Group20 Project1 Report

Diyang Zhang, Yuhang Zhang, Olivia Xu

February 11, 2020

1.Abstract

In project 1, we implemented the Logistic Regression(LR) and Naive Bayes classifiers(NB) on 4 datasets: ionosphere, adult, breast-cancer, and bank respectively. By analyzing their test performances, we found that the LR model achieved better accuracy and was faster to train in comparison to the NB model. Besides, we utilized both analytical and gradient descent(GD) approaches to realize the LR model, and found the closed-form analytical approach resulted in worse accuracy than GD.

2.Introduction

To begin with, we selected datasets breast cancer and bank marketing dataset in addition to the mandatory datasets.

Starting from examining our datasets, we removed potentially existing mal-features and mal-samples. Furthermore, we divided each dataset into training, validation, and testing set, sharing 80%,10%, and 10% of data samples severally.

Moreover, we constructed LR and NB models and tested their performance with both training and validation datasets. Last but not the least, we recorded the performance information and studied for their peculiarities and generalities. We explored the convergence speed of the LR-GD is negatively correlated to the learning rate. Besides, LR had higher accuracy on datasets ionosphere, but they experienced a comparable accuracy level on the other three datasets.

The k-cross validation accuracy level of either LR or NB is approximately 75 % on all datasets. Furthermore, both LG and NB achieved more than 88% accuracy on the dataset bank.

2.1 Logistic Regression

In our logistic regression model, we used optimization to find the optimal weight w for features x.

$$f_w(x) = \sigma(w^T x) = \frac{1}{1 + e^{-w^T x}}$$

We used threshold 0.5 for $f_w(x)$ to determine whether to output 1 or 0.
Two methods of optimizing the weight were employed.
[A] Full batch and stochastic gradient descent approximated the optimal weight using an iterative algorithm.

$$\nabla J(w) = \frac{\partial}{\partial w_1} J(w) + \frac{\partial}{\partial w_2} J(w) + ... + \frac{\partial}{\partial w_D} J(w)$$

We started from some w^0 and updated w. α is the step size.

$$w^{t+1} = w^t - \alpha \nabla J(w^t)$$

Gradient descent in general provided more accuracy than naive Bayes. Stochastic gradient descent provided less accuracy than full batch gradient descent on four datasets.
[B] The analytical method performed matrix inverse and multiplication for the algebraically optimal weight w.

$$w = (X^T X^{-1})^{-1} X^T y$$

where X N(size of training data)×D(number of features) and y is N × 1.

2.2 Naive Bayes

In the naive Bayes classifier, we trained our model to learn prior class probability and likelyhood in order to predict a posterior likelihood p($c|x$).
c stands for class, which is binary in all our four datasets. x represents matrix of features.

$$p(c|x) = \frac{p(c) \times p(x|c)}{p(x)}$$

Assuming that all features are conditionally independent.

$$p(x|c) = p(x1|c) \times p(x2|c) \times ...p(x_D|c)$$

If $p(c_0|x) > \frac{1}{2}$, then the result of prediction is 0. Else , output 1.
We used multinomial likelihood for categorical features and gaussian likelihood for numerical features. We found that the running time of naive Bayes is longer than logistic regression on large dataset.

3.Datasets

For all of the four datasets,we deleted data with empty features. We removed problematic and useless features, eg. where all values were identical. In order for the data to fit into our classification model , we distinguished categorical string features from numerical features and encoded them into integer values. We also randomly shuffled the rows to get more objective k-fold-cross-validation-accuracy.

[1]Source: Ionosphere
Goal : predict whether a radar returned from ionosphere is 'good' or 'bad'.

good	bad	total	features	removed fearures
0.64	0.36	350	34	1

note: Total is the total after removing bad samples.Feature is the number of features in the original data set.

[2]Source: Adult
Goal : predict whether income exceeds $50K/yr$ based on census data.

>=50k	<50K	total	features	removed
0.24	0.76	32560	14	0

[3]Source: Cancer
Goal : predict whether a patient has breast cancer.

yes	no	total	features	removed fearures
0.12	0.88	4520	16	0

[4]Source: Bank
Goal : predict whether a client subscribed a term deposit.

yes	no	total	features	removed fearures
0.24	0.76	285	9	0

4.Results

4.1 Comparison between Naive Bayes and Logistic Regression

LG performed better than NB on ionosphere and Adult, but NB performed better on Bank and Cancer. In general,the accuracy of LG and NB is on the same level.

