Comparison analysis between Logistic Regression and Naive Bayes in Machine Learning

Donghan LiuDingtao HuJifeng Wang
Mcgill University COMP551 Mini project1

***Abstract*—-In this project, we investigated the performance of two classification models - logistic regression (LR) and Naive Bayes (NB). Each method was implemented and tested with four bench mark datasets to evaluate the outcome accuracy. K-fold cross validation was applied during the evaluation and confusion matrix was used to obtain recallprecision and F1 score. Normalization is adopted to improve the predicting model. After comparing the outcome accuracy and time cost, we found that logistic regression is comparatively more time-consuming in the validation process and has slightly higher accuracy compared to naive Bayes method. Additionally,we investigated the contribution of each feature on the prediction results by Feature Importance for future exploration.**

I. Introduction

A. Preliminaries

Machine learning is a method used to make predictions or decisions by building mathematical models based on training data. To generate good predictions of data based on decision boundary, Logistic Regression (LR) and Naive Bayes (NB) are investigated to implement the model.[1]

LR: Logistic Regression is a discriminative learning method. Trained with the data to the learn the conditional distribution directly. To find such a series of weights, so that logistic function maximizes the most likelihood function, we adjust our initial weights through the gradient descend process in which the learning rate and number of iterations are selected manually. For improving the performance, we applied the dynamic weighting-based feature selection.[2] The primary characteristic of the method is that the feature is weighted according to its interaction with the selected features and the weight of features will be dynamically updated after each candidate feature has been selected.

NB: Naive Bayes is a probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between features.

B. Two Additional Data sets

1. Wine Quality Data Set [5]

The goal is to predict the quality of wine based on its chemical properties. For any sample with quality rating larger than or equal to 6, we counted its binary classification as positive(1), the rest was rated as negative(0).

2. Breast Cancer Wisconsin (Diagnostic) Data Set [4]

The goal for this data set is to predict whether a tumor is malignant or benign based on its various properties. For the sample that was classified with 1 as benign and the rest with 0 as malignant.

C. Task Description and Important Findings

The main goal of this project is to implement two algorithms to compares their performances in terms of result accuracy and time cost as well as investigating the possible improvements by manipulating the existing features in data sets.

This project was separated into three parts, the first task was to optimize the training of Logistic Regression by adjusting the learning rate. In this step we observed that there is peak on the accuracy with different learning rate when the number of fixed iterations is relatively large. After we have refined the LR model by the learning rate with the best performance, two models were compared with each other by their performances. Generally, NB has 2 percentages to 5 percentages lower accuracy among four data sets than LR and LR consumes more time (1s–2s) than NB on numerical data sets while NB consumes exponentially more time on categorical data set (Adult) but LR prevents the problem by hot-encoding. Then, we modified the feature and tried to improve the final accuracy of LR for wine quality data set which results in a subtle increase in accuracy.

The key taken away in this project was that by adjusting and manipulating the parameters of each learning model, we gained more understanding of the impact changing those parameters would have to the model and the interactive relationship between different parameters. As well as some possible approaches to improve the algorithms and data sets for better predicting.

II. Data Process

A. Data acquiring

Item with malformed data is removed from the data set prior to applying machine learning. All items are normalized by $(\frac{x-\bar{x}}{\sigma})$before the machine learning process.

