COMP 551 mini project 1

Abdulla Alseiari Kenji Marshall Mian Hamza

02/11/2020

Abstract

We implemented and compared two binary classification methods, namely Logistic Regression and Naive Bayes. We compared them across four data sets, and conducted experimentation on the effects of Logistic Regression learning rate, Elastic Net regularization, early stopping, training set size, and feature engineering. The Logistic Regression outperformed Naive Bayes on two of four data sets, which might be influenced by strong inter-feature correlation violating Naive Bayes assumptions. The Naive Bayes performed better on our smallest data set, where its strong biases are most potent. One data set showed equivalent performance.

1 Introduction

Logistic Regression and Naive Bayes are two different types of classifiers, where Logistic Regression is discriminative and Naive Bayes is generative. Discriminative classifiers learn the conditional probability of the label directly, learning parameters to map the input directly to the class [4]. Generative classifiers learn the joint probability distribution between the labels and the actual data, and then use Bayes' rule to learn the posterior probability and make predictions [4]. In computing the joint probability, Naive Bayes makes the strong and "naive" assumption that all features are conditionally independent. Comparing the two models is an important research field in machine learning, although in 2008, Xue and Titterington claimed that neither approach has been proven to be theoretically consistently better than the other. They claimed that "the choice depends on the relative confidence we have in the correctness of the specification of either p(y—x) or p(x, y) for the data." [6]. Since the two classifier model different probabilities, and make very different assumptions of the data, deciding between them will largely depend on the validity of those assumptions, and how accurately we can determine our desired probability.

The two data sets that were chosen for us are the Ionosphere and Adult data sets from the UCI machine learning repository, which are both binary classification problems[1]. The Ionosphere data consists of data collected from a phased-array of high-frequency antennas sending 17 pulses into the ionosphere at different time lags collected by radars in Goose Bay, Labrador [5]. This results in 34 features, representing the real and imaginary component of the returning signals. The database consists of 351 instances. The task was to predict if the signal is good or bad, where a "good" signal is one that demonstrates structure in the ionosphere. A reference neural network published by Sigillito et al achieved a test set accuracy of 98%. The second data set is the adult census data set [1]. This data set has 14 features of which 8 are categorical and 6 are numerical and 48842 examples. These features provide demographic information about people, and the goal is to predict if a person earns more than 50K annually. A reference solution using a Naive Bayes model is given by Ron Kohavi at 81.69% [2]. We chose two supplementary data set from the UCI Machine Learning Repository: Breast Cancer and Banknote Authentication. The breast cancer data set has 9 attributes and 286 instances [3]. The goal was to predict whether the patient had breast cancer recurrence. The banknote authentication data set has 4 numerical features that were properties of a wavelet transform of photographed paper bills [1]. The data set includes 1372 instances where the aim was to predict the authenticity of the banknote.

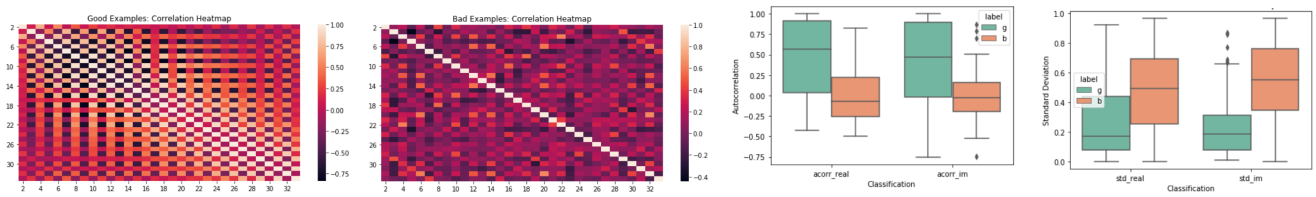


Figure 1: Insights derived from Ionosphere time series analysis. From left to right: correlation in good examples, correlation in bad examples, autocorrelation, standard deviation

2 Data Sets

For all four data sets, we explored the distributions by label for all categorical and numerical features. Subsets or transformations of numerical features had their correlation visualized between themselves and with the class. Any example with missing features was removed. Categorical variables were one hot encoded, and all numerical data was normalized to have a zero mean and a unit variance. Each data set also yielded specific insights throughout the data processing phase.

