Submitted by: Oluwasegun Somefun, somefuno@oregonstate.edu

AI534 — IA2 Homework Report Due Oct 29th 11:59pm, 2021

1 Part 1: 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 $\lambda \in [10^{-3}, 10^{3}]$. 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 value in this range is $\lambda = 10^{-3}$. This is illustrated in Figure 1

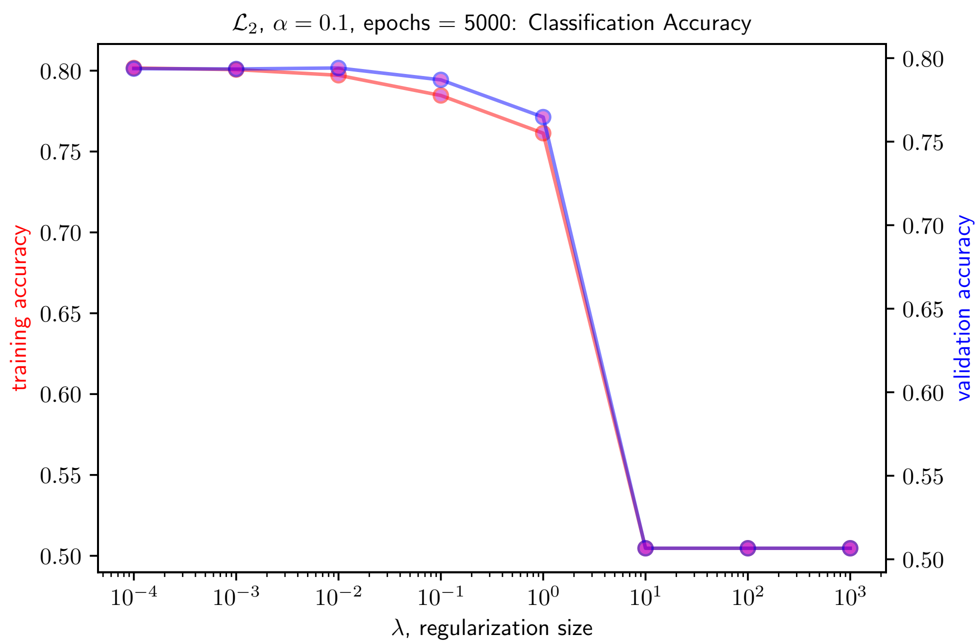

Figure 1: Comparison: class accuracy vs λ

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 illustarated below with $\lambda^* = 10^{-3}$, $\lambda_+ = 10^{-4}$ and $\lambda_- = 10^{-2}$ in Table 1. Interestingly, in this range there is no 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.

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 Figure 2, that as we change λ, the model sparsity increases. Increase in λ numerically forces the weights to be very close to zero. As λ is increased further, we expect the model sparsity to remain constant at a value equal or close to the number of features, which is 197 in this case.

Table 1: $\mathcal{L}_2$: Features (Top-5) with largest $|\omega|$

Feature	10^{-4}	10^{-3}	10^{-2}
Previously_Insured	-3.2264	**-2.9705**	-2.0701
Vehicle_Damage	2.2457	**2.2035**	1.9445
Policy_Sales_Channel_160	-1.8431	**-1.6671**	-1.3914
dummy	-1.0862	**-1.1732**	-0.9030
Policy_Sales_Channel_152	-0.8891	**-0.8424**	-0.6267

