Francis L. Bermillo
bermillf@oregonstate.edu

Implementation Assignment 2: Logistic regression with L2 and L1 regularizations
CS534 Machine Learning

Introduction

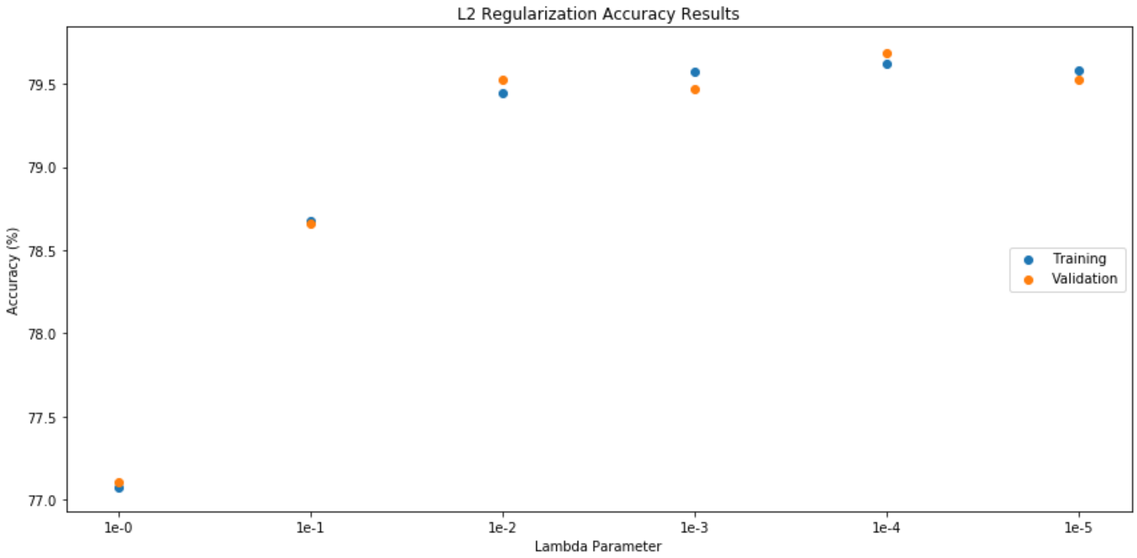
For this assignment, we implemented a logistic regression, which learned from a provided dataset and examined two different regularization methods: L2 (ridge) and L1 (Lasso). Experiments are ran for each regularization method with different lambda values (λ ∈ {10−i: i ∈ [0, 5]}). The results of each method are examined along with the influence of increasing or decreasing the value of lambda (λ).

Part 1 (45 pts): Logistic regression with L2 (Ridge) regularization.

(a) Implement Algorithm 1 and experiment with different regularization parameters λ ∈ {10−i: i ∈ [0, 5]}.

Lambda	Training Accuracy	Validation Accuracy
1e0	77.073	77.106
1e-1	78.674	78.663
1e-2	79.444	79.529
1e-3	79.574	79.467
1e-4	79.620	79.685
1e-5	79.582	79.529

(b) Plot the training accuracy and validation accuracy of the learned model as the λ value varies. 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?



- There is a positive trend in the training accuracy observed from the scatter plot as the lambda parameter becomes smaller. This makes sense because as lambda goes to zero, the less restriction it has on bounding how much the weights should update which may lead to overfitting.
- Similar to the training accuracy, the validation accuracy also has a positive trend as lambda becomes smaller.
- Based on the validation accuracy, the best lambda value would be 1e-4 for this experiment

(c) For the best model selected in (b), sort the features based on $|w_j|$. What are top 5 features that are considered important according to the learned weights? How many features have $w_j = 0$? If we use larger λ value, do you expect more, or fewer features to have $w_j = 0$?

Top 5 Features	
Previously_Insured	-3.09410
Vehicle_Damage	2.12246
Age	-1.21126
Driving_License	-1.04542
Policy_Sales_Channel_53	0.95172

- No features had a weight of 0 because L2 regularization squares the weight values in every update, because L2 penalizes the current weight on how much it changes which means that all weights are most likely going to be less than or greater than zero or close to zero, but not equal zero
- Using a larger lambda wouldn't change the number of zero elements in the weight vector due to the L2 regularization

Part 2 (45 pts). Logistic Regression with L1 (Lasso) regularization.