Since the Ionosphere data represented a real and imaginary time series of radar readings, we used this to help find structure in the data. For example, the inter-feature correlation, visualized in Figure 1showed how good examples had correlated time series (checkerboard pattern). This led us to examine autocorrelation as a discriminative feature (Figure 1 Also, by visualizing a collection of good/bad time series, it was identified that bad examples tended to have higher variance; this claim was empirically realized as well (Figure 1).

In the Adult data set, we had 3620 malformed examples that were dropped. In analysis, different features showed varying levels of discrimination for the labels. For example, age showed a clear delineation with older people earning more on average. The correlation between all numerical features is given in Figure 2. In one hot encoding, this resulted in 103 columns. In order to address the potential of models not handling high dimensionality, an alternative encoding was also presented using geographical regions instead of countries, that reduced the features to 74. These two structures are examined in experimentation. For Breast Cancer, 9 examples were malformed. This data was primarily categorical, but yielded some interesting relationships, such as between age, tumor malignancy, and label. This is shown in Figure 3. Banknote had no malformed examples. The correlation heatmaps for both of these data sets are shown in Figure 2.

The positive class distributions for the classes were 64.1%, 23.9%, 29.7%, and 44.5% for Ionosphere, Adult, Breast Cancer, and Banknote respectively.

Complete analysis of the data sets can be found in the accompanying Jupyter Notebook. For each data set, depending on the type of each feature, violin, box and whisker, bar, or scatter plots were used to visualize their distributions and how they depend on the classification. A collection of these figures are shown in Figure33.

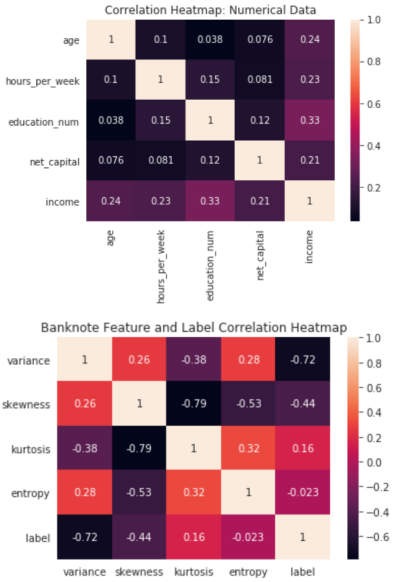


Figure 2: Feature correlations in Adult and Banknote data sets.

3 Results

While exploring the ionosphere and adult data sets, the analysis provided insight for potential methods of structurally altering the data sets. In the ionosphere data set, features like the standard deviation and autocorrelation of the real/imaginary time series showed to be discriminative. We first explore whether

