Regressions on MNIST Dataset

Applied Statistics DSE I1030

Juan Guerrero & Ivan Miller

Fall 2021

Assignment

Part 1: Find a least-squares binary classifier for handwritten MNIST digit set, i.e. determine if an image x is a digit k or not digit k. Pick a digit you like. First, extract the indices of all digit's with label k and randomly separate the samples into equal-sized training and testing groups. Second, do the same for the digits that are not labeled k. Use label $y_i = 1$ if x_i is digit k and $y_i = -1$ otherwise. Find β, and β_0 , such that $y_i \approx \hat{y}_i = \mathbf{sign}\,(\beta^T x_i + \beta_0) = \mathbf{sign}(\tilde{y}_i)$. Note that β can be visualized as a 2D image of the same 28x28 size as the digits in the MNIST data set. Display it. Compute the classification error rate and confusion matrices for the both the training and test sets. Reduce the number of parameters in the β using backward selection methods. Display β for the reduced model.
Use the plot() function to run a series of diagnostic tests for your regression. Plot "Residuals vs Fitted", "Normal Q-Q", "Scale-Location", and "Residuals vs Leverage" plots. (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/plot.lm). Briefly describe what you have observed.

Part 2: Repeat the above problem for all pairs of digits. For each pair of digits, report the classification error rates for the training and testing sets. The error rates can be formatted nicely into a triangular matrix. For storage and display efficiency, store the testing error in the lower triangle and the training error in the upper triangle. Display β and β_0 as a 2D images for pairs with lowest and highest error rates.

Part 3: Use logistic regression and compare your results with Parts 1 and 2.

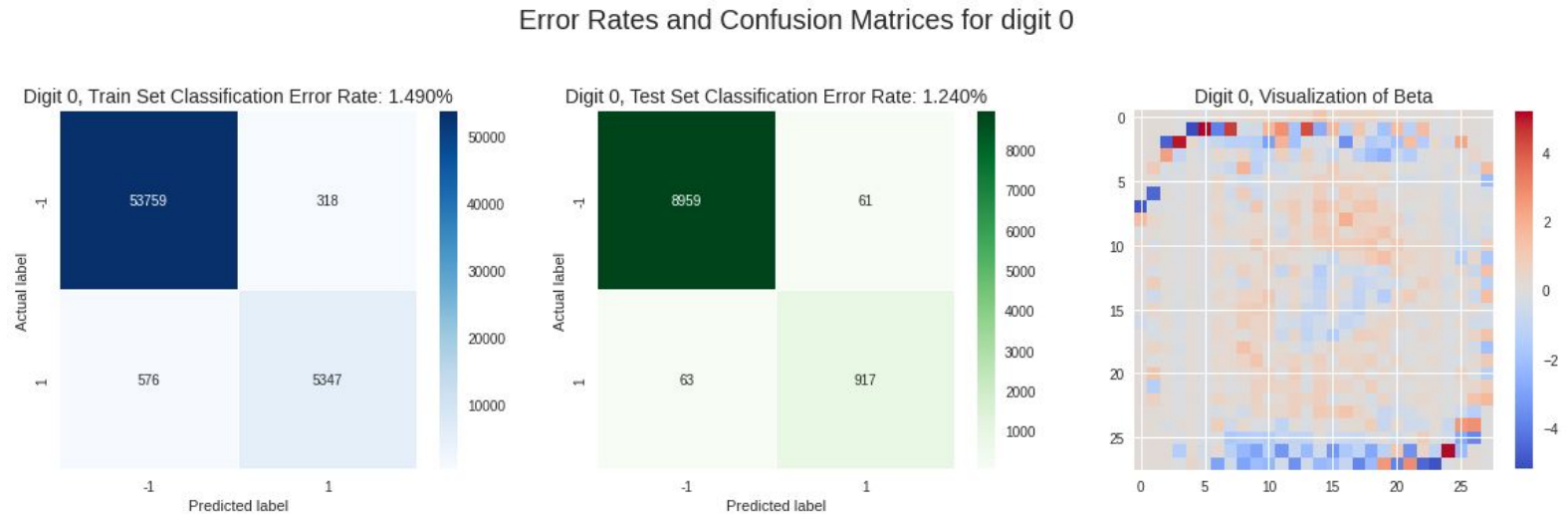
Part 4: Test for outliers in each group of 10 digits using Cook's distance. Visualize the 'mean' and identified outliers for each digit. Repeat Parts 1 - 3 with outliers removed.

Part 1

Recognizing One Digit with Least-Squares Binary Classifier

Least Squares Classifier for Digit 0

Comparison of error rates and confusion matrices between training and test sets after running the Least-Squares Binary Classifier on digit 0.



