COM6012 Assignment 1: Scalable Machine Learning

Question 1

No	Time Slot	Average Number of Requests (2 d.p.)
1	00:00:00-03:59:59	7000.64
2	04:00:00-07:59:59	5428.89
3	08:00:00-11:59:59	14319.43
4	12:00:00-15:59:59	17195.50
5	16:00:00-19:59:59	12956.75
6	20:00:00-23:59:59	9955.5

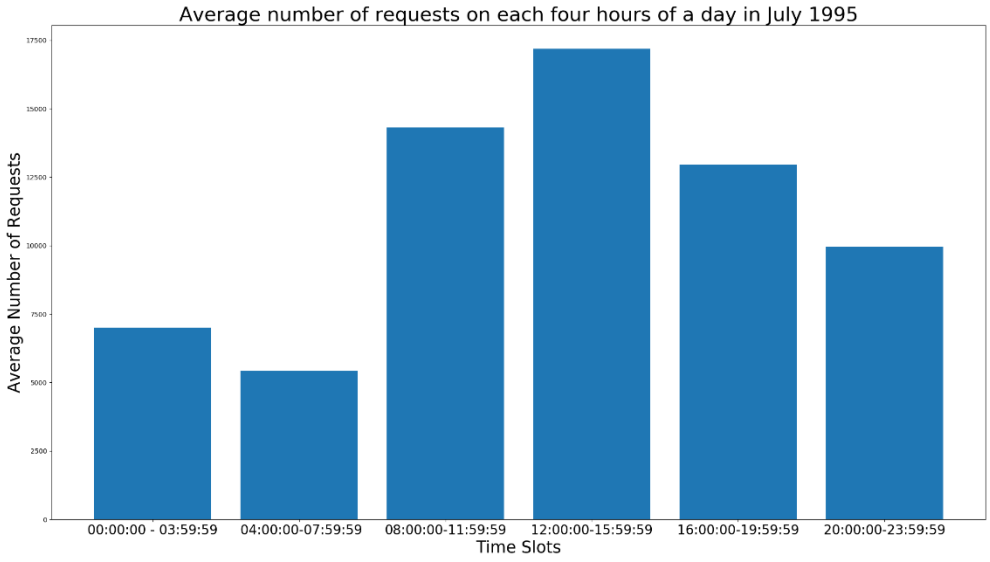

Figure 1 - Bar Chart to show Average Number of Requests for each time slot of a day in July 1995

Observations

It is expected that most of the requests being made occur during the working day i.e. 9 am – 5pm. It is expected that there is more traffic during the working hours of the day than during the usual American sleep hours. The data does follow this trend with 12:00:00 – 15:59:59 being the most popular time slot. Most of the web traffic is going to come from the USA as this is where most of the scientific community reside. Furthermore in 1995, the internet was in its infancy and so traffic from other countries is much less likely.

The data possibly follows a skewed normal distribution with a standard deviation of 4105.563 and a mean of 11,142.786. The skew of the data is 0.001566 which suggests the skew is to the right. However, to fully classify this as a normal distribution, more data points are required. This could be achieved by checking hour by hour instead of 4-hour blocks. The median is 11,456.125 and there is no mode for this dataset.

No	File Name	Number of Requests
1	/ksc.html	40219
2	/shuttle/missions/missions.html	24864
3	/shuttle/countdown/liftoff.html	22000
4	/shuttle/missions/sts-71/mission-sts-71.html	16717
5	/shuttle/missions/sts-70/mission-sts-70.html	16122
6	/shuttle/missions/sts-71/images/images.html	15897
7	/history/apollo/apollo.html	14472
8	/history/apollo/apollo-13/apollo-13.html	13768
9	/history/history.html	11816
10	/shuttle/countdown/countdown.html	8572
11	/shuttle/technology/sts-newsref/stsref-toc.html	7420
12	/software/winvn/winvn.html	6970
13	/shuttle/missions/sts-69/mission-sts-69.html	6967
14	/shuttle/missions/sts-70/images/images.html	6709
15	/shuttle/missions/sts-71/movies/movies.html	6308
16	/shuttle/missions/sts-70/movies/movies.html	6107
17	/history/apollo/apollo-13/apollo-13-info.html	5747
18	/facilities/lc39a.html	5260
19	/history/apollo/apollo-11/apollo-11.html	5004
20	/shuttle/countdown/lps/fr.html	4218

