B555 – Machine Learning

Submitted by : Nikita Bafna

Assignment 3: Generalized linear model using Logistic Regression

Introduction:

In this experiment, we implement Generalized linear model Algorithm using different likelihoods and generate the learning curves. The 3 different likelihood models used are:

1) Logistic model
2) Poisson model
3) Ordinal model

The following datasets (features and labels) were used:

1. For logistic model: A.csv, labels-A.csv

   usps.csv, usps.csv

2. For Poisson model: AP.csv, labels-AP.csv
3. For Ordinal model: AO.csv, labels-AO.csv

While conducting this experiment, we have considered value of Alpha as 0.1. Below algorithm was used:

Repeat 30 times

Step 1) Set aside 1/3 of the total data (randomly selected) to use as a test set.

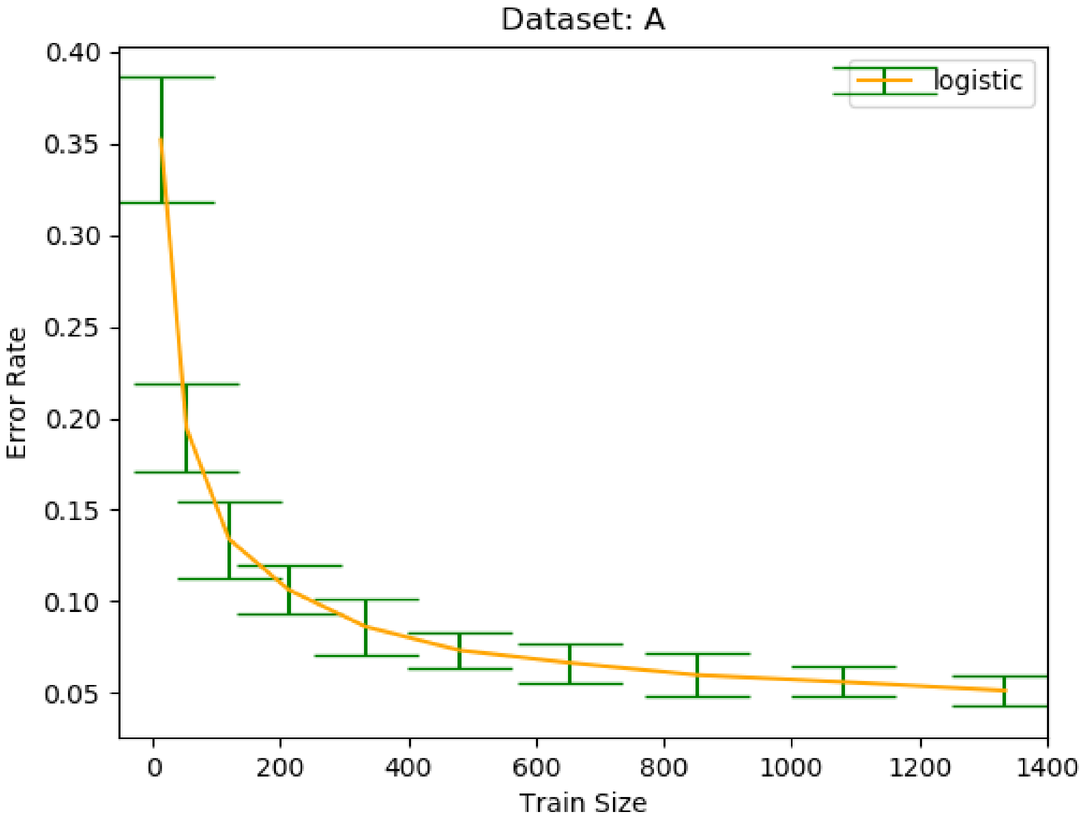
Step 2) Permute the remaining data and record the test set error rate as a function of increasing training set portion (0.1,0.2, ...,1 of the total size).