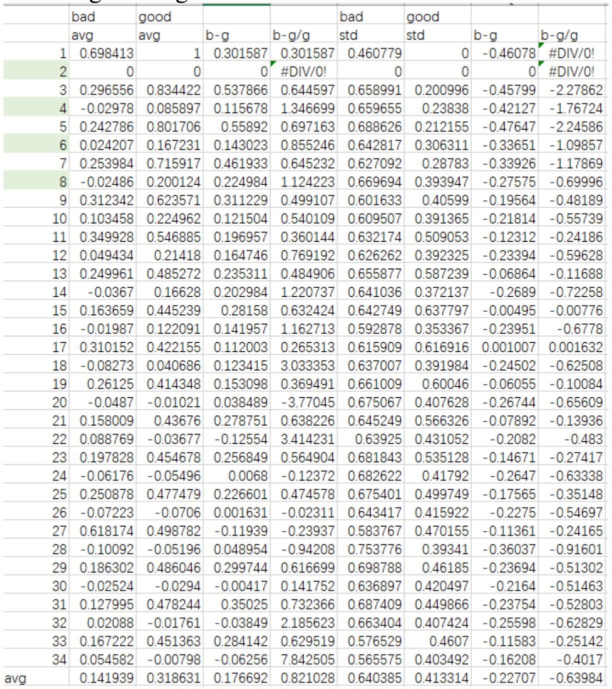
1) Ionosphere
   The goal is to predict the quality of data return by radar based on the 34 continuous attributes , which deviated from complex electromagnetic signal.[3]
2) Breast Cancer Wisconsin
   The goal is to predict whether the patient is malignant or benign based on 10 features of cell nucleus from the patient.[4]
3) Wine QualityThe goal is to determine the quality of wine based on 11 results from physicochemical tests. In order to perform binary classification, each wine

which quality is greater than 5 would be classified to high quality in this project.[5]

4) Adult
The goal is to determine whether a person makes over 50k a year by analyzing 13 detailed personal information, including education, race, native-country, and etc.[6]

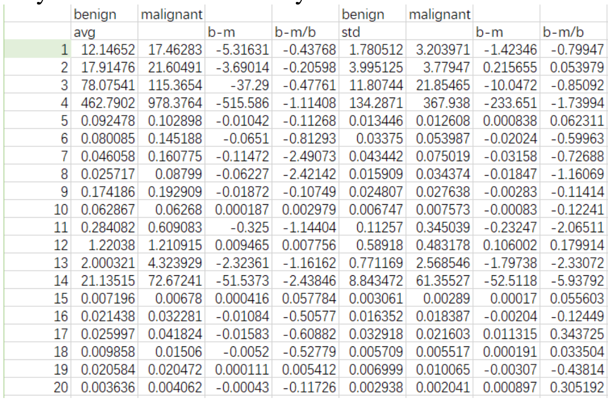
B. Data Analysis

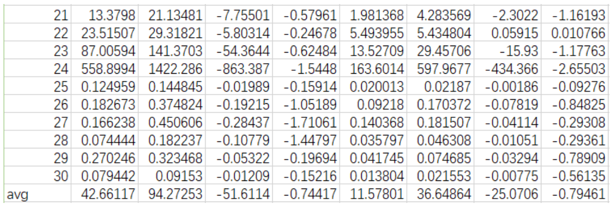
1) Ionosphere
The table below is the simple analysis of Ionosphere. After taking the average of all the attribute, one may observe that good signal's result is significantly higher than the bad one. What's more, the standard deviation of the good signal is also less than the bad one.