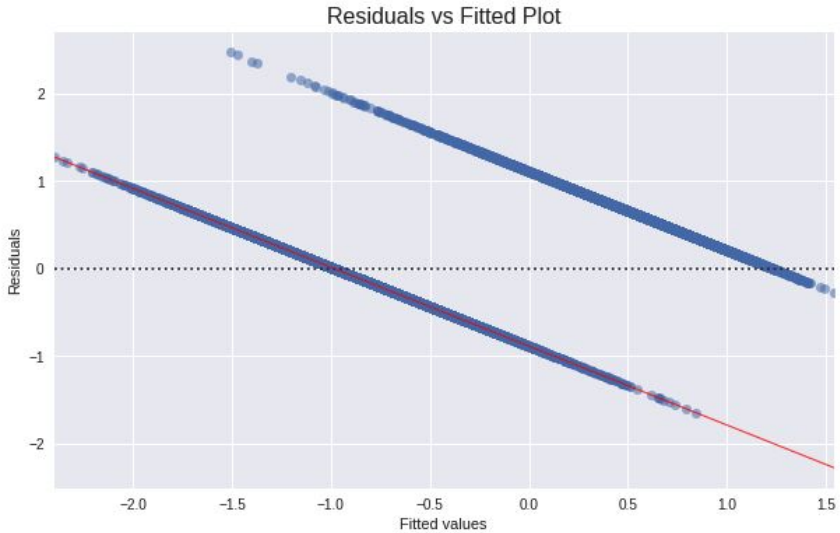
Backwards Selection Method Using a P-Value Approach

- Started off with 784 original number of features
- After executing the code we were able to reduce it to 189 features
- We used a threshold of 0.05

```
def Backward_Selection(X_df, y_df, threshold):

    ols_model = sm.OLS(y_df, sm.add_constant(X_df))
    fitted_model = ols_model.fit()

    max_index = np.argmax(fitted_model.pvalues.values[1:])

    if fitted_model.pvalues.values[1:][max_index] <= threshold:
        return {"columns": X_df.columns, "p-values": fitted_model.pvalues.values[1:]}
    else:
        return Backward_Selection(X_df.drop(columns=[X_df.columns[max_index]]),y_df, threshold)
```

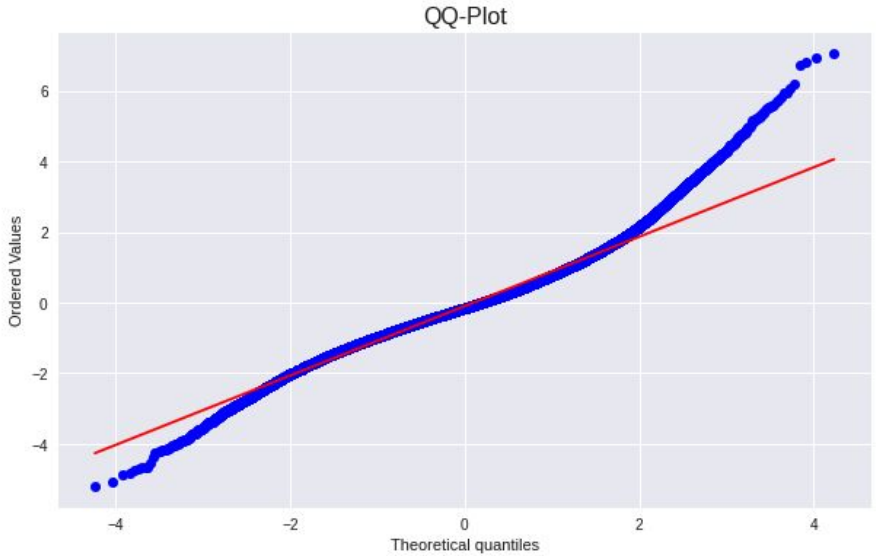
Diagnostic Plots: Residuals vs Fitted

An ideal Residuals vs Fitted plot will look like random noise; there won't be any apparent patterns in the scatterplot and the red line would be horizontal.



