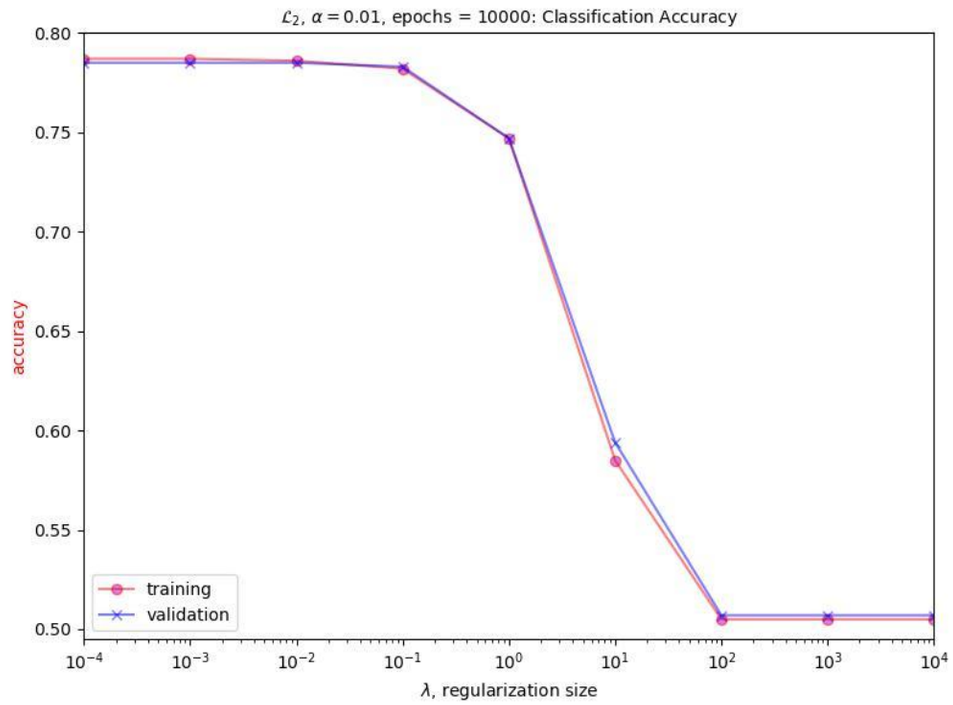
IA2 Report

G10 - Opeyemi Ajibuwa

Part 1: Logistic regression with L2 (Ridge) regularization

(a) What trend do you observe for the training accuracy as we increase λ? Why is this the case? What trend do you observe for the validation accuracy? What is the best λ value based on the validation accuracy?

The training class accuracy remain approximately constant for a while, then generally reduces as we increase λ ∈ [10^-2, 10^4]. This is the case because, increase in λ, directly increase the loss function and hence approximation error. The validation class accuracy also remain approximately constant for a while, then generally reduces as we increase λ within the given range. The best validation accuracy in this range is λ = 10^-2 with an accuracy of 78.54%. This is illustrated in the figures below.



(b) Do you see differences in the selected top features with different λ values? What is your explanation for this behavior?

The resulting top five features with respect to their weight magnitude are illustrated below with λ^{*} = 10^-2, λ_{-}= 10^-3 and λ_{+}= 10^-1. Interestingly, in this range, there is not much difference in the selected top 5 features. However, as the value of λ increases outside this range, the top features gradually begin to differ. This can be interpreted as automatic feature mapping during the regression process. The optimization process automatically associates the most important features with larger weights relative to the other features.