Results Evaluation:

Statistics and graphs for Logistic model:

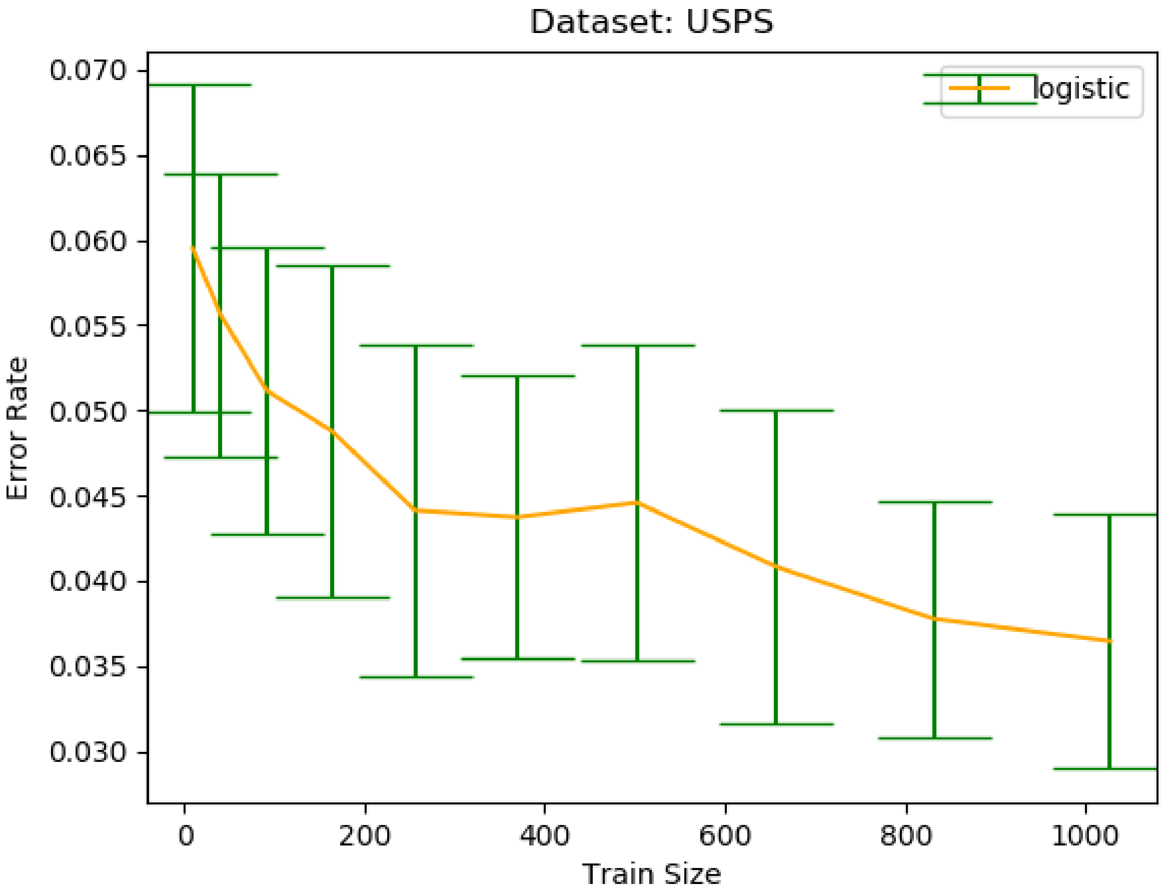
1. Dataset A

Train Size factor	Avg # of Iterations	Runtime in (seconds)
0.1	5.6	0.0121
0.2	5.3	0.0122
0.3	5.2	0.0127
0.4	5.15	0.0136
0.5	5.12	0.0146
0.6	5.1	0.0159
0.7	5.08	0.0183
0.8	5.125	0.0218
0.9	5.22	0.0263
1	5.3	0.0322