Ionosphere Training Accuracy

Test No.	Logistic Regression,k = 5	Naive Bayes, k = 5
1	0.75905109	0.737548777
2	0.759870144	0.700728398
3	0.758162554	0.763729377
4	0.759802554	0.777893217
5	0.759392329	0.757548876
Avg	0.759255734	0.747489729
Stdev	0.000694999	0.02989452

Ionosphere ValidationAccuracy

Test No.	Logistic Regression, k = 5	Naive Bayes, k = 5
1	0.760177458	0.750238824
2	0.759017243	0.753283479
3	0.750989911	0.743284873
4	0.759631517	0.748972379
5	0.759017288	0.755423875
Avg	0.757766683	0.750240686
Stdev	0.003819097	0.004639935

Ionosphere Training Accuracy

Test No.	Logistic Regression, k = 1	Naive Bayes, k = 10
1	0.759154374	0.747238485
2	0.758846393	0.769812399
3	0.759666494	0.721979823
4	0.759736091	0.757948321
5	0.759325197	0.727320033
Avg	0.75934571	0.744859812
Stdev	0.000367983	0.020191189

Ionosphere Validation Accuracy

Test No.	Logistic Regression, k = 1	Naive Bayes, k = 10
1	0.759563202	0.759563202
2	0.759017231	0.759017231
3	0.758744246	0.758744246
4	0.758914865	0.758914865
5	0.76154243	0.76154243
Avg	0.759556395	0.759556395
Stdev	0.001151762	0.001151762

Adult Training Accuracy

Test No.	Logistic Regression,k = 5	Naive Bayes, k = 5
1	0.75905109	0.737548777
2	0.759870144	0.700728398
3	0.758162554	0.763729377
4	0.759802554	0.777893217
5	0.759392329	0.757548876
Avg	0.759255734	0.747489729
Stdev	0.000694999	0.02989452

Adult Validation Accuracy

Test No.	Logistic Regression, k = 5	Naive Bayes, k = 5
1	0.760177458	0.750238824
2	0.759017243	0.753283479
3	0.750989911	0.743284873
4	0.759631517	0.748972379
5	0.759017288	0.755423875
Avg	0.757766683	0.750240686
Stdev	0.003819097	0.004639935

Adult Training Accuracy

Test No.	Logistic Regression, k = 1	Naive Bayes, k = 10
1	0.759154374	0.747238485
2	0.758846393	0.769812399
3	0.759666494	0.721979823
4	0.759736091	0.757948321
5	0.759325197	0.727320033
Avg	0.75934571	0.744859812
Stdev	0.000367983	0.020191189

Adult Validation Accuracy

Test No.	Logistic Regression, k = 1	Naive Bayes, k = 10
1	0.759563202	0.759563202
2	0.759017231	0.759017231
3	0.758744246	0.758744246
4	0.758914865	0.758914865
5	0.76154243	0.76154243
Avg	0.759556395	0.759556395
Stdev	0.001151762	0.001151762

Cancer Training Accuracy

Test No.	Logistic Regression,k = 5	Naive Bayes, k = 5
1	0.754716981	0.715649279
2	0.766777654	0.766481687
3	0.769811321	0.759378468
4	0.765889752	0.762264151
5	0.766777654	0.801479837
Avg	0.764794673	0.761050684
Stdev	0.005826454	0.030535097

Cancer Validation Accuracy

Test No.	Logistic Regression, k = 5	Naive Bayes, k = 5
1	0.774338473	0.779817494
2	0.750989911	0.76345325
3	0.758737864	0.759873472
4	0.766571483	0.732483829
5	0.754844851	0.772377592
Avg	0.761096516	0.761601128
Stdev	0.00938236	0.018043007

Cancer Training Accuracy

Test No.	Logistic Regression, k = 1	Naive Bayes, k = 10
1	0.76525	0.783875
2	0.762125	0.762125
3	0.755	0.77325
4	0.7705	0.78525
5	0.767875	0.775
Avg	0.76415	0.7759
Stdev	0.005983963	0.0093355

Cancer Validation Accuracy

Test No.	Logistic Regression, k = 1	Naive Bayes, k = 10
1	0.758718391	0.767823467
2	0.758718391	0.771233895
3	0.74706705	0.763475829
4	0.74708046	0.760823497
5	0.770450192	0.771038989
Avg	0.756406897	0.766879136
Stdev	0.009773885	0.004620966

