(CS534) Machine Learning

Implementation Assignment 2: Logistic Regression

A-young Kang (ID:933-610-350)

General Introduction.

In this assignment, I implemented and tested logistic regression with two different regularization methods: L2 (ridge) and L1 (Lasso). For both cases, I used learning rate $\alpha = 0.1$ and the upper limit of iterations is 10,000. I measured accuracy as the regularization parameter λ varies. Also, I compared the sparsity of the solutions.

Part 1. Logistic Regression with L2 (Ridge) Regularization.

(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 v? Why is this the case? What trend do you observe for the validation accuracy? What is the is the best λ value based on the validation accuracy?

Generally, for both the training data and validation data, as λ increases, the predication becomes less accurate as can be seen in Figure 1. As the magnitude of the w_j increases, there will be an increasing penalty on the cost function. The penalty is also dependent on the value of λ. Therefore, when λ is large, the gradient descent algorithm tries to avoid the risk of the overfitting more, while when λ is small, the algorithm focuses on fitting the data. Validation accuracy is highest at λ = 0.0001. (Table 1)

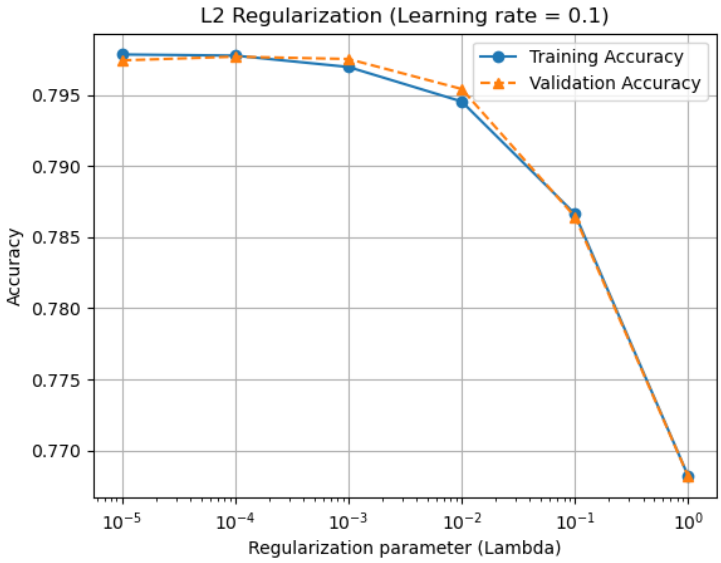

Figure 1. Training and Validation Accuracy with L2 Regularization (Learning rate = 0.1)

λ	0.00001	0.0001	0.001	0.01	0.01	1
Training Accuracy	0.79785	0.79777	0.79697	0.79454	0.78663	0.76819
Validation Accuracy	0.79743	0.79769	0.79753	0.79542	0.78643	0.76817

Table 1. Training and Validation Accuracy with L2 Regularization (Learning rate = 0.1)

(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 with the highest $|w_j|$ are shown in the table 2. For all λ's, the number of features with weight 0 is 3. If we use a larger λ value, more features will have $w_j = 0$ because it will focus on reducing the regularization term, which makes more w_j have 0 weights.

Feature	w_j
Previously_Insured	-3.499864
Vehicle_Damage	2.031472
Age	-1.661466
Policy_Sales_Channel_160	-1.670291
Policy_Sales_Channel_152	-0.79761

Table 2. Features with highest $|w_j|$ (Learning rate = 0.1)

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

(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 v? Why is this the case? What trend do you observe for the validation accuracy? What is the is the best λ value based on the validation accuracy?

Overall trend is similar to L2 regularization. For both the training data and validation data, as λ increases, the predication becomes less accurate as can be seen in Figure 2. As the magnitude of the w_j increases, there will be an increasing penalty on the cost function. The penalty is also dependent on the value of λ. Therefore, when λ is large, the gradient descent algorithm tries to avoid the risk of the overfitting more, while when λ is small, the algorithm focuses on fitting the data. Validation accuracy is highest at λ = 0.00001 (Table 3).

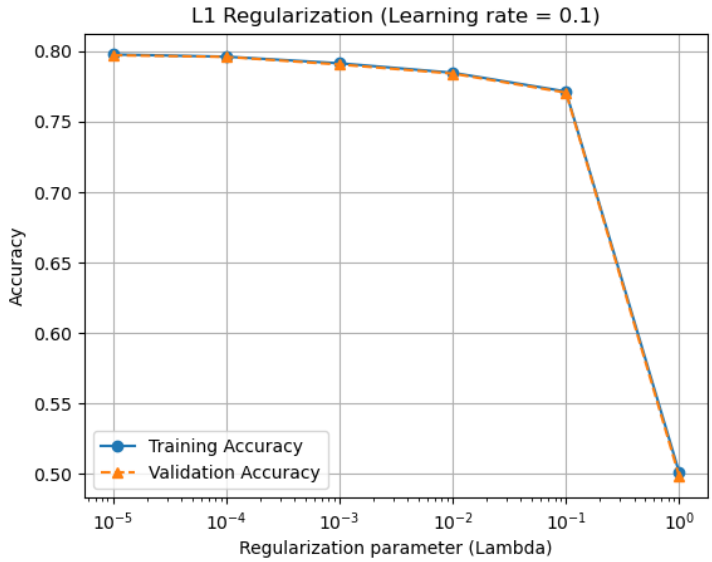

Figure 2. Training and Validation Accuracy with L1 Regularization (Learning rate = 0.1)

λ	0.00001	0.0001	0.001	0.01	0.01	1
Training Accuracy	0.79778	0.79623	0.79154	0.78480	0.77150	0.50087
Validation Accuracy	0.79724	0.79607	0.79065	0.78423	0.77070	0.49823

Table 3. Training and Validation Accuracy with L1 Regularization (Learning rate = 0.1)

(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 with the highest $|w_j|$ are shown in the table 4.

Feature	w_j
Previously_Insured	-3.608381
Vehicle_Damage	2.043292
Policy_Sales_Channel_160	-1.721087
Age	-1.707364
Policy_Sales_Channel_152	-0.81402

Table 4. Features with highest $|w_j|$ (Learning rate = 0.1)

The number of features that have $w_j = 0$ is shown in Table 5. Even if we use a larger λ value, the number of features with 0 weight will be 196 because our $\boldsymbol{w} = [w_0, w_1, \ldots, w_{196}]$ and bias term is not regularized (here, w_0 is a weight for the bias term).

λ	**0.00001**	**0.0001**	**0.001**	**0.01**	**0.01**	**1**
# of features with 0 weight	47	124	176	191	194	196

Table 5. The number of features with $w_j = 0$

(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.

When λ is small enough, the accuracies of logistic regression with both regularization methods are not very different (They are around 0.79). When we use L1 regularization and λ is 1, however, the accuracy is much less accurate. When we apply L2 or L1 regularization to our logistic regression problem, L1 tends to encourage the weights to shrink to zero while L2 tends to shrink weights evenly. This can be seen in our result where when we use L1 regularization, we have more sparse solutions than when we use L2 regularization. Particularly, when using L1 regularization and λ is 1, our algorithm shrinks all the weights toward zero except the bias term, which results in lower training and validation accuracies (0.50087 and 0.49823 respectively).