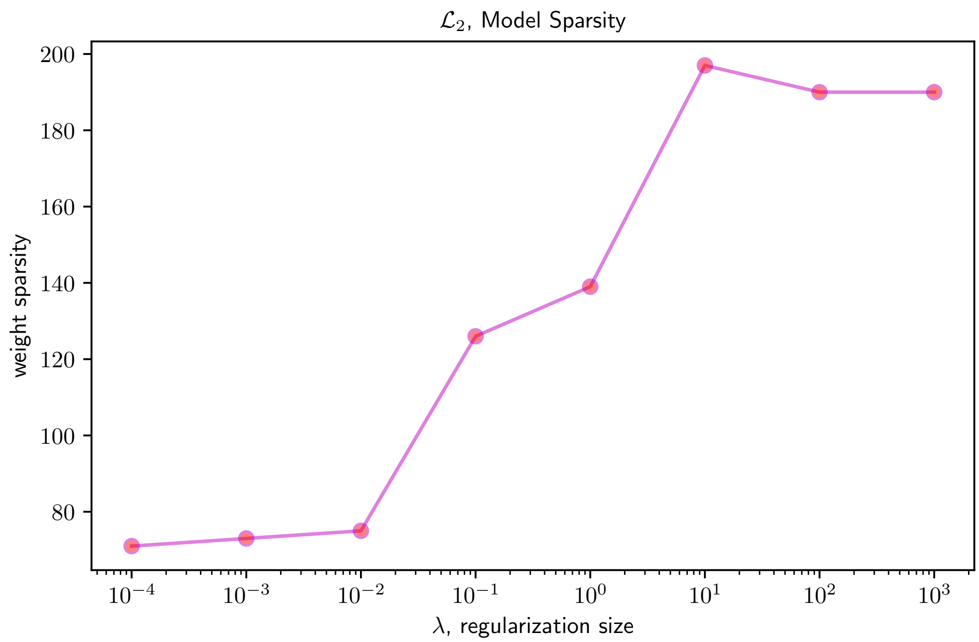

Figure 2: Model Sparsity vs λ

2 Part 2: 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 $\lambda \in [10^{-3}, 10^{3}]$. 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 value in this range is $\lambda = 10^{-3}$. This is illustrated in Figure 3

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 illustarated below with $\lambda^* = 10^{-3}$, $\lambda_+ = 10^{-4}$ and $\lambda_- = 10^{-2}$ in Table 2. Interestingly, in this range there is no 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

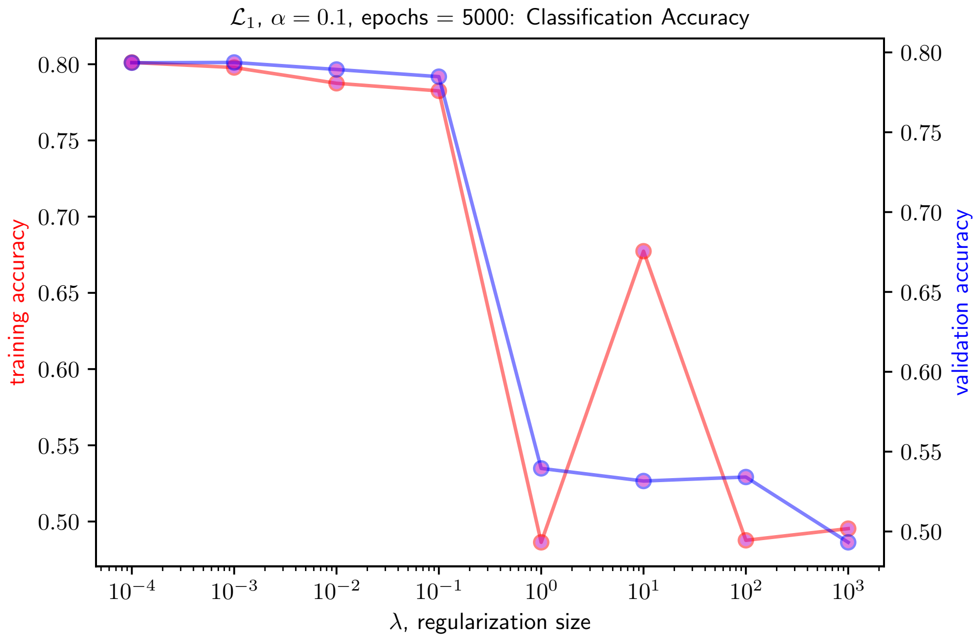

Figure 3: Comparison: class accuracy vs λ

Table 2: $\mathcal{L}_1$: Features (Top-5) with largest $|\omega|$

Feature	10^{-4}	10^{-3}	10^{-2}
Previously_Insured	-3.2506	**-3.1560**	-2.5235
Vehicle_Damage	2.2440	**2.2261**	2.0939
Policy_Sales_Channel_160	-1.8596	**-1.8089**	-1.2236
dummy	-1.1046	**-1.2292**	-1.0126
Policy_Sales_Channel_152	-0.9008	**-0.9347**	-0.7326

automatically associates the most important features with larger weights relative to the other features.