Bank Training Accuracy

Test No.	Logistic Regression,k = 5	Naive Bayes, k = 5
1	0.886210463	0.898743249
2	0.88301484	0.909234238
3	0.883991623	0.892376786
4	0.885476069	0.917297325
5	0.884994904	0.901233739
Avg	0.88473758	0.903777067
Stdev	0.001255513	0.009674093

Bank Validation Accuracy

Test No.	Logistic Regression, k = 5	Naive Bayes, k = 5
1	0.882526723	0.91987235
2	0.766571483	0.905377299
3	0.884245567	0.901342988
4	0.88572128	0.909234898
5	0.860809922	0.912347987
Avg	0.052694215	0.909635104
Stdev	0.012921222	0.007057074

Bank Training Accuracy

Test No.	Logistic Regression, k = 1	Naive Bayes, k = 10
1	0.88378064	0.897293748
2	0.883229272	0.901277237
3	0.882784735	0.899832477
4	0.884209152	0.901238473
5	0.884744495	0.902887874
Avg	0.883749659	0.900505962
Stdev	0.000775189	0.002096237

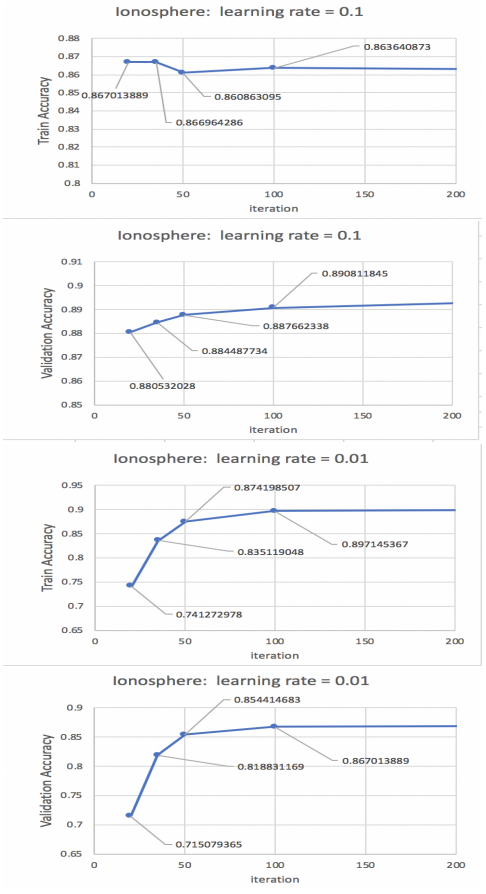
Bank Validation Accuracy

Test No.	Logistic Regression, k = 1	Naive Bayes, k = 10
1	0.886212747	0.897987235
2	0.882279794	0.901234475
3	0.884492509	0.909834892
4	0.886704013	0.91987235
5	0.883509371	0.902348724
Avg	0.884639687	0.906255535
Stdev	0.001844202	0.008761684

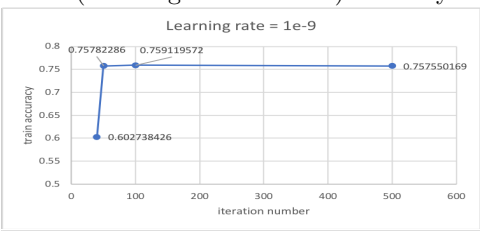
4.2 Logistic Regression under different hyperparameters

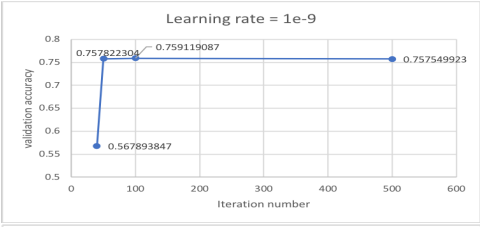
When iteration was below 100,accuracy increased significantly as iteration increased.But accuracy did not change much,when iteration was above 100.