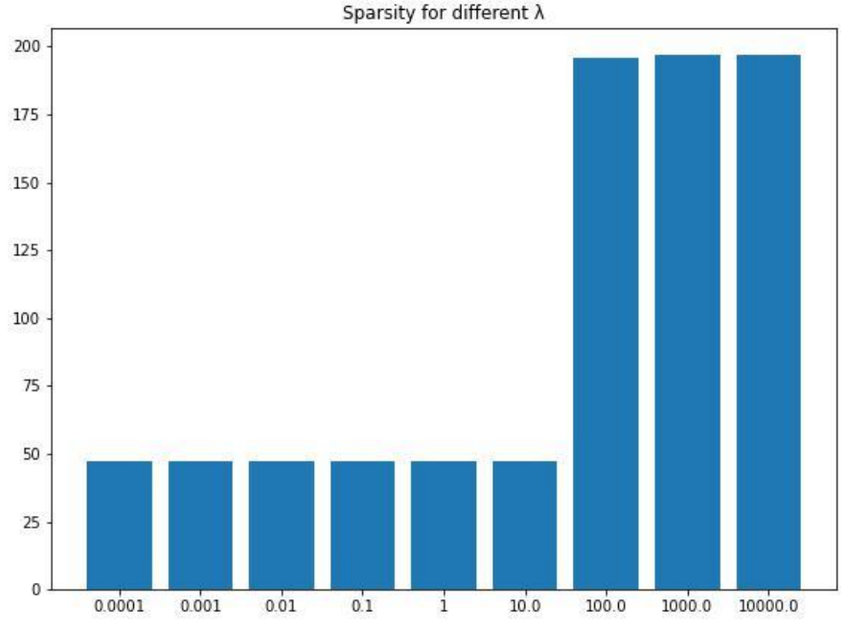
```
Top 5 features and their corresponding weights for lambda=0.001
Previously_Insured         -2.112892
Vehicle_Damage              1.836570
Policy_Sales_Channel_152   -0.687872
Vehicle_Age_1              -0.634887
Policy_Sales_Channel_160   -0.584882
dtype: float64

Top 5 features and their corresponding weights for lambda=0.01
Previously_Insured         -1.583421
Vehicle_Damage              1.466526
Policy_Sales_Channel_152   -0.545381
dummy                      -0.545281
Vehicle_Age_1              -0.495377
dtype: float64

Top 5 features and their corresponding weights for lambda=0.1
Vehicle_Damage              0.651357
Previously_Insured         -0.615228
dummy                      -0.427060
Policy_Sales_Channel_152   -0.250101
Vehicle_Age_1              -0.224794
dtype: float64
```

(c) What trend do you observe for the sparsity of the model as we change λ? If we further increase λ, what do you expect? Why?

It can be observed from the figure below, that as we increase λ, the model sparsity increases. Further increase in λ numerically forces the weights to be very close to zero. When λ reaches 1000, 10000, we observed the model sparsity becomes equal to the number of features, which is 200 in this case.



Part 2: Logistic regression with L2 (Ridge) regularization with noisy data

What are some the key differences do you observe comparing the results obtained using noisy training data to those of part 1? What do you think is the effect of regularization on the model's robustness to noise in the training set? Why?

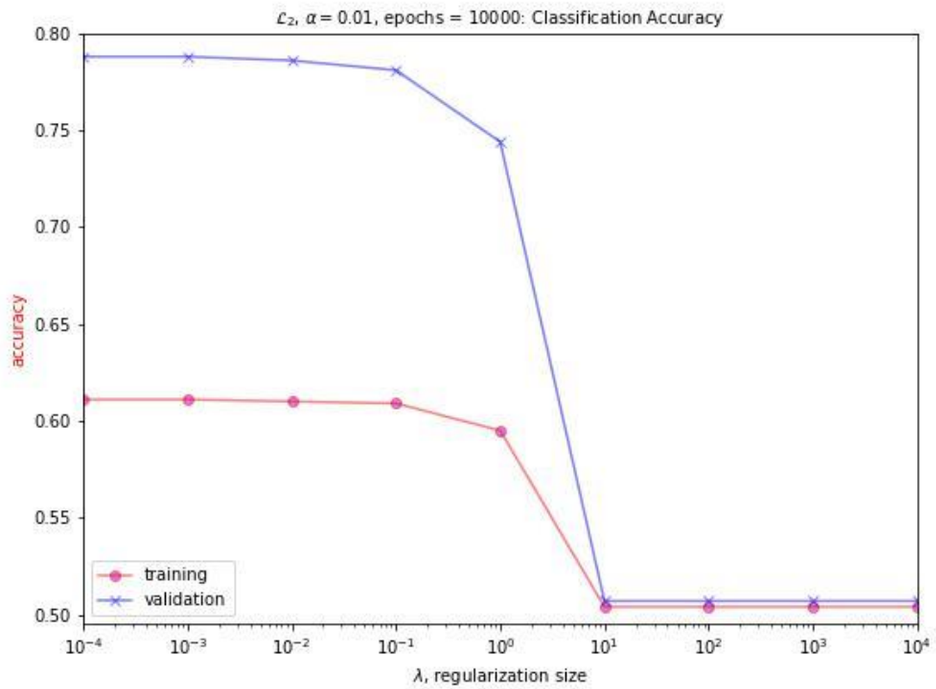
Some of the observed differences between the former model and the one trained with noisy data are:

1. The training accuracy dropped significantly in the model trained with the noisy data compared to the former model. This is expected since about 30% of the labels for the training data have been flipped and thus reducing the accuracy of the predictions of the model.