	bad	good			bad	good		
	avg	avg	b-g	b-g/g	std	std	b-g	b-g/g
1	0.698413	1	0.301587	0.301587	0.460779	0	-0.46078	#DIV/0!
2	0	0	0	#DIV/0!	0	0	0	#DIV/0!
3	0.296556	0.834422	0.537866	0.644597	0.658991	0.200996	-0.45799	-2.27862
4	-0.02978	0.085897	0.115678	1.346699	0.659655	0.23838	-0.42127	-1.76724
5	0.242786	0.801706	0.55892	0.697163	0.688626	0.212155	-0.47647	-2.24586
6	0.024207	0.167231	0.143023	0.855246	0.642817	0.306311	-0.33651	-1.09857
7	0.253984	0.715917	0.461933	0.645232	0.627092	0.28783	-0.33926	-1.17869
8	-0.02486	0.200124	0.224984	1.124223	0.669694	0.393947	-0.27575	-0.69996
9	0.312342	0.623571	0.311229	0.499107	0.601633	0.40599	-0.19564	-0.48189
10	0.103458	0.224962	0.121504	0.540109	0.609507	0.391365	-0.21814	-0.55739
11	0.349928	0.546885	0.196957	0.360144	0.632174	0.509053	-0.12312	-0.24186
12	0.049434	0.21418	0.164746	0.769192	0.626262	0.392325	-0.23394	-0.59628
13	0.249961	0.485272	0.235311	0.484906	0.655877	0.587239	-0.06864	-0.11688
14	-0.0367	0.16628	0.202984	1.220737	0.641036	0.372137	-0.2689	-0.72258
15	0.163659	0.445239	0.28158	0.632424	0.642749	0.637797	-0.00495	-0.00776
16	-0.01987	0.122091	0.141957	1.162713	0.592878	0.353367	-0.23951	-0.6778
17	0.310152	0.422155	0.112003	0.265313	0.615909	0.616916	0.001007	0.001632
18	-0.08273	0.040686	0.123415	3.033353	0.637007	0.391984	-0.24502	-0.62508
19	0.26125	0.414348	0.153098	0.369491	0.661009	0.60046	-0.06055	-0.10084
20	-0.0487	-0.01021	0.038489	-3.77045	0.675067	0.407628	-0.26744	-0.65609
21	0.158009	0.43676	0.278751	0.638226	0.645249	0.566326	-0.07892	-0.13936
22	0.088769	-0.03677	-0.12554	3.414231	0.63925	0.431052	-0.2082	-0.483
23	0.197828	0.454678	0.256849	0.564904	0.681843	0.535128	-0.14671	-0.27417
24	-0.06176	-0.05496	0.0068	-0.12372	0.682622	0.41792	-0.2647	-0.63338
25	0.250878	0.477479	0.226601	0.474578	0.675401	0.499749	-0.17565	-0.35148
26	-0.07223	-0.0706	0.001631	-0.02311	0.643417	0.415922	-0.2275	-0.54697
27	0.618174	0.498782	-0.11939	-0.23937	0.583767	0.470155	-0.11361	-0.24165
28	-0.10092	-0.05196	0.048954	-0.94208	0.753776	0.39341	-0.36037	-0.91601
29	0.186302	0.486046	0.299744	0.616699	0.698788	0.46185	-0.23694	-0.51302
30	-0.02524	-0.0294	-0.00417	0.141752	0.636897	0.420497	-0.2164	-0.51463
31	0.127995	0.478244	0.35025	0.732366	0.687409	0.449866	-0.23754	-0.52803
32	0.02088	-0.01761	-0.03849	2.185623	0.663404	0.407424	-0.25598	-0.62829
33	0.167222	0.451363	0.284142	0.629519	0.576529	0.4607	-0.11583	-0.25142
34	0.054582	-0.00798	-0.06256	7.842505	0.565575	0.403492	-0.16208	-0.4017
avg	0.141939	0.318631	0.176692	0.821028	0.640385	0.413314	-0.22707	-0.63984

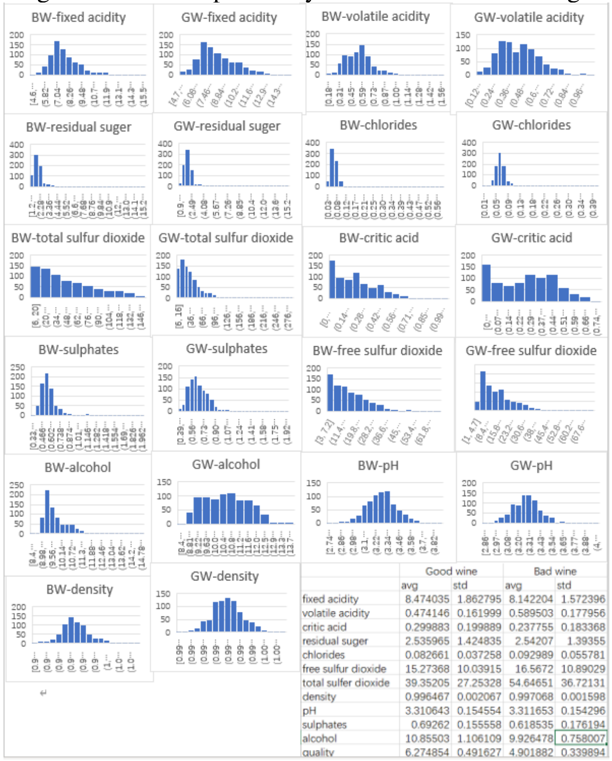
2) Breast Cancer Wisconsin
The table below is the simply analysis of Breast Cancer Wisconsin. After taking the average and standard deviation of all attributes, one may observe that the nucleus fro malignant patient is roughly double in size and 4 times in compactness and concavity. However, the diagnosis may not be effected by the fractal dimension.





