COMP135 - Project A - Problem 1

Emily Holt, Dylan Phelan

October 2020

1.A Dataset Exploration

	Training	Validation
Total Count	9817	1983
Positive Label Count	4913	1036
Fraction Positive	0.500	0.522

Table 1: Basic composition of MNIST Handwritten Digits data, broken down by training set & validation set

1.B Assess Loss and Error vs. Training Iterations

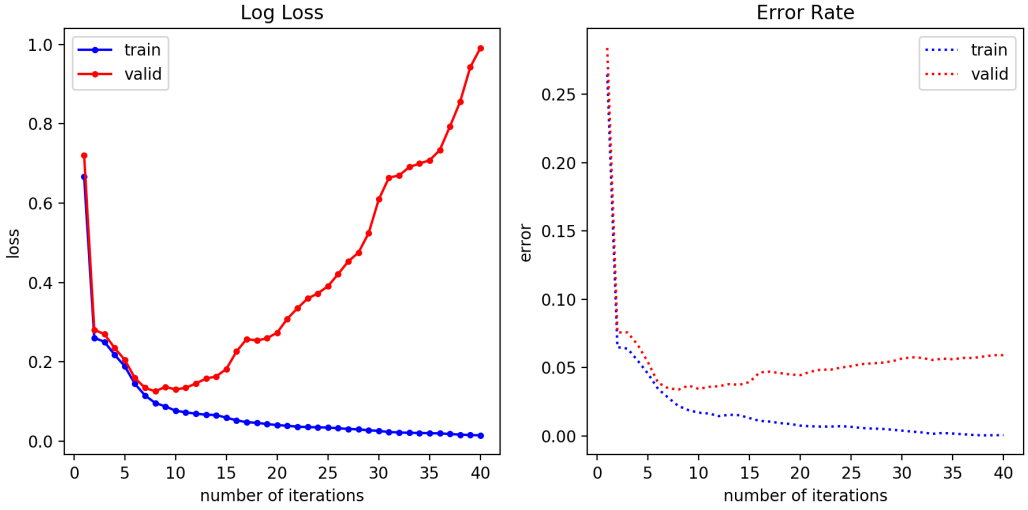

Figure 1: Two charts plotting loss and error as a function of the number of training iterations used in fitting the model. Note the continued reductions of error on the training-set as iterations increase, but the divergence of validation-set performance around 5-10 iterations

1.B Short Answer

These plots illustrate the performance of our models as the number of training iterations increases. We can see that in both the error and log-loss plots, the training error continues to decrease as the

number of iterations increases. However, we observe that between 5-10 iterations the model performs optimally with respect to the validation-set. Beyond 10 iterations we start to see large increases in error and log-loss for our validation set, indicating that our model suffers from overfitting at larger iterations.

1.C Hyperparameter Selection

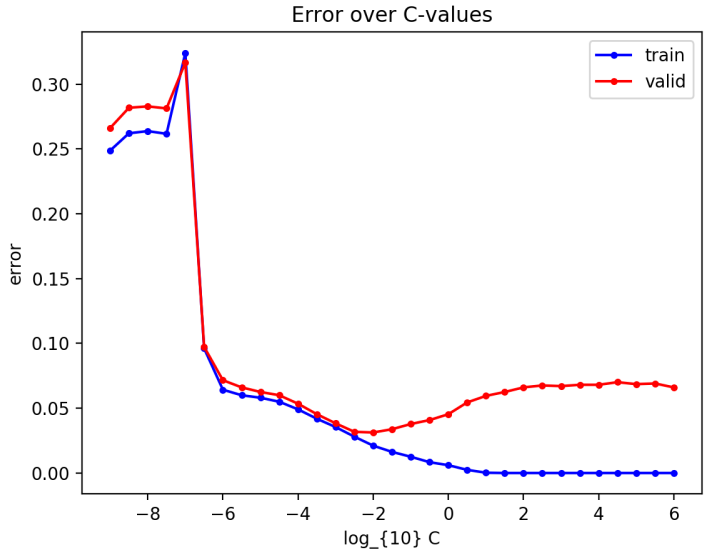

Figure 2: Logistic Regression model's error rate as a function of penalty values; all models trained with $max\_iter = 1000$. Observe that $Log_{10}(c) = -2$ is where we see the lowest error on our validation set. Transforming to our exponential form, the optimal C-value is 0.01 and is the hyperparameter value we should select.

