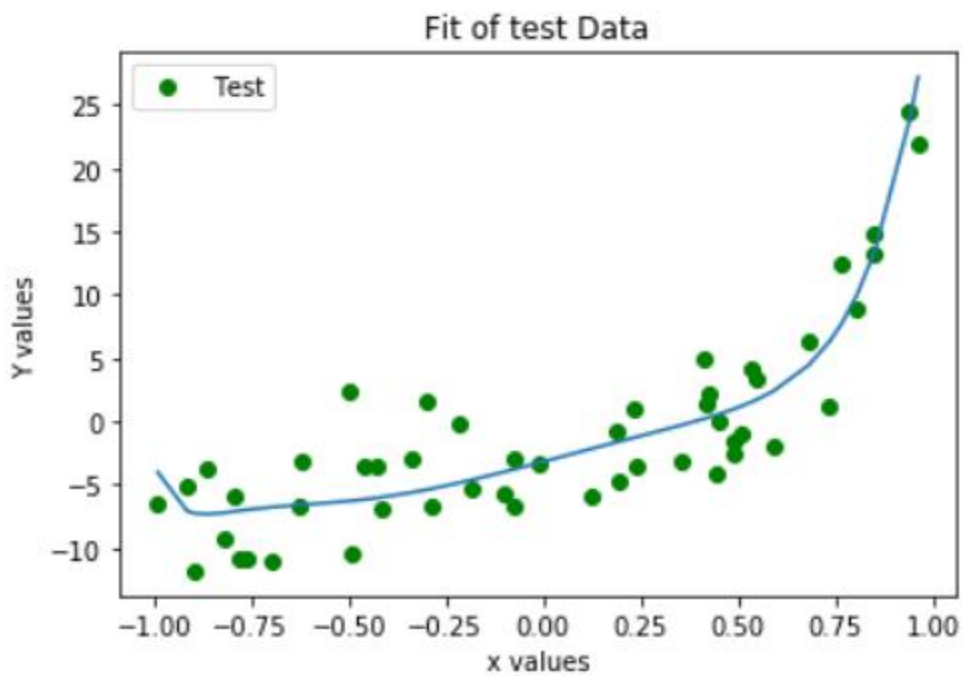


Figure 4: Test data fit using L2 regularization.

2.2 c) Visualizing the dataset using the new model we get the following plot.

Best fit of the data using L2 Regularization and the optimal lambda

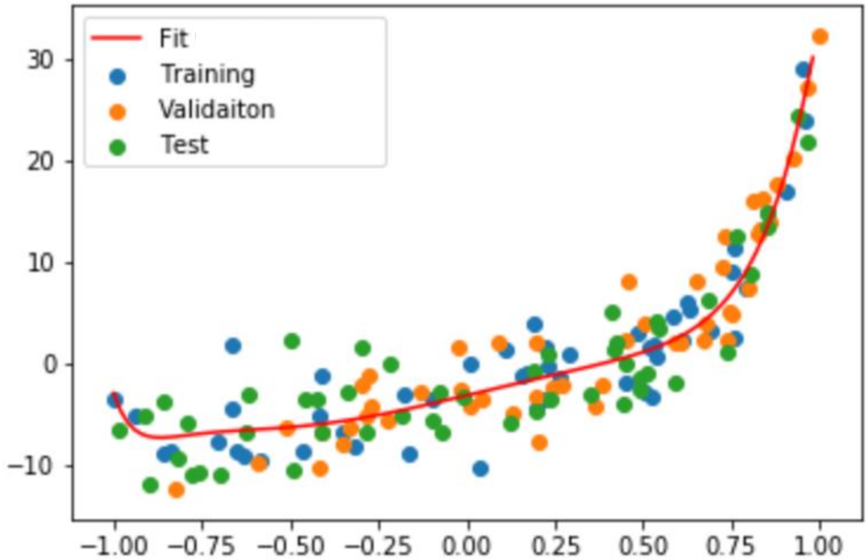


Figure 5: Fit with lambda equal to 0.0197

2.2 d) The model is not overfitting or underfitting anymore. L2 regularization helped improve the validation MSE.

2.3) While it is difficult to judge the exact degree of the polynomial from the plot, it appears to be a 2nd degree polynomial with a positive x^2 coefficient.

Question 3

3.1 a) The MSE is computed for every epoch and the plot of the validation MSE for every epoch is given below.