	benign	malignant			benign	malignant		
	avg		b-m	b-m/b	std		b-m	b-m/b
1	12.14652	17.46283	-5.31631	-0.43768	1.780512	3.203971	-1.42346	-0.79947
2	17.91476	21.60491	-3.69014	-0.20598	3.995125	3.77947	0.215655	0.053979
3	78.07541	115.3654	-37.29	-0.47761	11.80744	21.85465	-10.0472	-0.85092
4	462.7902	978.3764	-515.586	-1.11408	134.2871	367.938	-233.651	-1.73994
5	0.092478	0.102898	-0.01042	-0.11268	0.013446	0.012608	0.000838	0.062311
6	0.080085	0.145188	-0.0651	-0.81293	0.03375	0.053987	-0.02024	-0.59963
7	0.046058	0.160775	-0.11472	-2.49073	0.043442	0.075019	-0.03158	-0.72688
8	0.025717	0.08799	-0.06227	-2.42142	0.015909	0.034374	-0.01847	-1.16069
9	0.174186	0.192909	-0.01872	-0.10749	0.024807	0.027638	-0.00283	-0.11414
10	0.062867	0.06268	0.000187	0.002979	0.006747	0.007573	-0.00083	-0.12241
11	0.284082	0.609083	-0.325	-1.14404	0.11257	0.345039	-0.23247	-2.06511
12	1.22038	1.210915	0.009465	0.007756	0.58918	0.483178	0.106002	0.179914
13	2.000321	4.323929	-2.32361	-1.16162	0.771169	2.568546	-1.79738	-2.33072
14	21.13515	72.67241	-51.5373	-2.43846	8.843472	61.35527	-52.5118	-5.93792
15	0.007196	0.00678	0.000416	0.057784	0.003061	0.00289	0.00017	0.055603
16	0.021438	0.032281	-0.01084	-0.50577	0.016352	0.018387	-0.00204	-0.12449
17	0.025997	0.041824	-0.01583	-0.60882	0.032918	0.021603	0.011315	0.343725
18	0.009858	0.01506	-0.0052	-0.52779	0.005709	0.005517	0.000191	0.033504
19	0.020584	0.020472	0.000111	0.005412	0.006999	0.010065	-0.00307	-0.43814
20	0.003636	0.004062	-0.00043	-0.11726	0.002938	0.002041	0.000897	0.305192
21	13.3798	21.13481	-7.75501	-0.57961	1.981368	4.283569	-2.3022	-1.16193
22	23.51507	29.31821	-5.80314	-0.24678	5.493955	5.434804	0.05915	0.010766
23	87.00594	141.3703	-54.3644	-0.62484	13.52709	29.45706	-15.93	-1.17763
24	558.8994	1422.286	-863.387	-1.5448	163.6014	597.9677	-434.366	-2.65503
25	0.124959	0.144845	-0.01989	-0.15914	0.020013	0.02187	-0.00186	-0.09276
26	0.182673	0.374824	-0.19215	-1.05189	0.09218	0.170372	-0.07819	-0.84825
27	0.166238	0.450606	-0.28437	-1.71061	0.140368	0.181507	-0.04114	-0.29308
28	0.074444	0.182237	-0.10779	-1.44797	0.035797	0.046308	-0.01051	-0.29361
29	0.270246	0.323468	-0.05322	-0.19694	0.041745	0.074685	-0.03294	-0.78909
30	0.079442	0.09153	-0.01209	-0.15216	0.013804	0.021553	-0.00775	-0.56135
avg	42.66117	94.27253	-51.6114	-0.74417	11.57801	36.64864	-25.0706	-0.79461

3) Wine Quality
The graphs below indicates the comparison of attribute distribution between good wines and bad wines. The major difference is the total sulfur dioxide, which good wine contain 2/3 of total sulfur dioxide in bad wines. In addition, the standard deviation of alcohol for good wine is smaller than bad wine's; indicates that alcohol concentration in good wine comparatively centered at the average