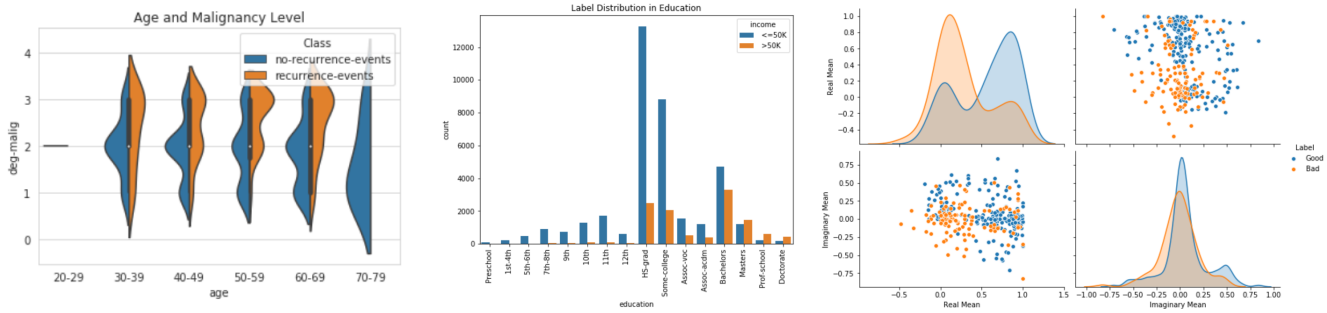


Figure 3: Sampling of plots for different datasets. Left (Breast Cancer): increase of tumor malignancy and recurrence events with age. Center (Adult): increase in income with education level. Right (Ionosphere): distribution of means for real/imaginary time series.

adding these new features will improve model performance via 5-fold cross-validation with Logistic Regression and Naive Bayes. By exploring every subset of the four features, the optimal arrangement was determined to be the inclusion of real autocorrelation and imaginary standard deviation. This led to an increase in validation accuracy from $89.2 \pm 2.8\%$ to $95.2 \pm 3.0\%$ for Logistic Regression, and from $86.6 \pm 3.6\%$ to $88.3 \pm 1.7\%$ for Naive Bayes. The Logistic Regression results are shown in Figure 4.

Figure 4: Impact of additional Ionosphere features on Logistic Regression accuracy in 5-fold cross validation. ar/ai: real and imaginary autocorrelation. sr/si: real and imaginary standard deviation.

In the adult data set, the native country feature generates 41 new features. Conversely, encoding countries by their geographical region instead will only generate 12. This dimensionality reduction could help mitigate overfitting, or it might lead to a loss of useful information. Via 5-fold cross-validation, this was shown to have a minimal effect, and slightly decreased the validation accuracy for both methods.

As such, the final Ionosphere data set included two additional features (real auto-correlation and imaginary standard deviation), and the final Adult data set used country encoding.

We also added an additional experiment focused on regularization. The Logistic Regression implementation has built-in hyper-parameters to generate Elastic Net regularization, which proposes a simple blend of Lasso and Ridge. The lambda penalty parameter is split between the L1 and L2 regimes by a mixing ratio. If this ratio is 0, then it is pure L2 regularization, and a value of 1 causes pure L1 regularization. We performed a grid search with cross validation over lambda and the mixing ratio for each data set to see how this would impact performance. The results are shown for the Ionosphere data set in Figure 8. The ideal parameters are a lambda of 0.001 with a mixing ratio of 0.3. This was able to increase validation accuracy to $95.4 \pm 3.3\%$. The Adult and Banknote data sets performed best without any regularization. The Breast Cancer data set showed ideal performance with a lambda of 0.01 and a mixing ratio of 0. This gave a validation accuracy of $76.1 \pm 4.0\%$. See the Appendix for more figures.

This experiment also implemented early stopping, where training was stopped early if validation cost increased for five successive iterations. For the Ionosphere data set, 4 out of 5 cross-validation folds stopped at or before approximately 20, 000 iterations. For Breast Cancer, all folds stopped before 2, 000 iterations.

