Programming Project – 3

Experiments with Classification Algorithms

Harika Kanakam

Algorithm 1 – 2 class Generative Model:

- Train dataset is considered as the 2/3rd part of the entire dataset and Model is trained 30 times for each training size.
- Weights are calculated for each training size by initializing alpha to 0.1 and w to a 0 vector.
- The first class is the labels that are 1s and the 2nd class is the labels that are 0s.
- Mean of both the classes are calculated and the S is calculated using S1 and S2.
- Weights W and Wo are calculated using mean and S.
- Once the weights are calculated, the error on test dataset is calculated for the remaining 1/3rd datasets 30 times of each train size.
- For each train size, the mean and standard deviations of 30 runs are calculated

Algorithm 2 – Bayesian Logistic regression:

- Train dataset is considered as the 2/3rd part of the entire dataset and Model is trained 30 times for each training size.
- Weights are calculated for each training size by initializing Sn to (1/alpha)I, alpha to 0.1 and w to a 0 vector.
- Wn+1 is calculated using the below formula:

$$w_{n+1} \leftarrow w_n - (\alpha I + \Phi^T R \Phi)^{-1}[\Phi^T (y - t) + \alpha w_n]$$

- To consider the final w, the stopping criteria is ||Wn+1- Wn||^2/ ||Wn||^2 < 0.001 or n>=100.
- Once the weights are calculated, the error on test dataset is calculated for the remaining 1/3rd datasets 30 times of each train size.
- For each train size, the mean and standard deviations of 30 runs are calculated

Algorithm 3 – Gradient Ascent:

- Train dataset is considered as the first 2/3rd part of the entire dataset and Model is trained 3 times.
- Weights are calculated by initializing alpha to 0.1 and eta to 0.001 and using the below formula:

$$w_{n+1} \leftarrow w_n - \eta[\Phi^T (y - t) + \alpha w_n]$$

- The stopping criteria for w is ||Wn+1- Wn||^2/ ||Wn||^2 < 0.001 or n>=6000.
- Time taken is calculated and w is stored for each 10 iterations of next step of W.
- Using the calculated time taken and W of train data, error on test data is calculated.

Task 1: Generative Vs Discriminative:

- Using the above two algorithms (1 and 2), the A, B and USPS datasets are trained with training fractions, f ∈ {0.1, 0.2, 0.3, . . ., 1.0} of the dataset size.
- After training the algorithm using train datasets with different sizes, error on test dataset is calculated for all the A, B and USPS datasets.

Task 1.1:

- Part 1 of the task is to run the 2 class Generative model for the datasets – A, B and usps.
- The model is trained using 2/3rd of the datasets and tested on the remaining 1/3rd of the datasets.
- The summary of the mean and standard deviation of the error rates for generative model is as below:

**********Error rates for the A Dataset**********

	Train Size	Mean Error Rate	SD error rate
0	0.1	0.251874	0.000000e+00
1	0.2	0.167916	0.000000e+00
2	0.3	0.163418	5.551115e-17
3	0.4	0.130435	2.775558e-17
4	0.5	0.091454	2.775558e-17
5	0.6	0.100450	0.000000e+00
6	0.7	0.103448	0.000000e+00
7	0.8	0.104948	4.163336e-17
8	0.9	0.101949	4.163336e-17
9	1.0	0.097451	1.387779e-17

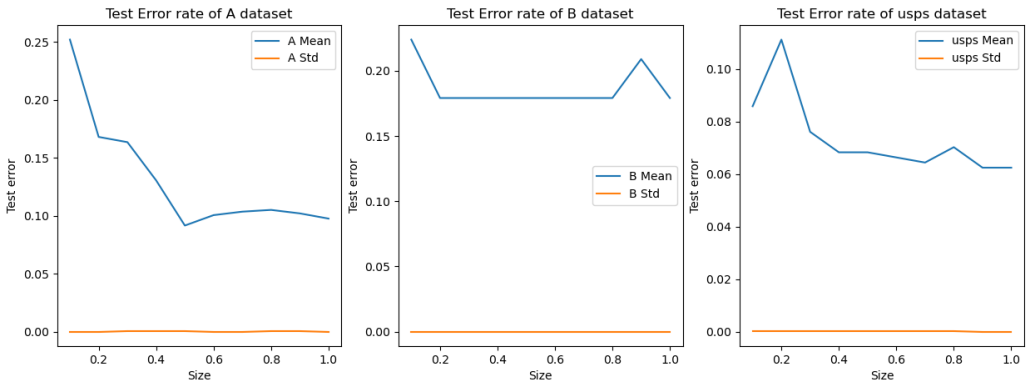
**********Error rates for the B Dataset**********

	Train Size	Mean Error Rate	SD error rate
0	0.1	0.223881	8.326673e-17
1	0.2	0.179104	5.551115e-17
2	0.3	0.179104	5.551115e-17
3	0.4	0.179104	5.551115e-17
4	0.5	0.179104	5.551115e-17
5	0.6	0.179104	5.551115e-17
6	0.7	0.179104	5.551115e-17
7	0.8	0.179104	5.551115e-17
8	0.9	0.208955	5.551115e-17
9	1.0	0.179104	5.551115e-17