Figure 2 - Pie Chart to show most accessed html files in July 1995

Observations

Ksc.html is accessed almost twice as much as the next most visited page. KSC refers to the Kennedy Space Centre. Ksc.html is the parent homepage, the user probably clicks on the desired subpage from here. There was a launch at the Kennedy Space Centre in July 1995 which could explain some of the web traffic. [1]

A reason why the Apollo pages were visited is because July 20th is the anniversary of the Apollo 11th launch. [2]

The data follows a logarithmic distribution as demonstrated by figure 3. Note that the data has not been shown as a line chart as the file counts are not strictly dependent on the previous file.

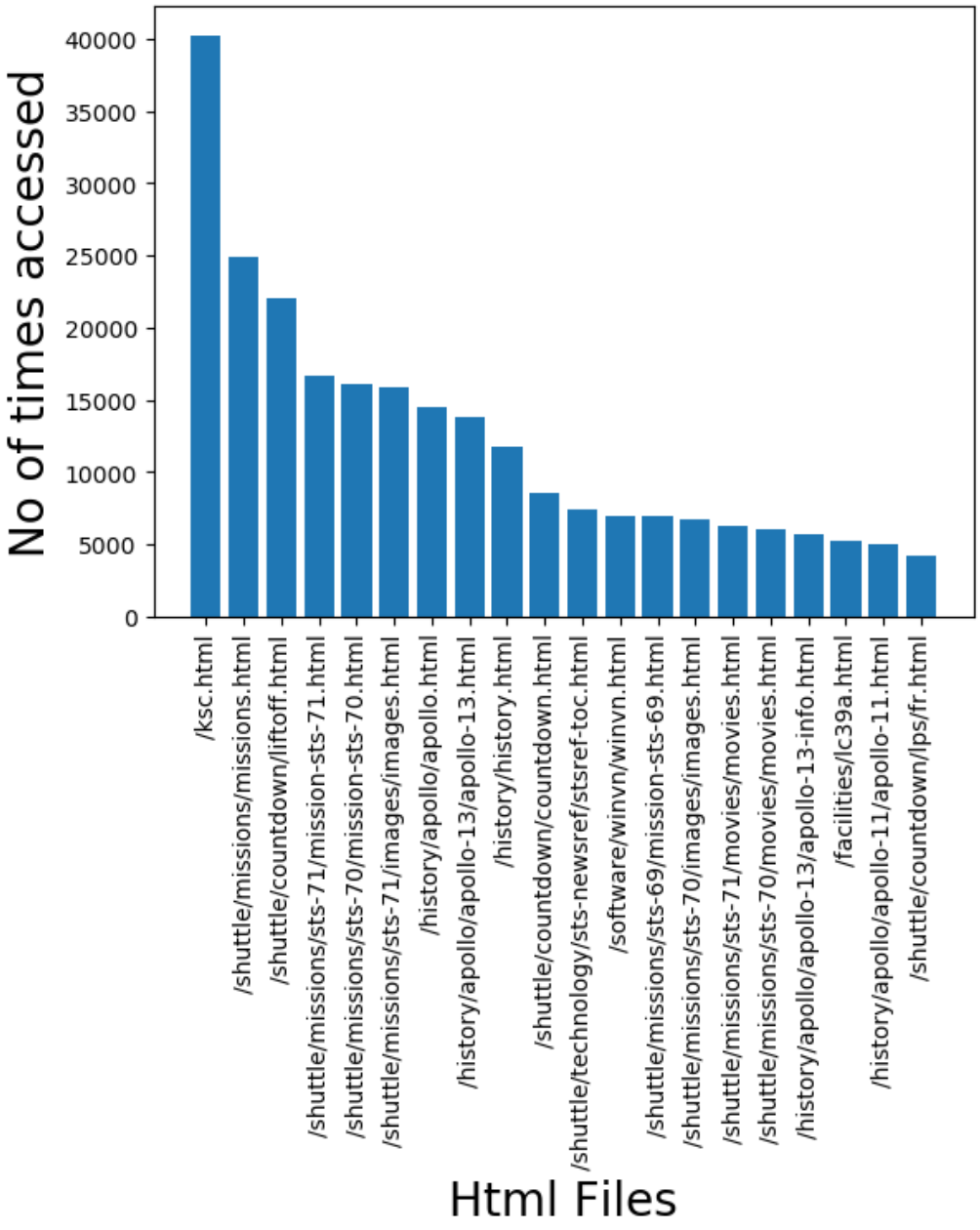

Figure 3 - Graph to show file counts for each html file

Question 2A

Parameters changed

ALS Strategy	Regularization Parameter	Max Number of Iterations	Cold Start Strategy
1	0.1	10	Drop
2	0.1	20	Drop
3	0.01	10	Drop

All other parameters were kept the same.

Results

No	Metric	ALS 1	ALS 2	ALS 3
1	Mean RMSE	0.8117	0.8091	0.8412
2	Mean MAE	0.6269	0.6239	0.6344
3	Standard Deviation of RMSE	0.0001316	0.00004956	0.0008961
4	Standard Deviation of MAE	0.0002335	0.0001443	0.0007281
5	RMSE (fold = 1)	0.8117	0.8416	0.8416
6	RMSE (fold = 2)	0.8118	0.8091	0.8420
7	RMSE (fold = 3)	0.8115	0.8090	0.8399
8	MAE (fold = 1)	0.6269	0.6239	0.6346
9	MAE (fold = 2)	0.6272	0.6240	0.6352
10	MAE (fold = 3)	0.6266	0.6237	0.6334

Note answers given to 4 significant figures.