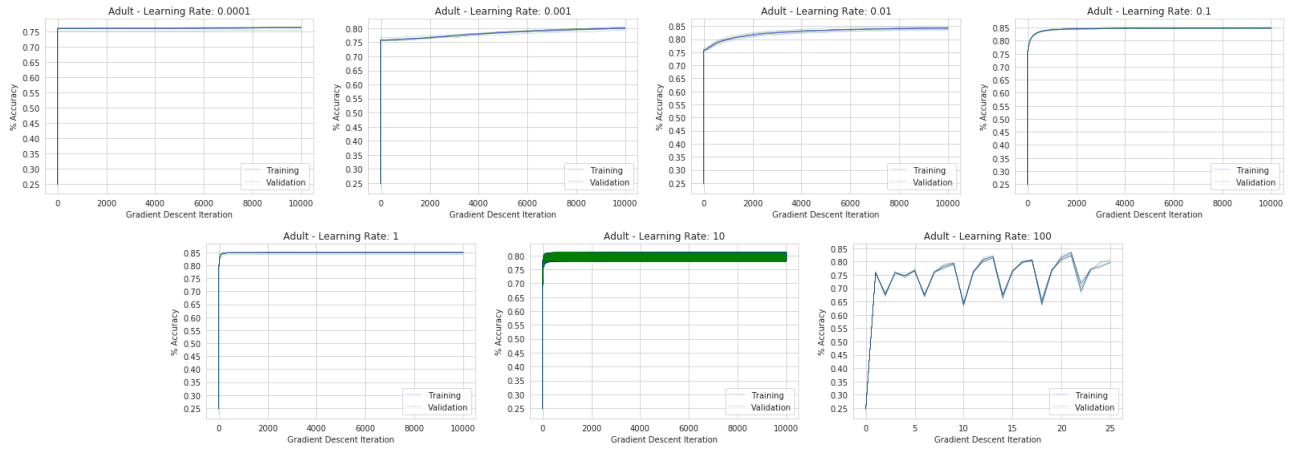


Figure 6: Logistic Regression for different learning rates for Adult data set

The Adult and Banknote data sets ran for the full 30 000 iterations. This effect is illustrated in Figure 5.

For Logistic Regression, the learning rate is an important hyper-parameter. Having a big learning rate will make the model overshoot the optima, causing it to oscillate back and forth as it tries to reach the ideal parameters. Having a small one increases the time it takes the model to reach the optimal solution. In this experiment, we plot the train and validation accuracy across all five folds in cross validation of Logistic Regression with varying learning rates to see how the model behaves. As you can see in (Figure 6), we chose 7 learning rates for our benchmark ranging from 0.0001 to 100. Higher learning rates showed oscillations in train and validation accuracy due to overshooting. The lower learning rate did not reach the optimal solution and progressed very slowly. For the Adult data set, which is shown in Figure 6, learning rates like 0.1 and 1 converge smoothly.

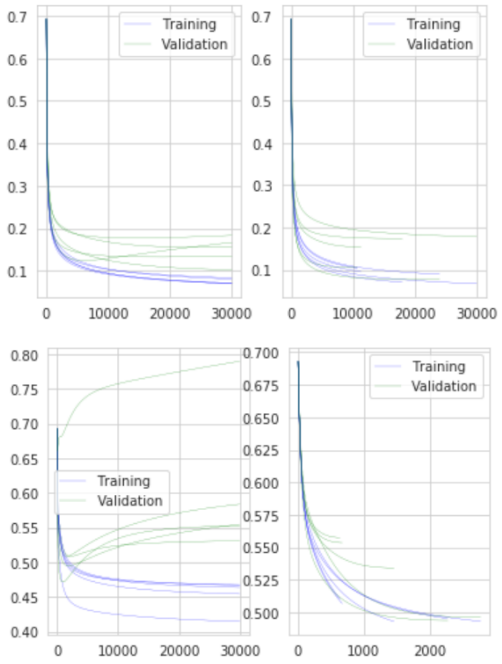


Figure 5: Cost Functions in Early Stopping: Ionosphere (top) and Breast Cancer (bottom)

Another experimental examined the performance of both models with varying training size. As shown in (Figure 7), the training set had a reduction in performance in both models while the validation and test sets had an improvement in performance as we increase the training size of the models.

Finally, we can also compare the raw performance of the models. Looking at the accuracy of the two methods (see Figure 9), we observed that the Logistic Regression model outperformed Naive Bayes for two out of four data sets, with Naive Bayes achieving higher accuracy on the Breast Cancer data set. The performance on the Adult data set was virtually equivalent. The validation accuracies across Ionosphere, Adult, Breast Cancer, and Banknote are given by 94.0 ± 2.5, 82.6 ± 3.8, 73.2 ± 6.0, and $97.4 \pm 1.0\%$ for Logistic Regression, and 88.0 ± 3.9, 82.8 ± 3.9, 74.0 ± 7.8, and $83.7 \pm 3.0\%$ for Naive Bayes. In the Adult and Breast Cancer data sets, Logistic Regression suffered from relatively low recall; as such, Naive Bayes achieved higher F1 scores.