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? Is this trend different from what you observed in 1(c)? Provide your explanation for your observation.

It can be observed from Figure 4, that as we change λ, the model sparsity trend is irregular. We expect that an increase in λ numerically forces the weights to be very close to zero. As λ is increased further, we expect the model sparsity to remain constant at a value equal or close to the number of features, which is 197 in this case. However, compared to the observations in 1c, the expected sparsity trend seems not to be the case. The reason may be due to the discontinous nature of lasso regularization.

3 Part 3: Kaggle Competition

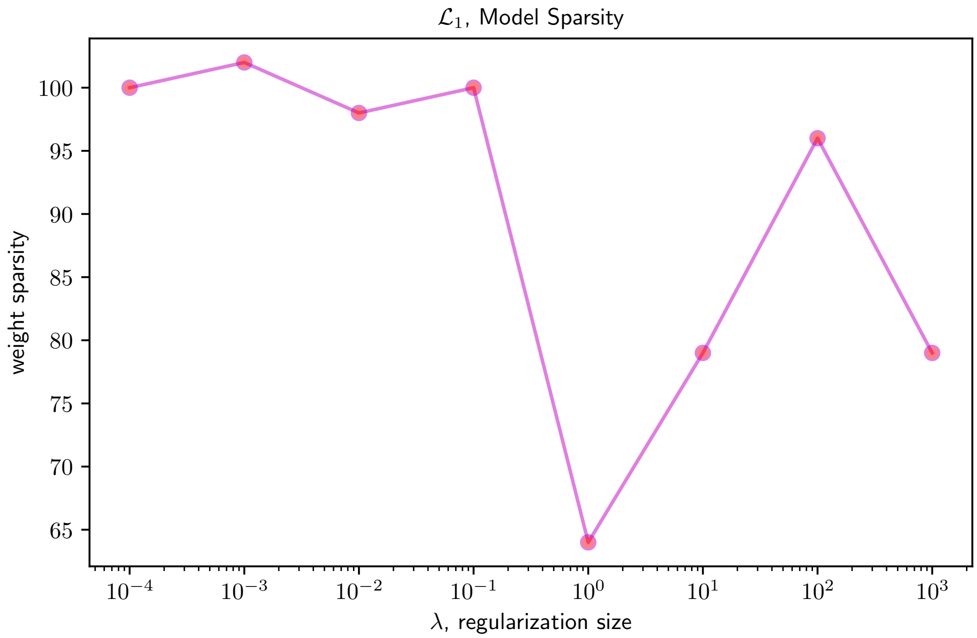

Figure 4: Model Sparsity vs λ

a.) For this part, you need to describe the methods you used and their performance on the public leaderboard. How do you handle the data usage? Do you treat the features differently from what was used for Part 1 and 2? What do you think is limiting the performance you can get on this dataset? The amount of data? the availability of features? the complexity of the algorithm?

The approach to this part is outlined below:

Methods

Learning rate: $\alpha = 0.1$,

Ridge-regularization: $\lambda = [10^{-4},\ 10^{-3}]$

Epochs: $[2500, 10000]$

Training was performed on a combined data of the train and dev datasets. The performance on the test-data seems to increase using the listed set-up above, with a balanced fitting found within the mid-range of the Epochs and λ-value listed above.

Data Preprocessing: The three numeric features in the dataset have low correlation with the **Response** class, so they were dropped.

Comments: Using logistic regression, the limiting performance out of this dataset may be due to the nature of the dataset itself, allowing a small local area of global convergence to a classification accuracy around 80%.