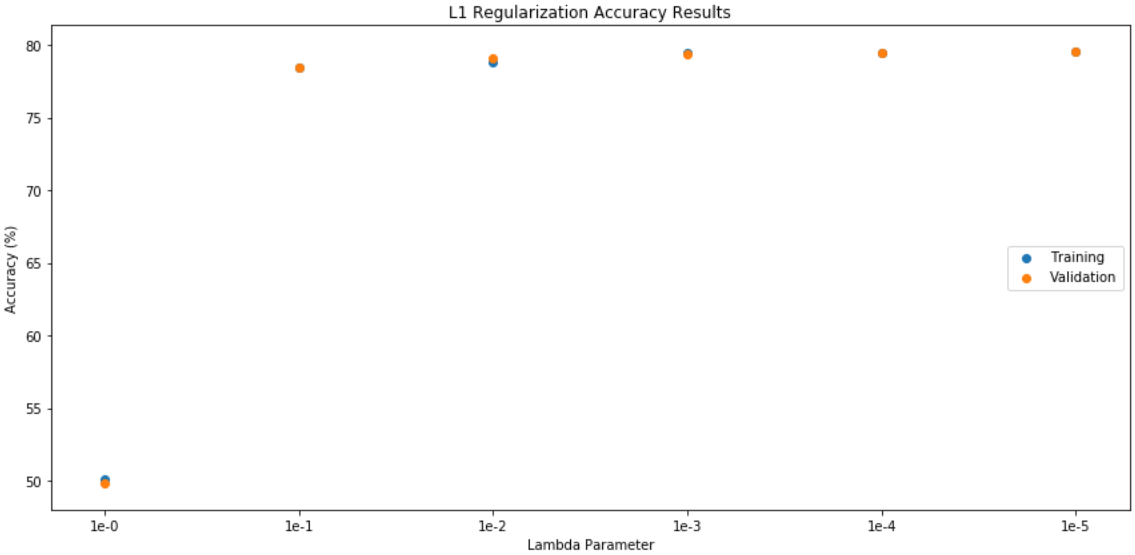
(a) Implement Algorithm 2 and experiment with different regularization parameters $\lambda \in \{10-i: i \in [0, 5]\}$.

Lambda	Training Accuracy	Validation Accuracy
1e0	50.087	49.823
1e-1	78.481	78.423
1e-2	78.856	79.046
1e-3	79.478	79.390
1e-4	79.487	79.458
1e-5	79.563	79.532

Francis L. Bermillo
bermillf@oregonstate.edu

Implementation Assignment 2: Logistic regression with L2 and L1 regularizations
CS534 Machine Learning

(b) Plot the training accuracy and validation accuracy of the learned model as the λ value varies. 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?



- Similar to Part 1, there is a positive trend in the training accuracy observed from the scatter plot as lambda becomes smaller. Again, just like explained in Part 1, as lambda goes to zero, the less restriction it has on bounding how much the weights should update. It can also be observed in this experiment, once lambda becomes smaller than 1e-2, the training accuracy improvements become insignificant.
- Similar to the training accuracy, the validation accuracy also has a positive trend as lambda becomes smaller and has insignificant improvement smaller than 1e-2.
- Based on the validation accuracy, the best lambda value would be 1e-5 for this experiment

(c) For the best model selected in (b), sort the features based on $|w_j|$. What are top 5 features that are considered important according to the learned weights? How many features have $w_j = 0$? If we use larger λ value, do you expect more, or fewer features to have $w_j = 0$?

Top 5 Features	
Previously_Insured	-2.78404
Vehicle_Damage	2.06001
Age	-1.09393
Policy_Sales_Channel_96	0.98315
Policy_Sales_Channel_115	0.98062

- No features had a weight of 0 because for these set of weights due to the size of lambda, however, using a larger lambda would add more feature weights that are equal to zero since satisfying the condition in L1 regularization:

$$\mathbf{w_j \text{ if } | \, w_j \, | < \alpha\lambda \text{ else } w_j \leftarrow 0}$$

  would be much harder due to the larger lambda value

(d) Compare and discuss the differences in your results for Part 1 and Part 2, both in terms of the performance and sparsity of the solution.

- In Part 1 (L2 Regularization), regardless of the lambda value, the accuracy of both training and validation performed well across all tested lambda values. It can be observed that unlike with L2, L1 Regularization from Part 2 performs poorly when the lambda value is large ($\lambda = 1$). However, in both parts, the performance improves as the lambda value gets smaller.
- Unlike L2 Regularization, L1 contained more 0 elements within the feature weights as the lambda value becomes larger, this is due to how L1 calculates its penalty when updating the weights. L1 restricts the weight update to be either 0 or $| \, wj \, | < \alpha\lambda$ which creates for a sparser matrix, on the other hand, L2 only penalizes weight updates that are large which will drive the weight away from ever being 0.