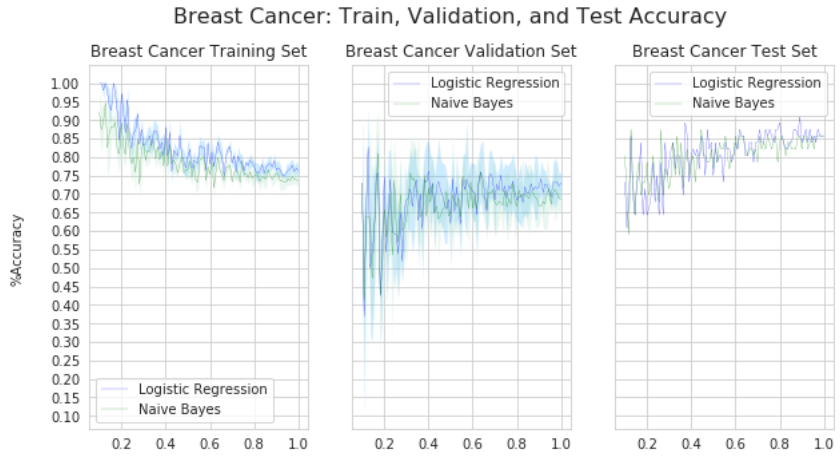


Figure 7: Train, Validation, and Test Accuracy (with error) as a function of training size

4 Discussion and Conclusion

In summary, the experiments yielded a collection of useful insights. Firstly, Elastic Net regularization provided little benefit to validation performance. This might be because Logistic Regression is a relatively simple model, and naturally doesn't fit complex, non-linear irregularities in the training data. The two smallest data sets, Ionosphere and Breast Cancer, did show increases in performance both through Elastic Net and in early stopping. This makes sense, as less data allows the Logistic Regression weights to model non-representative trends more easily. In performance comparisons, we saw that Logistic Regression had low recall on the Adult and Breast Cancer data sets; this is likely because these two data sets have the largest class imbalance (more negative examples), and thus this would allow a model to get away with lower sensitivity. Finally, Naive Bayes underperformed on Ionosphere and Banknote data. However, as seen in Figures 1 and 2, the Ionosphere and Banknote data showed strong inter-feature correlation. This violates the independence assumption of Naive Bayes, and might be a reason for why it didn't perform as well. Conversely, Naive Bayes' ability to succeed with the Breast Cancer data set may be because the data set is the smallest, and this makes Naive Bayes' strong assumptions and biases more valuable. Overall, a comparison between Logistic Regression and Naive Bayes models did not establish either model as superior. The two methods solve the same problem in two different ways, and the model preference is mainly dependent on the data itself.

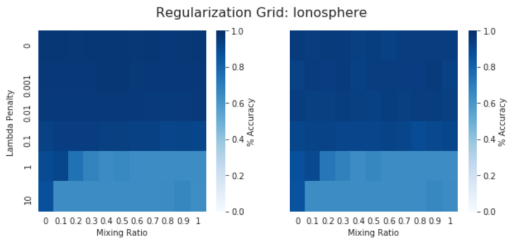


Figure 8: Regularization hyperparameter grid search for Ionosphere data set

5 Statement of Contributions

Abdulla Alseiari, Kenji Marshall and Hamza Mian contributed to the project and the write-up. Kenji implemented the models and contributed to both the analysis and the experimentation. Abdulla ran experiments one and three while Hamza analyzed the datasets, and helped implement the regression algorithms.