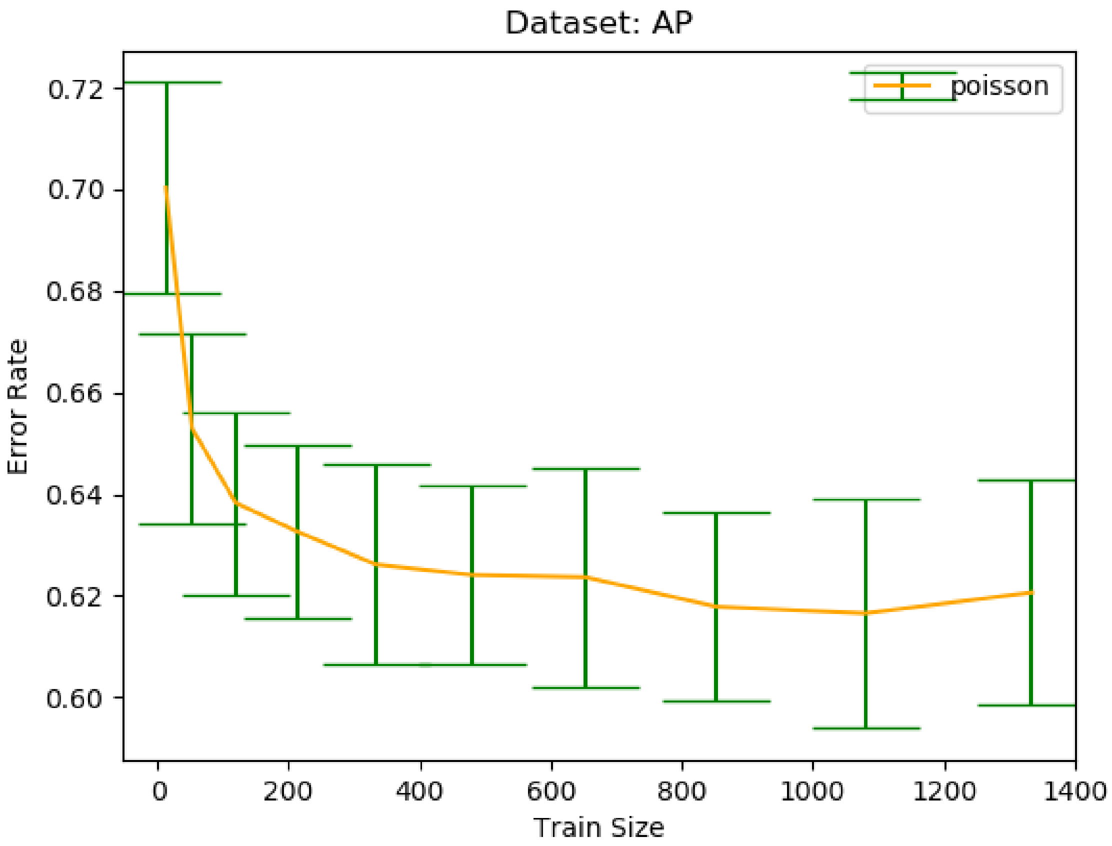
2. Dataset USPS:

Train Size factor	Avg # of Iterations	Runtime in (seconds)
0.1	9.46	0.133
0.2	9.71	0.140
0.3	9.82	0.147
0.4	9.91	0.155
0.5	9.96	0.164
0.6	10	0.179
0.7	10.03	0.191
0.8	10.058	0.202
0.9	10.059	0.213
1	10.06	0.225