Epochs until convergence for validation: 5000

Final Validation MSE: 0.2644356376134

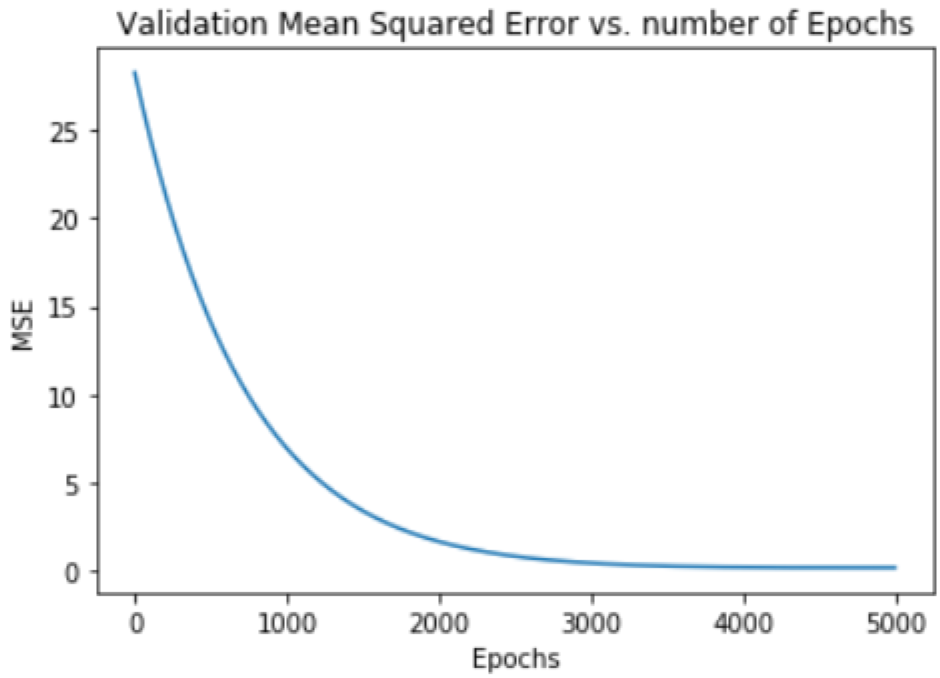


Figure 5: Training, Validation MSE for every epoch trained.

3.1 b) The MSE for validation & trianing is shown above in Figure 5. The two plots overlap.

3.2 a)

Table 2: Learning Rate and performance

Learning Rate	Validation MSE	Number of Epochs
1	DIVERGES TO INFINITY	DIVERGES
1e-1	0.98	5
1e-2	0.08	42
1e-3	0.08	343
1e-4	0.08	2665
1e-5	0.08	4578
1e-6	0.08	7857

3.2 b) Chosen Model: λ = 1e-3 best tradeoff between minimum MSE and convergence. MSE 0.08

3.3) Visualization is in the next page.

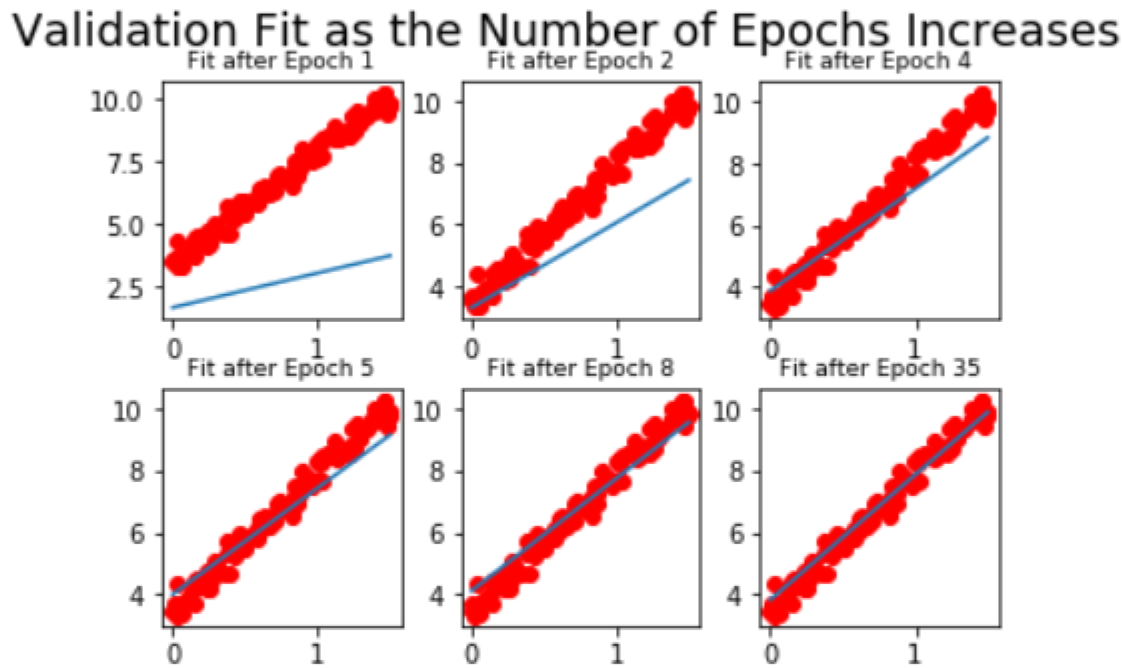


Figure 6: Model evolution

Question 4

4.1 a), b) c) Using the mean makes the dataset susceptible to very large outliers that would bring the mean up and introduce bias to the data. The median is a better option that is immune to large outliers, which therefore improves performance.

4.2 The average 5-fold validation error is: 0.02 The parameters learned are saved as a txt in the code submission.

4.3 a) From the various lambda values, the optimal one is 1.9 with MSE equal to 0.0197.