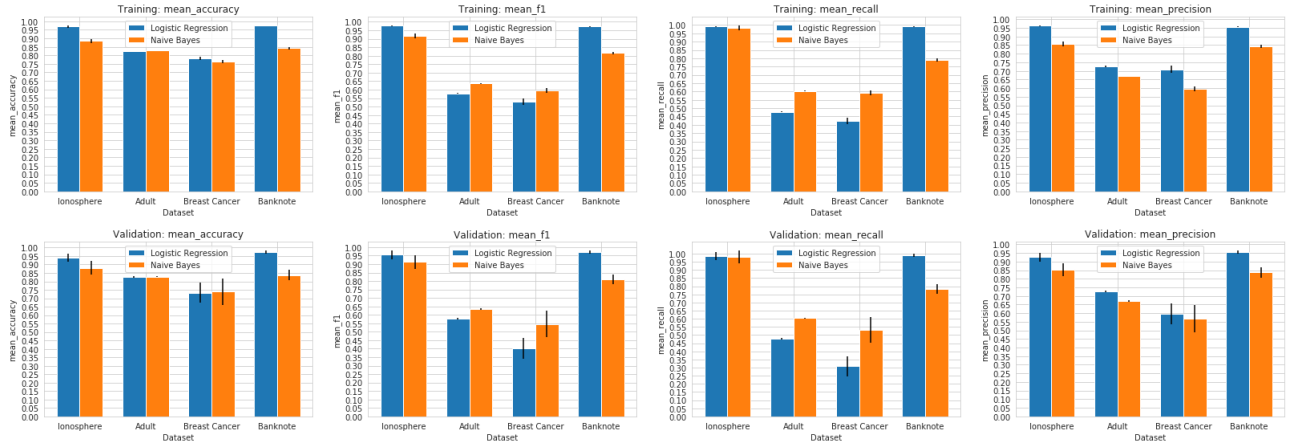


Figure 9: Logistic Regression vs Naive Bayes accuracy, F1, recall, and precision scores

References

[1] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[2] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 202–207. AAAI Press, 1996.

[3] Ryszard S Michalski, Igor Mozetic, Jiarong Hong, and Nada Lavrac. The multi-purpose incremental learning system aq15 and its testing application to three medical domains. *Proc. AAAI 1986*, pages 1–041, 1986.

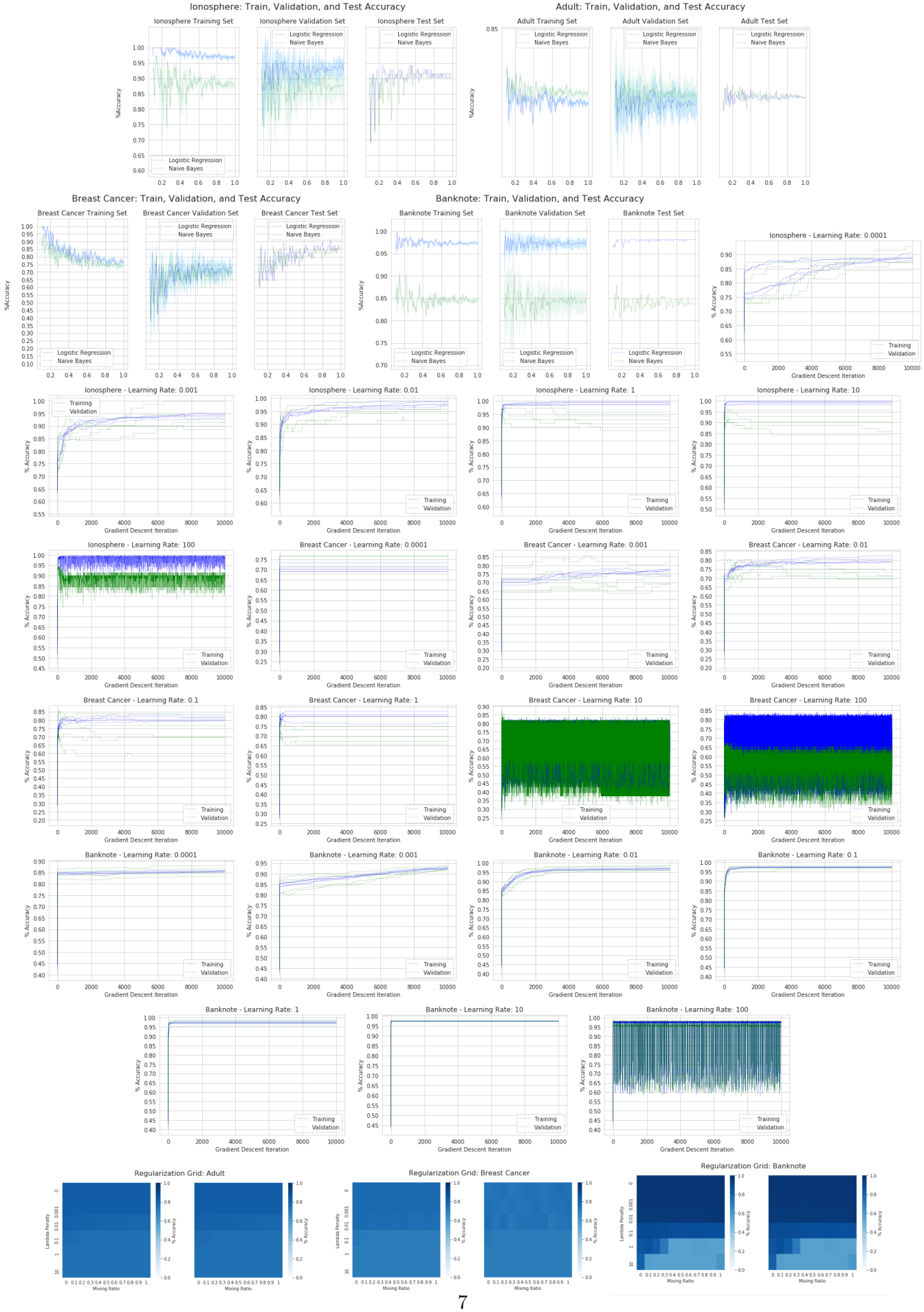
[4] Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.