	Good wine		Bad wine	
	avg	std	avg	std
fixed acidity	8.474035	1.862795	8.142204	1.572396
volatile acidity	0.474146	0.161999	0.589503	0.177956
critic acid	0.299883	0.199889	0.237755	0.183368
residual suger	2.535965	1.424835	2.54207	1.39355
chlorides	0.082661	0.037258	0.092989	0.055781
free sulfur dioxide	15.27368	10.03915	16.5672	10.89029
total sulfer dioxide	39.35205	27.25328	54.64651	36.72131
density	0.996467	0.002067	0.997068	0.001598
pH	3.310643	0.154554	3.311653	0.154296
sulphates	0.69262	0.155558	0.618535	0.176194
alcohol	10.85503	1.106109	9.926478	0.758007
quality	6.274854	0.491627	4.901882	0.339894

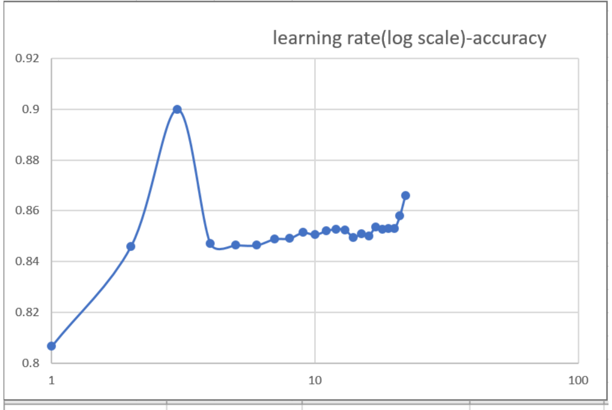
4) Adult
According to the data set, we may hardly construct a relationship between income and other features except age, which would effect education and many other features by common sense, and working hours per week.

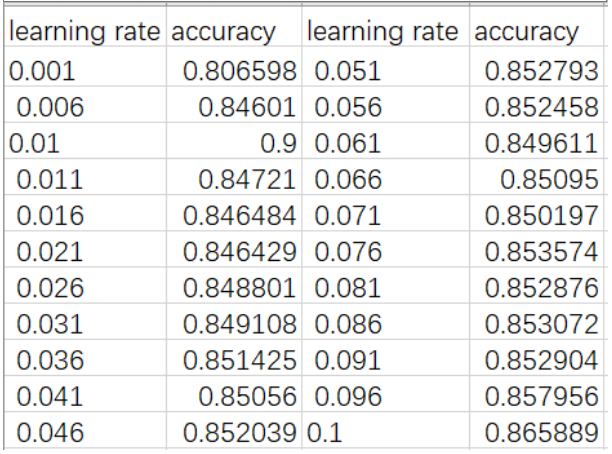
III. Results

A. Adjusting the learning rate for logistic regression model

The task started by deciding the suitable learning rate for logistic regression algorithm and observe the influences of changing in learning rate to the prediction performances of model. To keep the comparison between large and small learning rate unaffected by the number of iteration, we initialized the learning rate at 0.001 to ensure it reaches the area of local minimum before training ended. A lab is designed to test and compare the average accuracy between models with 22 learning rates increases from 0.001 to 0.1.

The iteration of training was fixed to 2000. Ionosphere data set is used for this lab and the accuracy is evaluated by using k-fold validation with 5-folds. The results of the test are shown in the figure below. It turns out that the overall change of learning rate does not have significant effect to the outcome accuracy, all results are fluctuated between 0.8 and 0.9.Specifically, we can see that as the learning rate increase, there will be a point where the accuracy stops increasing and starts to decrease rapidly, where the point is at 0.01 by our adjusting process. And after the point 0.011 the learning rate-accuracy line rises in a small range and boosts slightly at last two points. In practice, our learning rate should ideally be somewhere to the left to the highest point of the graph. [5] In this case, we pick the point 0.01.