1.D Analysis of Mistakes

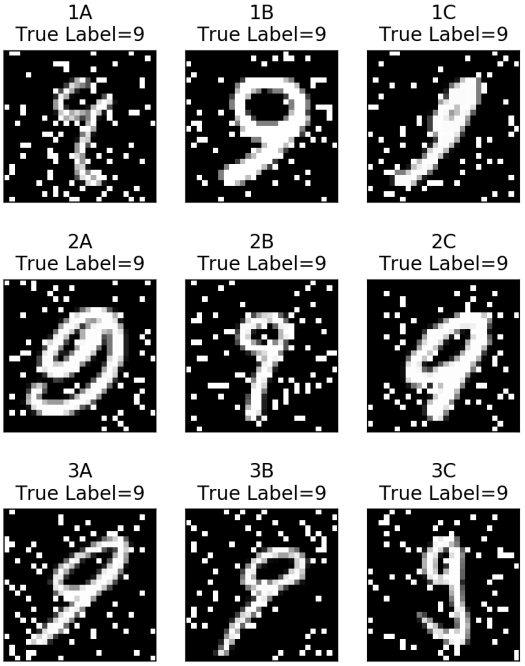

Figure 3: A sampling of falsenegatives from out LR classifier, each labeled with a alphanumeric identifier corresponding to its row-column position in the grid

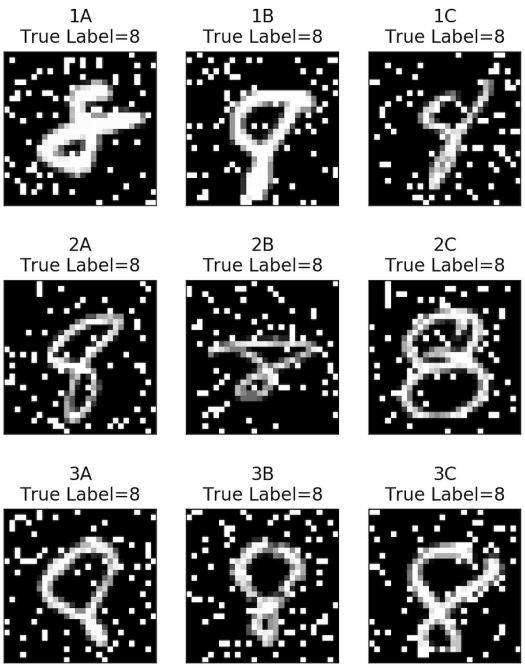

Figure 4: A sampling of falsepositives from out LR classifier, each labeled with a alphanumeric identifier corresponding to its row-column position in the grid

1.D Short Answer

In the case of false positives, the classifier gets confused when the lower loop of the eight is precariously thin/small (i.e. 1B, 1C, 2A, 2B, 3A, 3B) and when there's a gap in one of the loops (i.e. 1A, 1C, 3C).

In the case of false negatives, the classifier gets confused when the down-stroke of the nine is angled on a diagonal (i.e. 1B, 1C, 2B, 2C, 3A, 3B), the down-stroke of the nine ends in a curly flourish (i.e. 1A, 2A, 3C), and when the main loop of the nine is inordinately large (i.e. 1B, 2A, 2C).

1.E Interpretation of Learned Weights

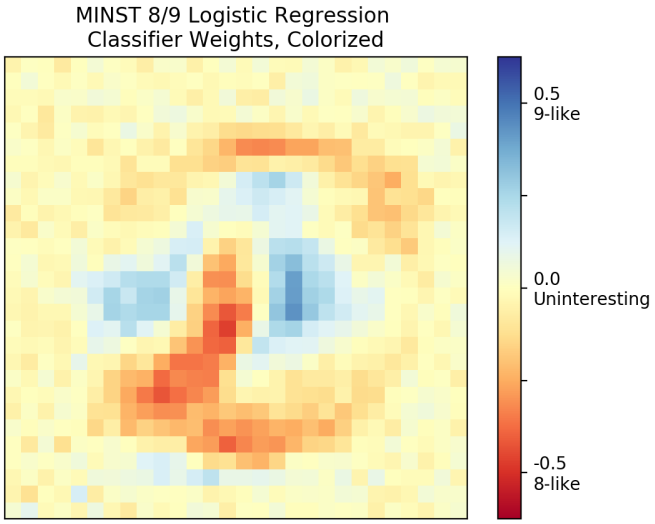

Figure 5: Heatmap of LR-classifier weights

1.E Short Answer

Areas with negative weights include the outer-curves of the image and a sickle-shaped region starting at true center, diagonally stretching down towards the lower left corner, then curving back up and over towards the right side of the image to connect with that upper-, outer-curve.

Areas with high positive weights include the middle-left and middle-right of the image, encased inside the outer-curves with negative weights, as well as a small region just below the head of the sickle in the bottom-left of the image.

With respect to the negative weights, the loops of the eights tend to balloon out, resulting in the outer curves. Additionally, the negative diagonal region is due to the final-stroke of the eight typically cutting straight across looped regions, closing the figure where the initial pen-placement occurs. With respect to the positive weights, the down-stroke of the nine typically cuts through a region of pixels that a loopy eight normally doesn't. Additionally, we believe 9's with a large main loop will have a tendency to touch pixels in the middle-left that, on the final diagonal stroke of the eight, typically aren't used in 8 images.