2. Unsurprisingly, the validation accuracy remains the same and is comparable to what was obtained under the former model. An obvious explanation for this behavior is because of the regularization parameter. The L2 regularizer ensures that our model that does not overfit on the noisy training data so that it can have good generalization performance when tested on new, unseen data.

3. Another notable difference is the decrease in the size of the feature weights learned in the model trained with the noisy data. There is a noticeable reduction in the size of the weights of the top-5 features in the model trained with the noisy data. This behavior can be attributed to the effect of the regularization parameter which acts proportionately to the size of the learned weights.

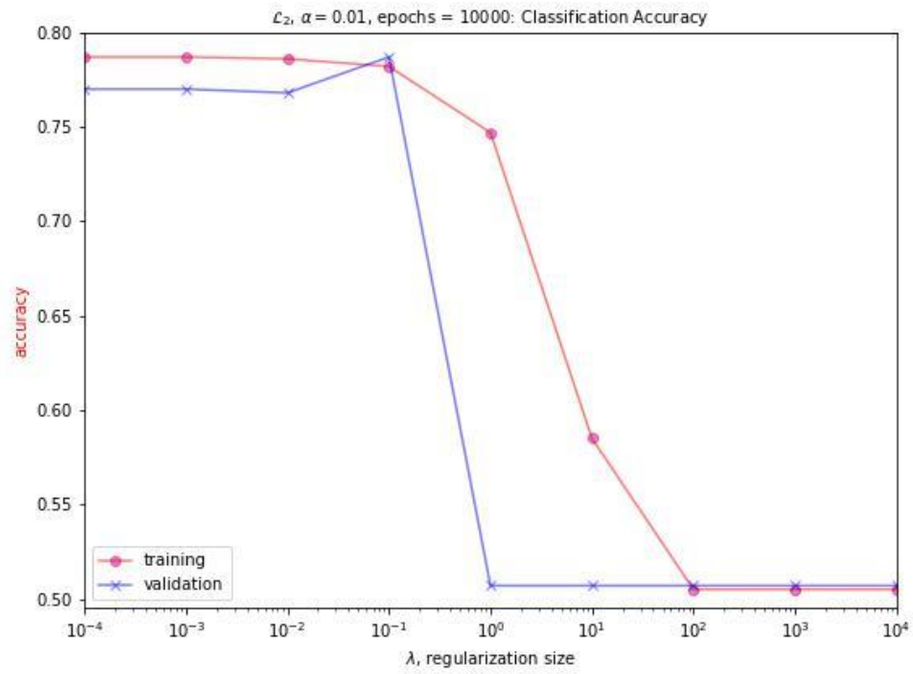
Regularization penalizes the feature weights coefficients of the noisy training instances to prevent our model from picking up the noise or peculiarities or imagine a false pattern where there is none. It does this so that it can keep our model robust and generally improve its generalization performance to unseen data.



Part 3: Logistic regression with L1 (Lasso) regularization

(a) What trend do you observe for the training accuracy as we increase λ? Why is this the case? What trend do you observe for the validation accuracy? What is the best λ value based on the validation accuracy?

The training class accuracy remain approximately constant for a while, then generally reduces as we increase λ ∈ [10^-2, 10^4]. This is the case because, increase in λ, directly increase the loss function and hence approximation error. The validation class accuracy also remains approximately constant for a while (for λ from 10^-4 to 10^-1), then it sharply reduced and stayed approximately constant at this value for the rest of the regularization parameters in these range. The best validation accuracy in this range is lambda = 10^-3 with an accuracy of 76.97%. This is illustrated in the figure below.



(b) Do you see differences in the selected top features with different λ values? What is your explanation for this behavior?

The resulting top five features with respect to their weight magnitude are illustrated below with λ values of 10^-4, 10^-3, and 10^-2. Interestingly, in this range, there is not much difference in the selected top 5 features. However, as the value of λ increases outside this range, the top features gradually begin to differ. This can be interpreted as automatic feature mapping during the regression process. The optimization process automatically associates the most important features with larger weights relative to the other features.