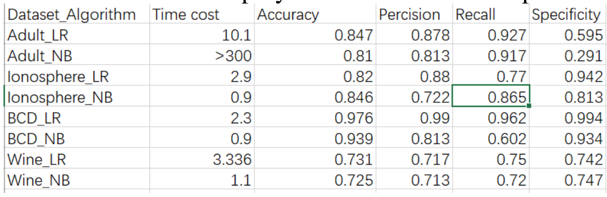




learning rate	accuracy	learning rate	accuracy
0.001	0.806598	0.051	0.852793
0.006	0.84601	0.056	0.852458
0.01	0.9	0.061	0.849611
0.011	0.84721	0.066	0.85095
0.016	0.846484	0.071	0.850197
0.021	0.846429	0.076	0.853574
0.026	0.848801	0.081	0.852876
0.031	0.849108	0.086	0.853072
0.036	0.851425	0.091	0.852904
0.041	0.85056	0.096	0.857956
0.046	0.852039	0.1	0.865889

B. Comparing between two algorithms

Then do time cost, outcome accuracy, precision, recall, and specificity comparison between two algorithms. We would employ 5-fold cross validation with four given data sets to determine the desired outcome. We also employ normalization as pretreatment



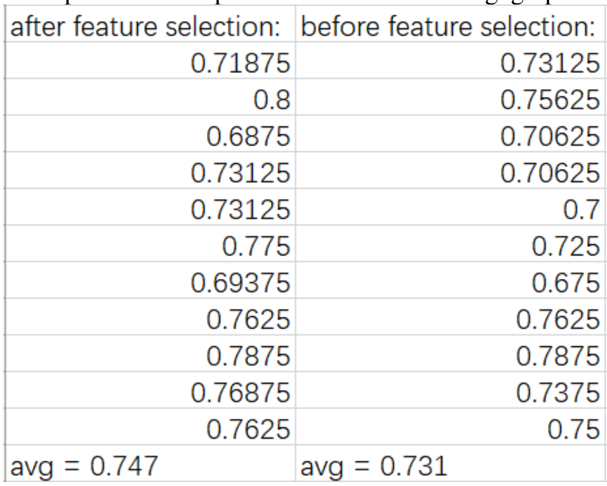
Dataset_Algorithm	Time cost	Accuracy	Percision	Recall	Specificity
Adult_LR	10.1	0.847	0.878	0.927	0.595
Adult_NB	>300	0.81	0.813	0.917	0.291
Ionosphere_LR	2.9	0.82	0.88	0.77	0.942
Ionosphere_NB	0.9	0.846	0.722	0.865	0.813
BCD_LR	2.3	0.976	0.99	0.962	0.994
BCD_NB	0.9	0.939	0.813	0.602	0.934
Wine_LR	3.336	0.731	0.717	0.75	0.742
Wine_NB	1.1	0.725	0.713	0.72	0.747

As shown above, Naive Bayes outmatch Linear Regression in time efficiency with three data set. All of the three data set have only numeric terms in their attributes. However, the time cost of Naive Bayes would enormously grow if literal classification presents as attributes.

Comparing the outcome accuracy, Linear Regression is slightly higher than Naive Bayes in most of the tests. Moreover, in order to further investigate the diversity between two algorithm, we calculate precision, recall, and specificity. It turns out that Logistic regression beat Naive Bayes in most of the cases, especially when Naive Bayes was struggling with literal classifications in the adult data set.

C. Feature editing

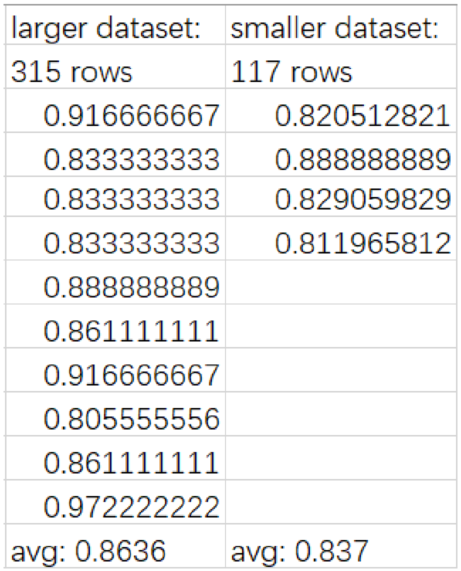
Next, we focused on further improving the outcome accuracy by deleting features from data sets.Ionosphere data set is used in this section. During the process, we added a weight on the gradient descent on Ionosphere's LR model and listed the weight of each feature in trained model to figure out which one takes significantly larger portion than others, and the following step is to take away the features that has less significant weight than others.After the feature selection the model was trained for 11 times while in each time the model was trained with the same selected features. However, as they did not affect the training model much due to their weight, deleting those features turns out having no significant impact to the model accuracy. The improvement is in the range of 1 percent to 2 percents as the following graph shows.