**********Error rates for the USPS Dataset**********

	Train Size	Mean Error Rate	SD error rate
0	0.1	0.085770	4.163336e-17
1	0.2	0.111111	1.387779e-17
2	0.3	0.076023	1.387779e-17
3	0.4	0.068226	1.387779e-17
4	0.5	0.068226	1.387779e-17
5	0.6	0.066277	2.775558e-17
6	0.7	0.064327	1.387779e-17
7	0.8	0.070175	1.387779e-17
8	0.9	0.062378	0.000000e+00
9	1.0	0.062378	0.000000e+00

The plots of the error rates for generative model are as below:



- For all the datasets, it is observed for almost all the training datasets as the training size increases, the mean and standard deviation of the test error is decreasing.
- As the Standard deviation values are small the decline in standard deviation is clearly not visible.

Task 1.2:

- Part 2 of the task is to run the Bayesian Linear Regression for the datasets – A, B and usps.
- The model is trained using 2/3rd of the datasets and tested on the remaining 1/3rd of the datasets.
- The summary of the mean and standard deviation of the error rates for Bayesian linear regression is as below:

**********Error rates for the A Dataset**********

	Train Size	Mean Error Rate	Std Error Rates
0	0.1	0.349325	1.110223e-16
1	0.2	0.184408	0.000000e+00
2	0.3	0.137931	2.775558e-17
3	0.4	0.098951	4.163336e-17
4	0.5	0.083958	0.000000e+00
5	0.6	0.068966	1.387779e-17
6	0.7	0.061469	0.000000e+00
7	0.8	0.053973	1.387779e-17
8	0.9	0.040480	0.000000e+00
9	1.0	0.046477	1.387779e-17

The summary of dataset B

	Train Size	Mean Error Rate	Std Error Rates
0	0.1	0.492537	0.000000e+00
1	0.2	0.492537	0.000000e+00
2	0.3	0.208955	5.551115e-17
3	0.4	0.194030	2.775558e-17
4	0.5	0.164179	5.551115e-17
5	0.6	0.164179	5.551115e-17
6	0.7	0.194030	2.775558e-17
7	0.8	0.208955	5.551115e-17
8	0.9	0.194030	2.775558e-17
9	1.0	0.194030	2.775558e-17

The Summary of Dataset usps:

	Train Size	Mean Error Rate	Std Error Rates
0	0.1	0.463938	1.665335e-16
1	0.2	0.463938	1.665335e-16
2	0.3	0.136452	2.775558e-17
3	0.4	0.485380	1.110223e-16
4	0.5	0.541910	1.110223e-16
5	0.6	0.532164	2.220446e-16
6	0.7	0.524366	1.110223e-16
7	0.8	0.499025	0.000000e+00
8	0.9	0.491228	5.551115e-17
9	1.0	0.522417	1.110223e-16

The plots of the error rates for Bayesian linear regression are as below:

Summary:

- It is observed from both the graphs that for almost all the training datasets as the training size increases, the mean and standard deviation of the test error is decreasing.
- As the Standard deviation values are small the decline in standard deviation is clearly not visible.

- As weights are calculated for each step and the optimize weights are considered for Bayesian linear regression, the error rate is less when compared to the generative model.
- Though the error rate for Bayesian linear regression is more when the train size is at the minimal when compared to the generative model, Bayesian linear regression performs better when there is decent amount of training data to train the model.
- As the data in dataset A is uniformly distributed, we could see a gradual decrease in error rate as the training data increase.
- But for the datasets B(for which class is generated from multiple gaussians with differing covariance structure) and dataset usps(which represent 16×16 bitmaps of the characters 3 and 5) which are not uniformly distributed the graph is not gradually decreasing or increasing and there is no uniformity in the graph structure.

Task 2: Newton's Method Vs Gradient Ascent:

- Using the above two algorithms (2 and 3), the A and USPS datasets are trained with $2/3^{rd}$ of the dataset size.
- After training the algorithm using train datasets with different sizes, error on test dataset is calculated for the A and USPS datasets.
- Part 1 of the task is to run the 2 class Generative model for the datasets – A, B and usps.
- The model is trained using $2/3^{rd}$ of the datasets and tested on the remaining $1/3^{rd}$ of the datasets.
- Part 2 of the task is to run the Bayesian Linear Regression for the datasets – A, B and usps.
- The model is trained using $2/3^{rd}$ of the datasets and tested on the remaining $1/3^{rd}$ of the datasets.