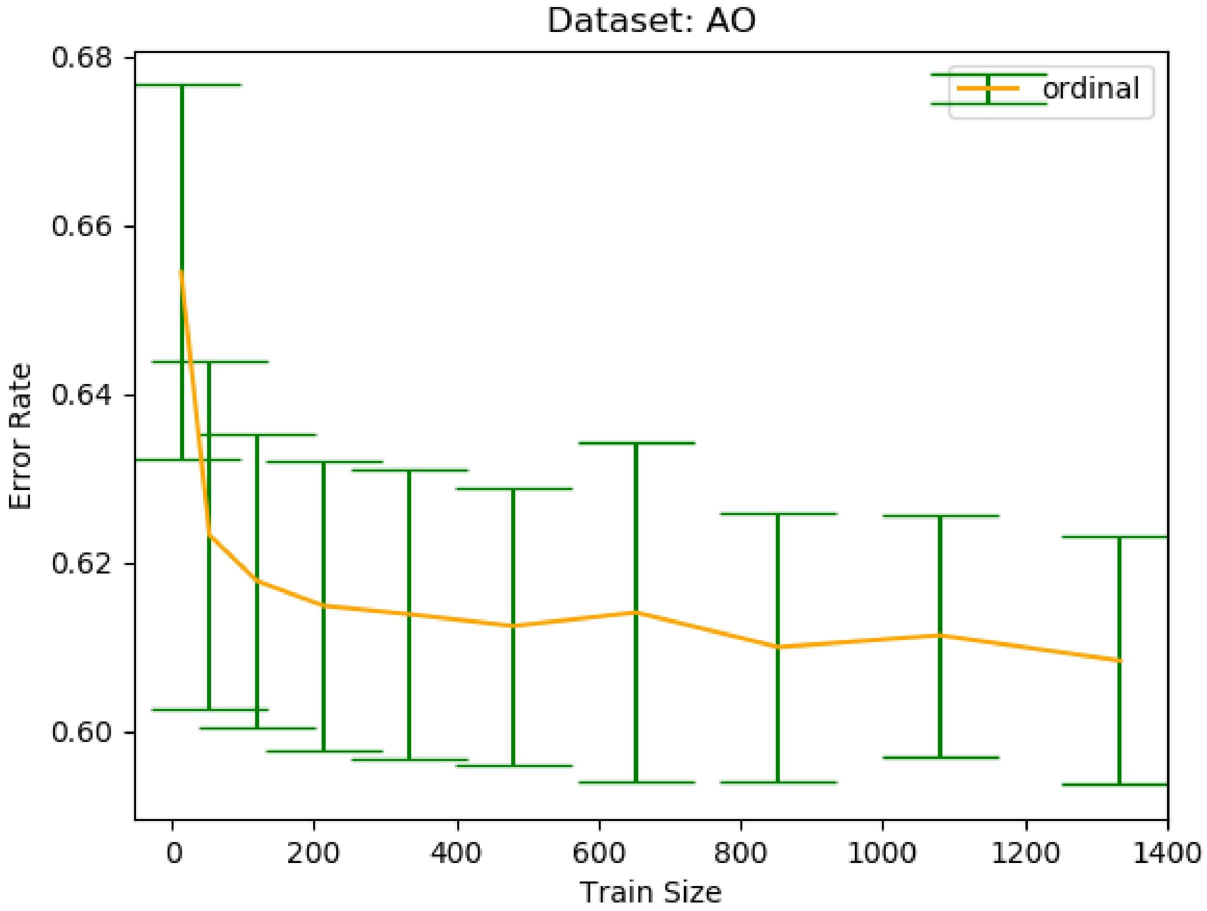
3. Dataset AP:

Train Size factor	Avg # of Iterations	Runtime in (seconds)
0.1	11.06	0.030
0.2	10.38	0.031
0.3	9.92	0.0312
0.4	9.54	0.032
0.5	9.30	0.035
0.6	9.12	0.038
0.7	8.99	0.041
0.8	8.86	0.045
0.9	8.76	0.050
1	8.68	0.055



4. Dataset AO:

Train Size factor	Avg # of Iterations	Runtime in (seconds)
0.1	6.8	0.030
0.2	6.833	0.031
0.3	6.855	0.0312
0.4	6.875	0.032
0.5	6.893	0.035
0.6	6.9	0.038
0.7	6.914	0.041
0.8	6.920	0.045
0.9	6.929	0.050
1	6.936	0.055



Observations :

1. As a general trend across all the dataset we observe that as the training set size increases the testing accuracy (less error) increases also the standard deviation shown by the error bar decreases.
2. We can see that the dataset A converges in less number of iterations compared to others.
3. The run time increases as the training size increases.
4. Logistic model takes comparatively less time compared to other models.
5. As the training size increases, logistic model captures the best performance giving least error.
6. The standard deviation and error rate are captured in the graphs.
7. For the logistic model, for Dataset A and USPS, dataset A took less time compared to USPS as it had less number of features.
8. Ordinal model takes more time to complete the entire algorithm compared to other models.

EXTRA CREDITS:

I think model selection can be implemented using Stratified cross validation with 10 folds in the similar fashion we did in Bayesian linear regression. I didn’t have enough time to compute results for the extra credits.