after feature selection:	before feature selection:
0.71875	0.73125
0.8	0.75625
0.6875	0.70625
0.73125	0.70625
0.73125	0.7
0.775	0.725
0.69375	0.675
0.7625	0.7625
0.7875	0.7875
0.76875	0.7375
0.7625	0.75
avg = 0.747	avg = 0.731

D. Data Size versus Accuracy

At last we investigated the relationship between the accuracy as a function of the size of data set. The LR model of Ionosphere data set is used in this section.By comparing the accuracy on training of larger data set (with 315 attributes) and smaller data set (with 117 attributes), it is observed that the accuracy has positive correlation with the expansion of data size. The double growth on the volume of data leads to an obvious increase on the accuracy which is of 3 percents approximately. Under the analysis, this result is caused by decrease of contingency factors due to the increase in data size. [8] This result is limited to the lack of the changing of data size. The future investigation will be concentrate on expanding the experimental group, the overfitting of large size and exploring the relativity between training data size and the number of features in both models.



larger dataset:	smaller dataset:
315 rows	117 rows
0.916666667	0.820512821
0.833333333	0.888888889
0.833333333	0.829059829
0.833333333	0.811965812
0.888888889	
0.861111111	
0.916666667	
0.805555556	
0.861111111	
0.972222222	
avg: 0.8636	avg: 0.837

IV. Discussion and Conclusion

In this assignment, we investigate two supervised learning method and discovered that Naive Bayes algorithm is less time consuming compared to logistic regression model for numerical data set, which leads to a signicant time saving during training process. One should regards the considerably starting time and carefully choose proper learning rate before using Logistic regression. In addition, it is necessary to adopt proper strategies like normalization, decaying learning rate, estimated number of iterations etc. to ensure the weights are properly trained and eventually reaches its target. Although Naive Bayes performs better in time complexity than Logistic Regression in most of the cases; it is generally less accuracy, precision, recall, and sensitivity comparing to the performance of a well-trained Logistic Regression.

In addition, enough emphasis should be laid on the modification on feature selection which results in a negligible improvement on accuracy during our process. The strategies like adding more features to the model by some non-linear combinations (x^2,$x * y$ etc.) for the features that have significant weights and the method of Forward Selection and Backward Elimination aiming to generate a good subset should be considered as the solutions to tackle the problem. There might be some other approaches to improve the accuracy of the models, such as Hyperparameter tuning, regularization and improvement on normalization if the distribution of data set is not approximate to normal. These could be the areas left for further investigation.

References

[1] C. M. Bishop, Pattern recognition and machine learning. New York: Springer, 2006.

[2] Sun, X.,Liu, Y.,Xu, M., Chen, H.(2013). Knowledge-Based Systems(Page541-549). Retrieved from https://www.science direct.com/science/article/pii/S0950705112002699

[3] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/ionosphere]. Irvine, CA: University of California, School of Information and Computer Science.

[4] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://https://archive.ics.uci.edu/ml/datasets/Breast +Cancer+Wisconsin+(Diagnostic)]. Irvine, CA: University of California, School of Information and Computer Science.

[5] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://https://archive.ics.uci.edu/ml/datasets/Wine]. Irvine, CA: University of California, School of Information and Computer Science.

[6] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://https://archive.ics.uci.edu/ml/datasets/Adult]. Irvine, CA: University of California, School of Information and Computer Science.

[7] Zulkifli, H (2018) Understanding Learning Rates and How It Improves Performance in Deep Learning Retrieved from https://towardsdatascience.com/understanding-learning-rates-and-how-it-improves-performance-in-deep-learning-d0d4059c1c10

[8] D. Udhayakumarapandian (2016). Data Size versus Accuracy: Performance by different Data Mining Tools. Retrieved from http://ijarcet.org/wp-content/uploads/IJARCET-VOL-5-ISSUE-3-577-580.pdf