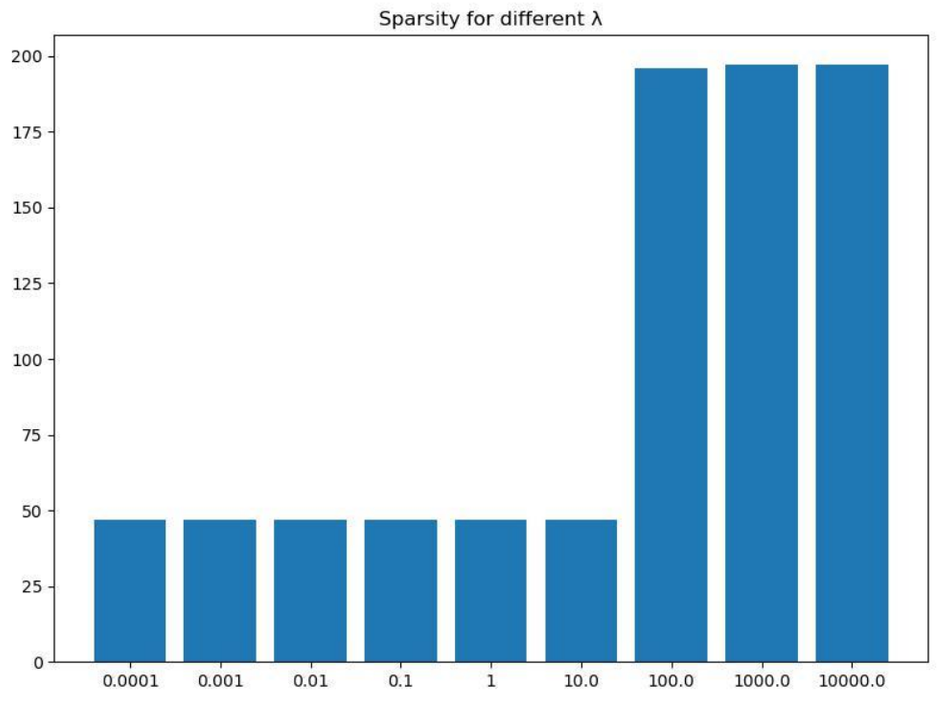
```
Top 5 features and their corresponding weights for lambda=0.0001
Previously_Insured         -2.192550
Vehicle_Damage              1.890561
Policy_Sales_Channel_152   -0.708503
Vehicle_Age_1              -0.655143
Policy_Sales_Channel_160   -0.609874
dtype: float64

Top 5 features and their corresponding weights for lambda=0.001
Previously_Insured         -2.112892
Vehicle_Damage              1.836570
Policy_Sales_Channel_152   -0.687872
Vehicle_Age_1              -0.634887
Policy_Sales_Channel_160   -0.584882
dtype: float64

Top 5 features and their corresponding weights for lambda=0.01
Previously_Insured         -1.583421
Vehicle_Damage              1.466526
Policy_Sales_Channel_152   -0.545381
dummy                      -0.545281
Vehicle_Age_1              -0.495377
dtype: float64
```

(c) For different values of λ , compute the `sparsity' of the model as the number of weights that equal zero and plot it against λ.

As λ increases, the weights becomes sparser; if λ increases, sparsity is expected to increase even more. These trends are similar to what we have seen in 1(c). However, in L1 regularization, the sparsity increases rapidly from λ >= 10^-4, which is faster than L2 regularization. But as λ increases towards infinity, both penalties for weights become too heavy and all feature weights will become 0.



(c) What are the key differences between the two regularization methods observed on this data set? Specifically, which method achieves the best validation accuracy? Which method is more sensitive to the choice of the regularization parameter for this data? Which method produced sparser feature weights? What are the advantages and disadvantages of each method in general?

1. Using the same learning rate of 10^-2 for both regularization types, L1 regularization led to a faster convergence on the training dataset than L2 regularization.
2. L1 regularization is also observed to be more robust on the noisy data than the L2 regularization was. Since L2 regularization takes the square of the weights, so the cost of outliers presents in the data increases exponentially. L1 regularization takes the absolute values of the weights, so the cost only increases linearly.

The ridge logistic regression method has the best validation accuracy. The ridge logistic regression method is more sensitive to the choice of regularization parameter. Lasso regularization produces sparser feature weights than ridge regularization.

The major advantage of ridge logistic regression is coefficient shrinkage and reducing model complexity. Lasso: Along with shrinking the weights coefficients, lasso performs feature selection as well. One disadvantage of ridge logistic regression is the high computational overheads with searching for the choice of an ideal regularization hyperparameter to reduce size of the feature space. A benefit of lasso regression is its ability to drive the feature weights to zero and produce a sparser feature space which correspondingly lead to a computationally efficient regression process than ridge regression. A limitation of lasso regression is that if there are two or more highly collinear features then LASSO regression select one of them randomly which is not good for the interpretation of data.