[5] V G Sigillito, S P Wing, L V Hutton, and K B Baker. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Tech. Dig*, vol. 10:262–266, 1989. in.

[6] Jing-Hao Xue and D Michael Titterington. Comment on "on discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes". *Neural processing letters*, 28(3):169, 2008.

6 Appendix

Appendices


Ionosphere: Train, Validation, and Test Accuracy
Ionosphere Training Set
Ionosphere Validation Set
Ionosphere Test Set
Adult: Train, Validation, and Test Accuracy
Adult Training Set
Adult Validation Set
Adult Test Set
Breast Cancer: Train, Validation, and Test Accuracy
Breast Cancer Training Set
Breast Cancer Validation Set
Breast Cancer Test Set
Banknote: Train, Validation, and Test Accuracy
Banknote Training Set
Banknote Validation Set
Banknote Test Set
%Accuracy
Logistic Regression
Naive Bayes
Ionosphere - Learning Rate: 0.0001
Ionosphere - Learning Rate: 0.001
Ionosphere - Learning Rate: 0.01
Ionosphere - Learning Rate: 1
Ionosphere - Learning Rate: 10
Ionosphere - Learning Rate: 100
Breast Cancer - Learning Rate: 0.0001
Breast Cancer - Learning Rate: 0.001
Breast Cancer - Learning Rate: 0.01
Breast Cancer - Learning Rate: 0.1
Breast Cancer - Learning Rate: 1
Breast Cancer - Learning Rate: 10
Breast Cancer - Learning Rate: 100
Banknote - Learning Rate: 0.0001
Banknote - Learning Rate: 0.001
Banknote - Learning Rate: 0.01
Banknote - Learning Rate: 0.1
Banknote - Learning Rate: 1
Banknote - Learning Rate: 10
Banknote - Learning Rate: 100
% Accuracy
Gradient Descent Iteration
Training
Validation
Regularization Grid: Adult
Regularization Grid: Breast Cancer
Regularization Grid: Banknote
Lambda Penalty
Mixing Ratio