## Generative Models On The 0-1 Classes

Using the generative models, Naive Bayes and Gaussian Discriminant Analysis (GDA), we predicted the classes of the 0-1 samples of the test set by training our models on the 0-1 samples of the training set. There are 12665 samples from the training set having a "0" or "1" class, and there are 2115 samples from the test set having a "0" or "1" class. Using only the "0" and "1" classes, our multinomial classification problem became a binomial classification problem. We regarded the "0" class equivalent to "False" and the "1" class equivalent to "True".

### Naive Bayes

Initially, the Naive Bayes model was ineffective due there being 256 possible "bins" for each of the $28^2 = 784$ features for each of the 12665 samples from the training set belonging to the "0" or "1" class. Therefore, for any training sample $(x, y)$ in this set, $P(x|y = 0)$ and $P(x|y = 1)$ were either 0 or approximately 0. Due to rounding error of the program, both probabilities were interpreted as 0, making the model have no predictive power.

However, after transforming the feature values so that they each could fall into one of only two "bins", the Naive Bayes model became much more effective. For each sample $i$ and each feature $j$, $x_j^{(i)} \leftarrow 1$ if $x_j^{(i)} > 127$, and $x_j^{(i)} \leftarrow 0$ otherwise.

The Confusion Matrix for our Naive Bayes model is shown below:

```
Begin
Time Elapsed: 4.578144311904907 Seconds
           ActualTrue  ActualFalse
PredTrue         1130            3
PredFalse           5          977

Process finished with exit code 0
```

From the 2115 samples of the test set, only 8 were falsely predicted.

### Gaussian Discriminant Analysis (GDA)

The GDA model relies on the covariance matrix of the feature values of the training set along with its determinant and its inverse. Due to the large number of samples used in the training set along with the possible size of the feature values being at most "255", the determinant of the covariance matrix of the raw training data was computed to be positive or negative infinity. To fix this, we used a Z-score normalization on the feature values of the data. This transformation made our determinant finite.

After the z-score transformation, we had to deal with the issue of having a zero determinant, which resulted in our covariance matrix being singular. To make our covariance matrix $\Sigma \in \mathbb{R}^{dxd}$ become invertible, we removed rows and columns that were approximately 0 in $\mathbb{R}^d$. Thus, for a fixed $q \in \mathbb{R}$, we kept the components $i$ of the covariance matrix $\Sigma$ for which $|\Sigma[i,:]| > 10^q$. We tried various values of $q$. If $q$ was too small, our covariance matrix would still have a zero determinant. However, if $q$ was too large, we would remove all components of $\Sigma$, resulting in an empty matrix. We found that $q = 0.5$ was a value that worked to get a nonsingular, nonempty covariance matrix. After finding a matrix $\overline{\Sigma}$ that approximated the original covariance matrix $\Sigma$ which had a nonzero, finite determinant, we were able to predict the classes of the test set well.

The Confusion Matrix for our GDA model is shown below:

```
Begin
Time Elapsed: 4.961873531341553 Seconds
           ActualTrue  ActualFalse
PredTrue         1133            5
PredFalse           2          975

Process finished with exit code 0
```

From the 2115 samples of the test set, only 7 were falsely predicted.

### Model Comparison

Both of the models did very well. The GDA model did slightly better predicting the "1" classes as "1" classes, whereas the Naive Bayes model did slightly better predicting the "0" classes as "0" classes. However, the differences are negligible when considering the size of the data set that was tested on.