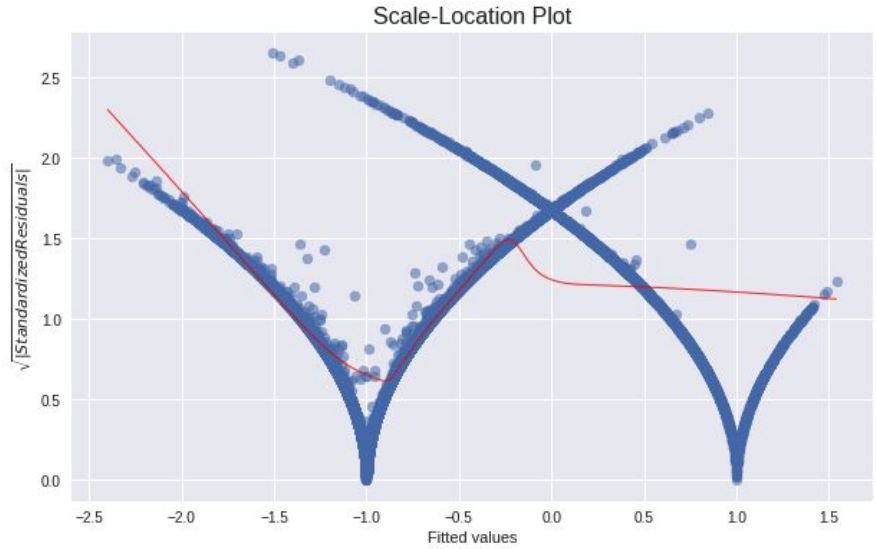
Diagnostic Plots: QQ-Plot

The plot shows the distribution of the residuals. A good normal QQ plot has all of the residuals lying on or close to the red line.



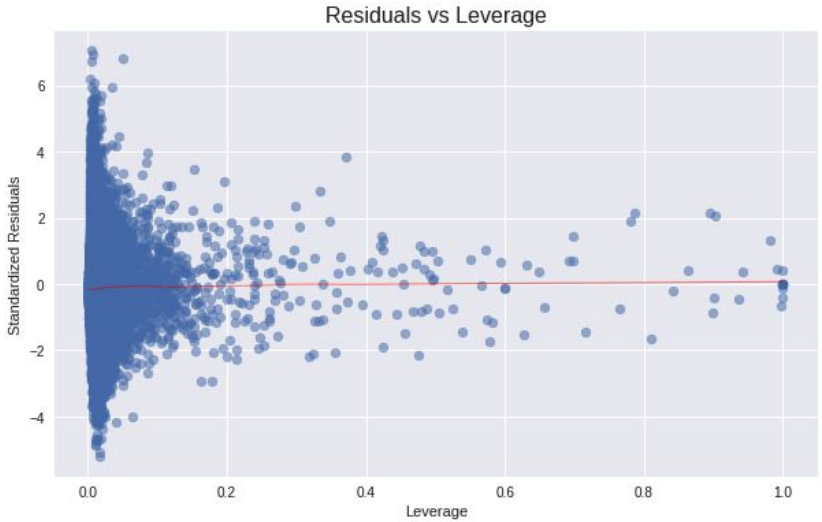
Diagnostic Plots: Scale-Location Plot

This plot is a way to check if the residuals suffer from non-constant variance, aka



Diagnostic Plots: Residuals vs Leverage Plot

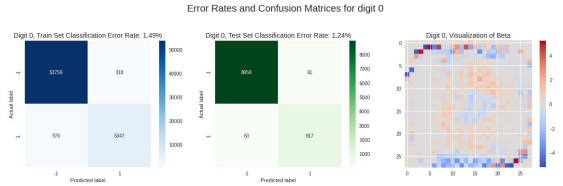
Unlike outliers, which have an unusually large y value, leverage points have extreme x values. This may not seem so bad at face value, but it can have damaging effects on the model because the β coefficients are very sensitive to leverage points. The purpose of the Residuals vs Leverage plot is to identify these problematic observations.

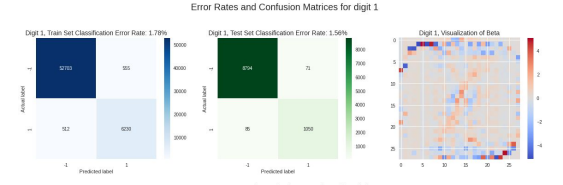


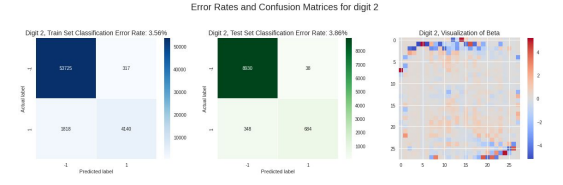
Part 2

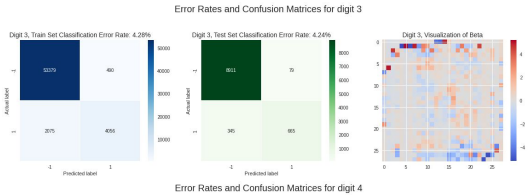
Using Least-Squares Binary Classifier on All Digits

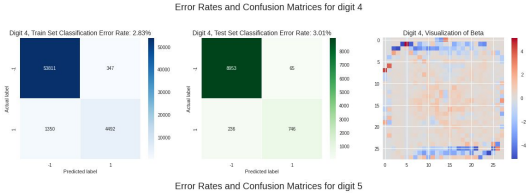
Least Squares Classifier on all 10 Digits