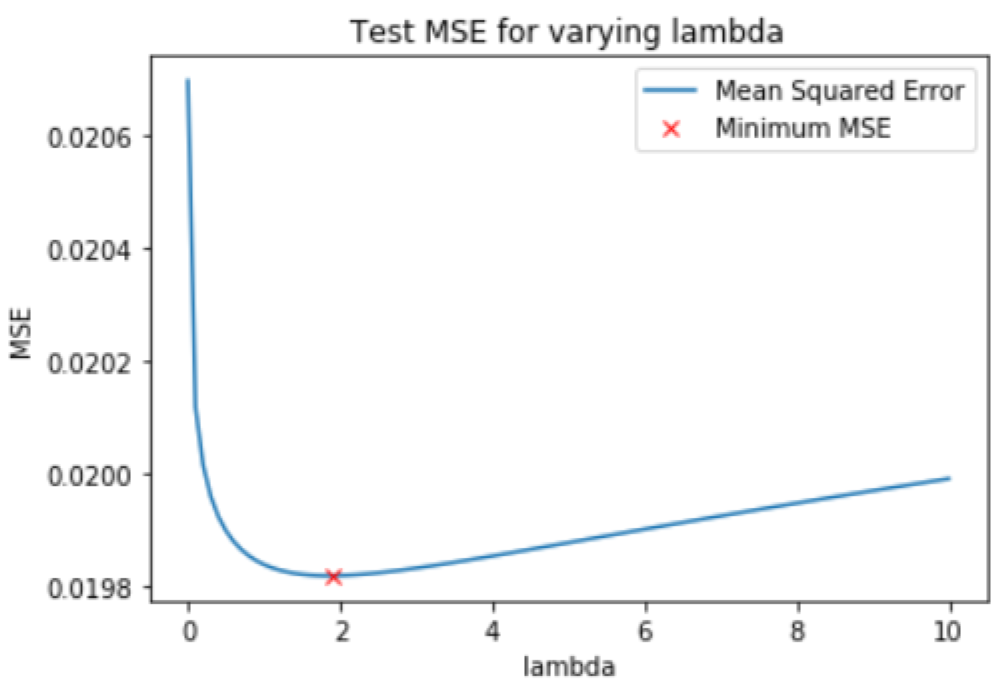
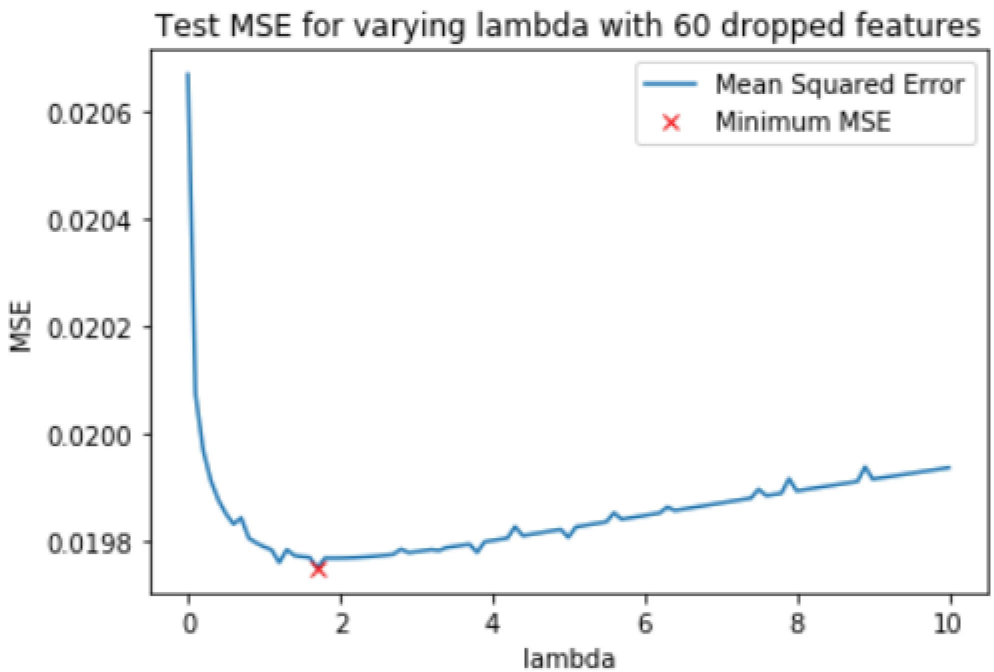


Figure 7: Optimal Lambda

4.3 b) The information of this experiment can be used for feature selection. To do this we sort the weight values for all the features in order of absolute magnitude. We can then proceed to drop the smallest weight since they affect our distribution the least. As a result, we should expect to lose small amount of accuracy but have a much simpler to compute model.

4.3 c) Plotted below is the MSE for a reduced feature L2 regression.



Feature 8: Model performance with 60 dropped features.

4.4) The result of feature selection was that our overall error increased marginally (by less than 2%) while the amount of features we used was half. This means that this is a very successful technique.