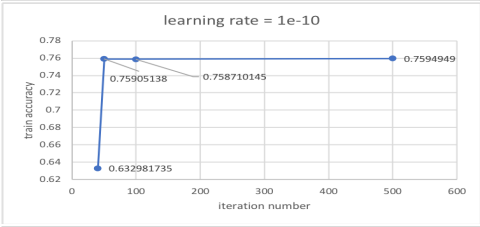
Ionosphere(Training&Validation)Accuracy-Iteration

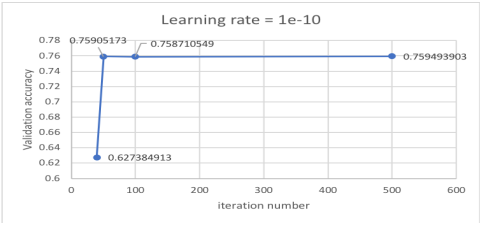


Adult (Training & Validation)Accuracy-Iteration

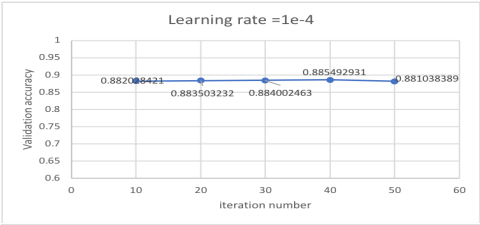


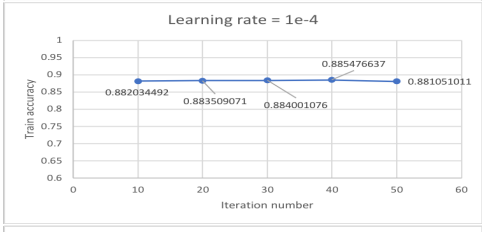


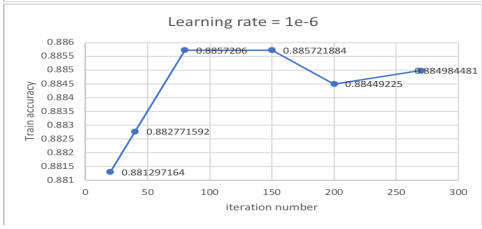


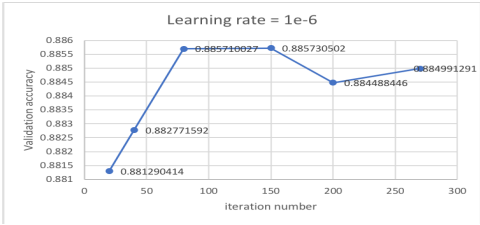


Bank (Training & Validation)Accuracy-Iteration

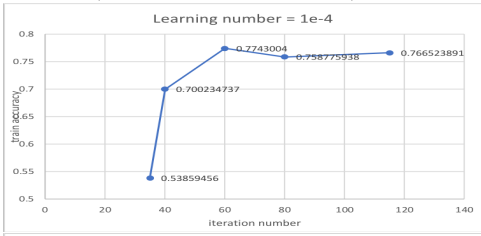


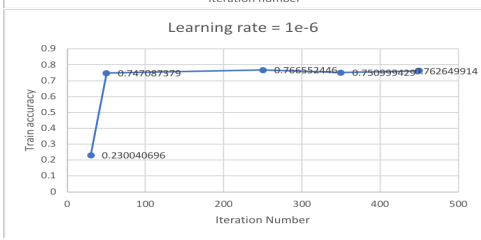


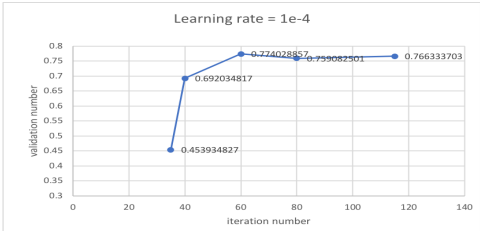


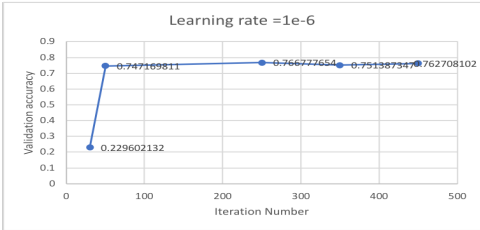


Cancer (Training & Validation)Accuracy-Iteration



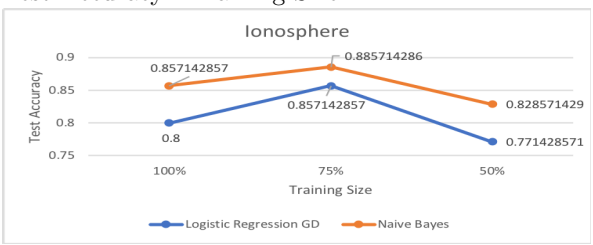


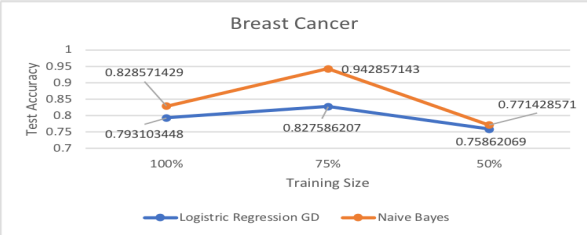




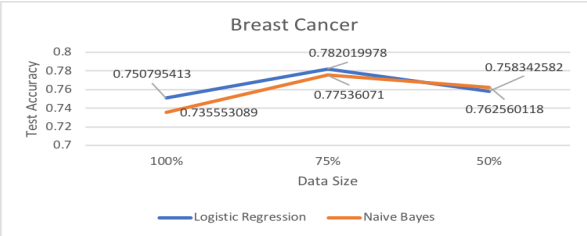
4.3 Relation of training size and accuracy

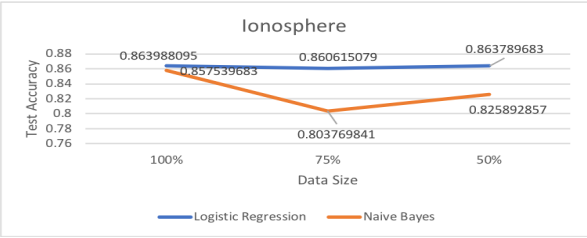
Test Accuracy - Training Size





K-Cross-Validation Accuracy - Training Size





4.4 Stochastic Gradient V.S. Full Batch Gradient

Stochastic gradient descent provided less accuracy than full batch Gradient Descent.

Training Accuracy

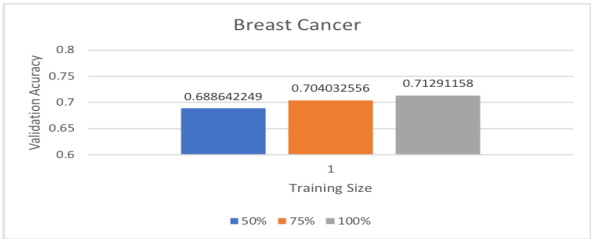
Data Sets:	Gradient Descent Accuracy	Stochastic Gradient Descent Accuracy
ionosphere	0.867769827	0.735763268
adult	0.759597326	0.759324284
breast-cance	0.770359848	0.743016098
bank	0.884998408	0.881787597

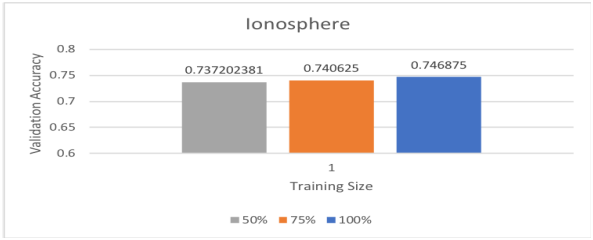
Validation Accuracy

Data Sets:	Gradient Descent Accuracy	Stochastic Gradient Descent Accuracy
ionosphere	0.888352772	0.743345103
adult	0.760211786	0.758300637
breast-cance	0.770434028	0.76265873
bank	0.885475559	0.88621281

4.5 Logistic Analytical Matrix Closed-Form Accuracy

The accuracies of analytical method are generally lower than using Logistic Regression GD or Naive Bayes. And we note that the variation for accuracies for different training size is relatively small.





5.Discussion and Conclusions

In conclusion, we noticed the Logistic Regression and Naive Bayes model performed comparably on all datasets. As for the Logistic Regression, full batch gradient descent resulted in better prediction accuracy compared to the stochastic gradient descent, but the analytical method generated the worst accuracy level. In further developnment, the analytical approach of Logistic Regression shall be investigated in order to accomplish a more reliable prediction. We also learned that a large iteration step would lead to higher accuracy in gradient descent. However, a change in learning rate from 0.1 to 0.01 didn't make a significant variance. The second issue we shall consider in future development is the Naive Bayes runtime issue. When calculating the accuracy using K-fold on dataset "adult", our Naive Bayes algorithm regularly took over 20 minutes to generate its accuracy result. Such an issue diminished the competence and effectiveness of the Naive Bayes Model. Therefore, the algorithm of Naive Bayes should be investigated further in order to improve the Naive Bayes efficiency.

6.Statement of Contributions

All three team members made even and adequate contributions to the project in terms of the time spend and the code written.