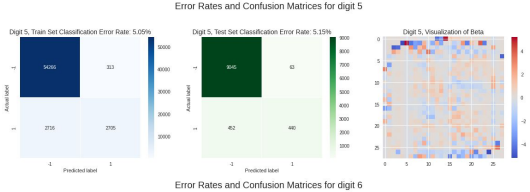


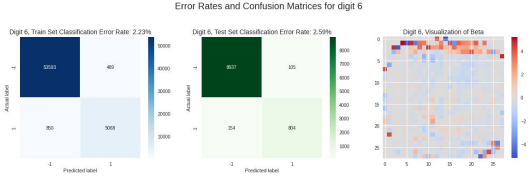


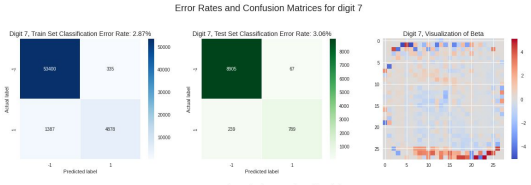


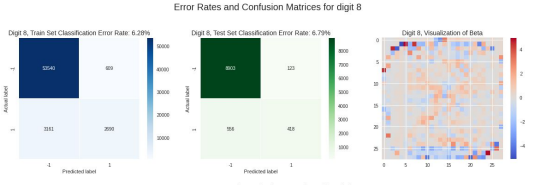


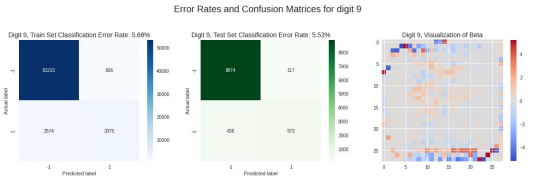










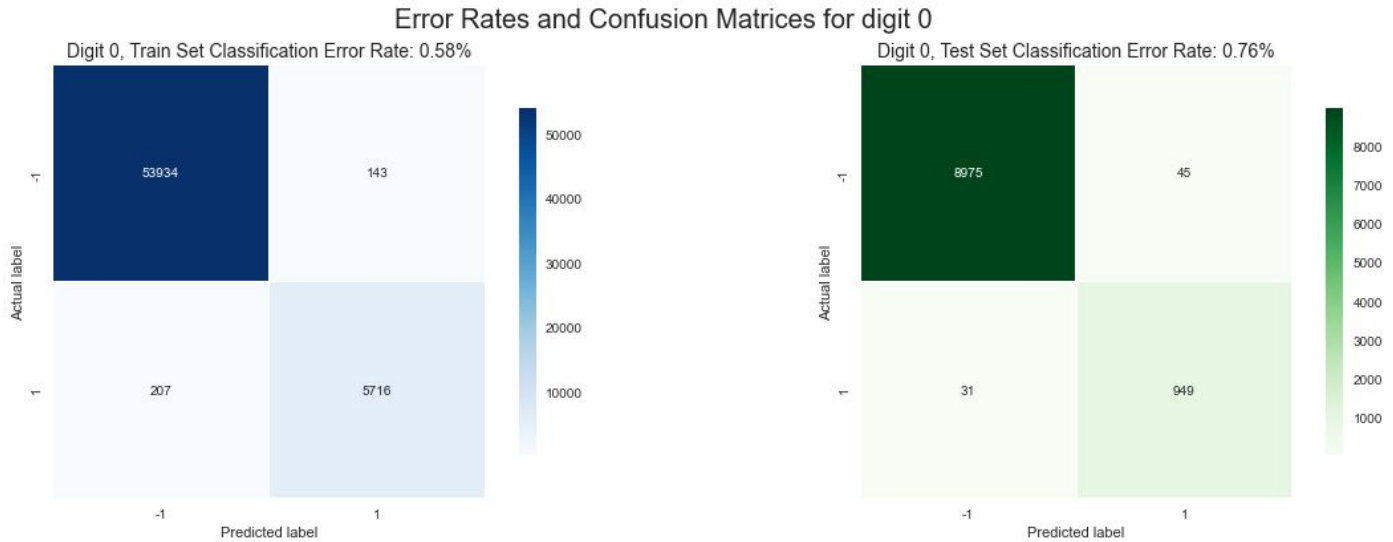


Part 3

Logistic Regression

Running Logistic Regression on Digit 0

Switching to Logistic Regression and running it on one digit as a test, showed lower error rates when compared to the least squares binary classifier on the same digit.



Comparing Error Rates
Between OLS and Logistic Regression

Visualizing error rates for two classifiers on one plot allows us to make a conclusion that Logistic Regression is a solution that is much better suited for the classification problem when compared to the Least Squares binary classifier.

Accuracy Score and Confusion Matrix for All Digits

Visualizing confusion matrix for the whole dataset allows us to identify the pairs of digits that the logistic regression struggles with the most:

- 3 vs 5
- 3 vs 8
- 4 vs 9
- 7 vs 9