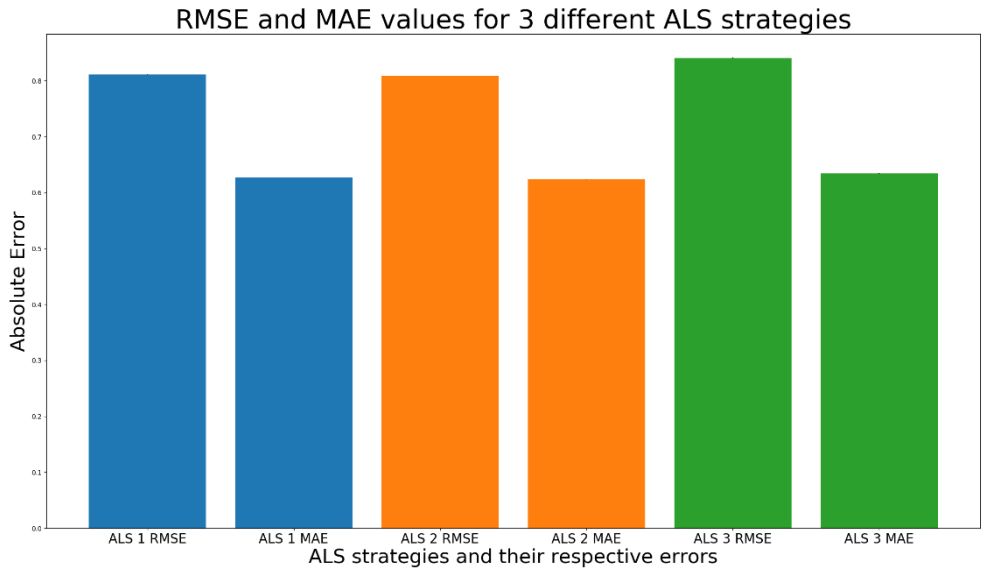

Figure 4 - Bar Chart to show RMSE and MAE values for different ALS Strategies

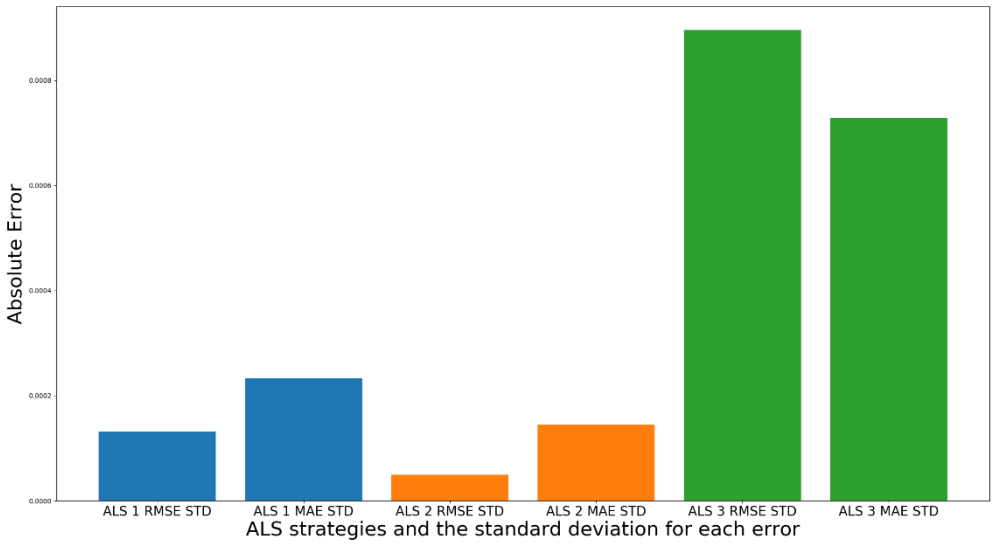

Figure 5 - Bar Chart to show standard deviation of RMSE and MAE for 3 different ALS strategies

Observations

The best overall strategy was ALS 2 using regularization parameter of 0.1 and 20 max iterations. This would make one think that the regularization parameter is less important that the max number of iterations. In order to test this hypothesis another ALS strategy was run using a regularization parameter of 0.01 and max number of iterations of 20.

Interestingly using a regularization parameter of 0.01 and 20 iterations leads to a mean RMSE of 0.8381 and a mean MAE 0.6291. Using a smaller regularization parameter made it a worse performer compared to ALS 2. This could lead to the conclusion that ALS converges quickly and that the error is more dependent on the number of iterations. Typically, ALS would converge ~20 iterations. For the best performance the regularization parameter should decrease monotonically as the iteration progresses. [3] Furthermore, if a repeated random sub-sampling method was used for the cross-validator method instead of k-folds, then the number of iterations would be completely independent of the training and test split. [4]

If the value of k was increased, the model would be far more accurate. This is because the model would be trained on many different subsets. Looking at the data obtained it seems that some of the strategies performed slightly worse in between folds, for example, ALS 3 performed worse between fold 1 and 2 for both the RMSE and the MAE. As a random subset of the data is being taken and only one-fold has occurred, this is perfectly plausible. It is possible that in this split the algorithm chose a split of data containing many anomalies. The only way to reduce the chance of the model performing worse overall is to increase the number of k.

ALS Strategy	Regularization Parameter	Max Number of Iterations	Cold Start Strategy
4	0.01	20	Drop

No	Metric	ALS 4
1	Mean RMSE	0.8381
2	Mean MAE	0.6291
3	Standard Deviation of RMSE	0.0006228
4	Standard Deviation of MAE	0.0004840
5	RMSE (fold = 1)	0.8378
6	RMSE (fold = 2)	0.8390
7	RMSE (fold = 3)	0.8375
8	MAE (fold = 1)	0.6288
9	MAE (fold = 2)	0.6298
10	MAE (fold = 3)	0.6287

Question 2C

The top 3 clusters for ALS 1 are:

Cluster	Rank	Tag Number	Tag	No of Movies with Tag in cluster
1	1	742	Original	5723
1	2	646	Mentor	5723
1	3	270	Criterion	5723
2	1	742	Original	4050
2	2	972	Storytelling	4050
2	3	1104	Weird	4050
3	1	742	Original	3949
3	2	646	Mentor	3949
3	3	468	Great Ending	3949

The top 3 clusters for ALS 2 are:

Cluster	Rank	Tag Number	Tag	No of Movies with Tag in cluster
1	1	742	Original	4630
1	2	972	Storytelling	4630
1	3	646	Mentor	4630
2	1	742	Original	4458
2	2	646	Mentor	4458
2	3	188	Catastrophe	4458
3	1	742	Original	4081
3	2	646	Mentor	4081
3	3	195	Chase	4081

The top 3 clusters for ALS 3 are:

Cluster	Rank	Tag Number	Tag	No of Movies with Tag in cluster
1	1	742	Original	18687
1	2	646	Mentor	18687
1	3	445	Good	18687
2	1	742	Original	3814
2	2	807	Predictable	3814
2	3	646	Mentor	3814
3	1	270	Criterion	3795
3	2	742	Original	3795
3	3	1008	Talky	3795

Using tags.csv to see how many times users rated movies with tags identified by genome-tags

Tag	Movie Count in tags.csv
Original	527
Mentor	68
Criterion	55
Storytelling	63
Weird	287
Great Ending	47
Catastrophe	1
Chase	490
Predictable	775
Talky	95

Observations

Note that every movie has every tag (1128 tags) and hence the number of movies with tag per cluster is the same. [5]

The tag Original appears in every single cluster for every single version of ALS. Looking at movies.csv this makes perfect sense as the movies used for this data are generally older in nature and thus more likely to be an original movie. Furthermore, looking at tags.csv many users have listed many films as original (527 tags containing the word original). There are also a couple of international movies which are more likely to be rated as original for the average movie lens user. Also, the older the movie the more likely it is to have more ratings thus potentially increasing its chance at a high similarity score for each tag.

The tags generally obtained by the clustering algorithm seem to be mostly positive. The exceptions are predictable, talky and weird. However, some users may prefer movies of that nature. There are also neutral tags like criterion. Criterion refers to an American video distribution company who license classic and contemporary films. The large amount of classic and contemporary films could be correlated with the large number of original movies returned. Mentor is also a neutral tag which refers to films which have a character who mentors another character, for example, Kung Fu Panda.

ALS 3 cluster 1 captures 18687 movies. There are 62432 movies in the dataset meaning that cluster 1 accounts for 30% of the movies. Likewise, ALS 1 cluster 1 accounts for 9.2% and ALS 2 cluster 1 accounts for 7.41%. ALS 3 performed better at separating out clusters despite the same clustering algorithm being used for each split. This is likely due to the lower regularization parameter used, thus allowing for a tighter spread of data.

References

[1] Wikipedia, "Kennedy Space Center Launch Complex 39," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Kennedy_Space_Center_Launch_Complex_39. [Accessed 05 03 2020].

[2] NASA, "July 20, 1969: One Giant Leap For Mankind," NASA, 20 07 19. [Online]. Available: https://www.nasa.gov/mission_pages/apollo/apollo11.html. [Accessed 05 03 2020].

[3] Y. Chen, W. Sun, M. Xib and J. Yuan, "A Seminorm Regularized Alternating Least Squares Algorithm For Canonical Tensor Decomposition," *Journal of Computational and Applied Mathematics,* vol. 347, no. 1, pp. 296-313, 2019.

[4] G. Seif, "Why and How to do Cross Validation for Machine Learning," Towards Data Science, 24 05 2019. [Online]. Available: https://towardsdatascience.com/why-and-how-to-do-cross-validation-for-machine-learning-d5bd7e60c189. [Accessed 06 03 2020].

[5] J. Vig, S. Sen and J. Riedl, "The Tag Genome: Encoding Community Knowledge to Support Novel Interaction," *ACM Transactions on Interactive Intelligent Systems,* vol. 2, no. 3, p. 13, 2012.