The summary of the mean and standard deviation of the error rates for Newton and Gradient Ascent is as below:

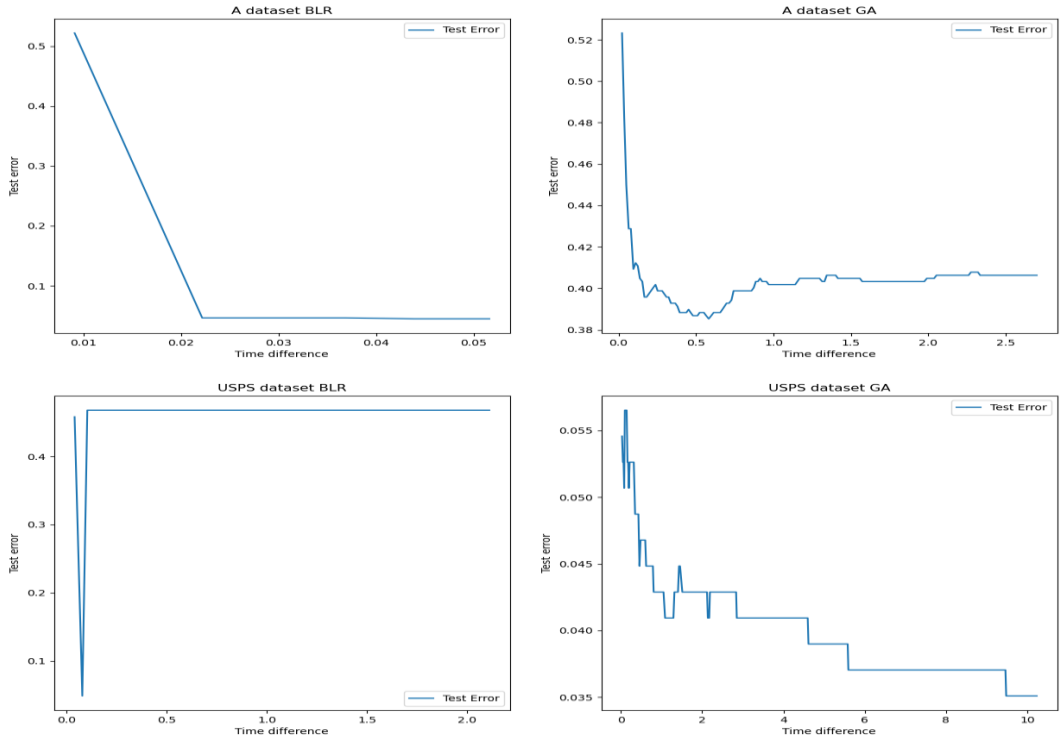
```
**************** Gradient Ascent Summary for dataset A *******
     Avg Time Diff  Error Rates
0         0.017537     0.523238
1         0.028088     0.482759
2         0.038006     0.449775
3         0.048212     0.428786
4         0.058097     0.428786
..             ...          ...
177       1.945876     0.406297
178       1.956537     0.406297
179       1.967327     0.406297
180       1.977601     0.406297
181       1.988227     0.406297

[182 rows x 2 columns]
**************** BLR Summary for dataset A ******************
   Avg Time Diff  Error Rates
0       0.007756     0.521739
1       0.014994     0.046477
2       0.021334     0.046477
3       0.026891     0.046477
4       0.032770     0.044978
5       0.037673     0.044978

**************** Gradient Ascent Summary for dataset USPS ***
    Avg Time Diff  Error Rates
0         0.012202     0.054581
1         0.023150     0.052632
2         0.033188     0.052632
3         0.043140     0.050682
4         0.053856     0.056530
..             ...          ...
497       5.301111     0.035088
498       5.311325     0.035088
499       5.322124     0.035088
500       5.332165     0.035088
501       5.342654     0.035088

**************** BLR Summary for dataset USPS ******
    Avg Time Diff  Error Rates
0         0.014229     0.458090
1         0.028176     0.048733
2         0.040092     0.467836
3         0.052246     0.467836
4         0.066002     0.467836
..             ...          ...
95        1.154440     0.467836
96        1.167527     0.467836
97        1.179567     0.467836
98        1.191250     0.467836
99        1.204129     0.467836
```

The plots for the A and usps datasets for Newtons methos and gradient ascent against time difference and the error rate is as below:



Summary:

- During the implementation value of w as well as the wall clock time after each update in Newton's method and every 10 iterations for gradient ascent are stored.
- For the Newton's methos, the computational cost is more as the computational cost for the inverse of hessian matrix is more.
- But for newtons method, the required weight is calculated fast in lesser steps whereas in gradient ascent the time to calculate the weight is less but the number steps taken is more when compared to newton's methos.
- In both the methods, as the time increases, the training on the model increases and hence the error rate is decreasing.
- Though there is decrease in error rate for usps in newton's method, as an overview there is decrease in error rate with increase in time.
- For the A dataset, in both newton's method and gradient ascent, there is a decrease in the error rate as the time increase.
- From the graph, it is clear that, for the dataset A there is gradual decrease in both the methods as it is uniformly distributed but for usps dataset there are to-and-fros in the graph for both the methods as data represent 16×16 bitmaps of the characters 3 and 5