BME 2211: Predicting Mask Type for a Person in the COVID Testing Line

Andrew Sifferlen, Amelia Harvey, Madison Sanborn, Grace Jolin, Khushi Soni

Methods

Data Collection

To collect our data, we observed WPI students and faculty members waiting in the COVID testing line outside of the Harrington Auditorium. Each of our group members were responsible for collecting twenty observations. Each observation was one individual person who was observed for 5 minutes. Information was collected on the date, time, and observation number of each person. For each person we observed the type of mask they were wearing, how many times they touched their face, and whether or not they were on their cell phones, wearing headphones, or talking to someone else in line. The data collected was recorded in an Excel spread sheet. The data on cell phone and headphone usage as well as if the person was talking with someone else was reported in logical formatting, with yes being denoted as a 1 and no being a 0. The type of mask was represented as 1 (cloth) or 0 (disposable). We collected a total of 101 observations with 7 features. When looking at our data we had a good range of values for the times our data was collected, 9 am to 3 pm. However, we did not have as good of a range of values for the amount of times a person touched their face, most people never touched their face while in line and the amount only ranged to a max value of 8. We believe the amount of times a person would touch their face was typically so low because people are cognizant of being careful not to touch their face when they are on campus and especially cognizant of this when in the COVID testing line. Originally, we had hoped there would be a larger range of number of face touches to produce more variation for predictions.

From this data we wanted to determine if the amount of times someone touches their face and the time of day that they receive their test is indicative of the type of mask they are wearing, or if the type of mask they are wearing could predict how many times they touch their face.

Data Analysis

To analyze our data we used the time of day someone was in line and the amount of times they touch their face to predict the class labels for the type of mask they were wearing a cloth (1) or a disposable/surgical mask (0). To do this we performed classification on the data. To do this we used MATLAB's internal function 'classify' to find the weight vector and threshold value. After finding these values we plotted a histogram of the data points projected over the projection vector. If the classification model was a good fit for this set of data, there would have been two peaks to either side of the threshold value. However, our histogram of data points projected over the projection vector had only one peak and was right skewed.

Even though a classification model may not have been the best fit for this data we continued to use the classification model. We started by classifying the points as above or below the threshold value we then plotted the class predictions from our model overlaid on the original data. In this plot (Figure 1) the x's were the original data values in blue and red to represent the different data classifications and the o's were the class predictions from our model, so that a blue x was a person who wore a disposable mask, and a blue o represented a prediction for a disposable mask and a red x represented a person who wore a cloth mask and a red o represented a prediction for a cloth mask. Using this method any circle that mismatched the color of the x inside represented an incorrect prediction and matching colors represented a correct prediction.

To determine the goodness of fit of our model we determined the true "positive," false "positive," true "negative," and false "negative" amounts. In this model a "positive" is wearing a cloth mask and a "negative" is wearing a disposable mask. Using those numbers, we calculated accuracy, sensitivity, and specificity of our model. Accuracy is a measure of correct predictions over total data. Sensitivity is a measure of correct predictions for wearing a cloth mask compared to all people who actually wore a cloth mask. And specificity is a measure for correct predictions of people wearing a disposable mask compared to all people who actually wore a disposable mask.

Results

The classification technique was not able to model the data incredibly well. After running the 101 data points through the analysis, there were 29 True Positives, 9 False Positives, 43 False Negatives, and 20 True Negatives (Here, like mentioned above, "positive" means wearing a cloth mask, and "negative" means wearing a disposable mask). And just like in class false positive or false negative means they were incorrectly classified. From this, the accuracy was calculated to be 48.51%, the sensitivity of the data was calculated to be 40.28%, and the specificity of the data was calculated to be 68.97%. Considering the model accuracy was below 50%, it is fair to say that our model did not classify the data well enough. Looking at the figures below, Figure 2 shows a histogram of the data points projected onto the weight vector. The threshold value that was calculated from the data was 0.5001, but as the graph shows, the peaks are not equally distributed on either side of the threshold value, as they would be in a better model, and the values itself are right-skewed. Looking at Figure 1, while the technique was able to correctly predict many people who would be wearing a cloth mask (red crosses in red circles), it is clearly visible that it was not able to predict people who would be wearing a disposable mask, and it incorrectly predicted many other points as well. Also, by looking at the graph, it is clearly visible that there is no real distinction between the two classifications (disposable versus cloth mask), which again shows that the technique used was not the best to model our data.

The reason our techniques were not able to model our data well was because there was not much variation in the data itself. Because we only observed people for five minutes at a time, it was difficult to produce a variation in the number of times people touched their face. Additionally, thinking back the COVID line is one location people are probably cautious to not touch their face which could be why we did not have high variation in number of face touches. If we were to redo the data collection, we would have collected another feature that varied with time and was not based on labels like the bulk of our features were. We also would have tried to collect more data on one person for a longer time, so we could have possibly counted a greater number of face touches. Another way could be to collect data on the same people on different days

Figures

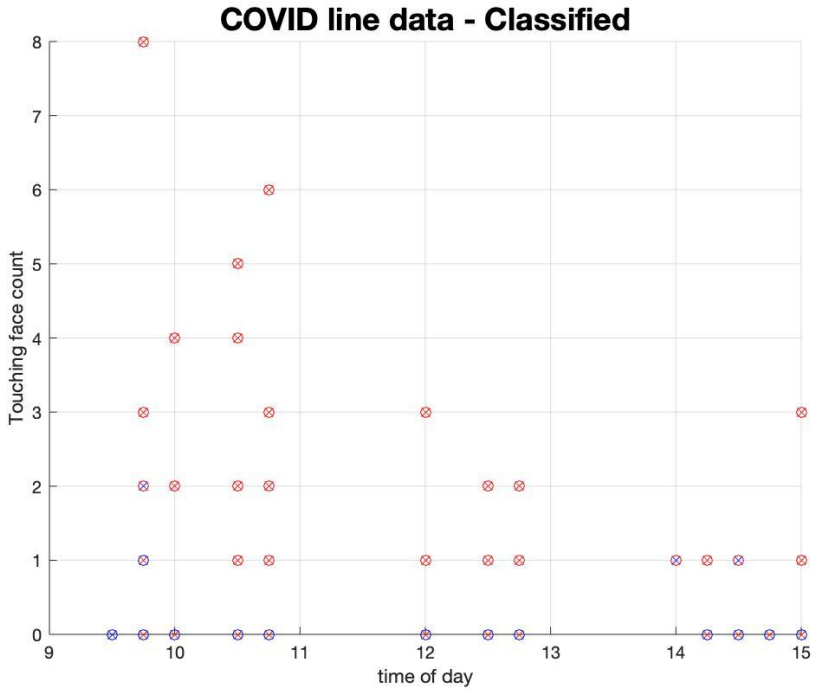

Figure 1. This graph shows the result of the threshold classification on the data points and the actual data points. The classifier utilized a weighted vector and a threshold value to evaluate the predicted response values y_hat. Red x's represent the actual data values for a person in a cloth mask and a red o represents a prediction for a cloth mask based on the model. Similarly, a blue x represents the actual data for a person in a disposable mask and a blue o represents a prediction for a disposable mask. x's and o's that match represent correct predictions and mismatches represent incorrect predictions.

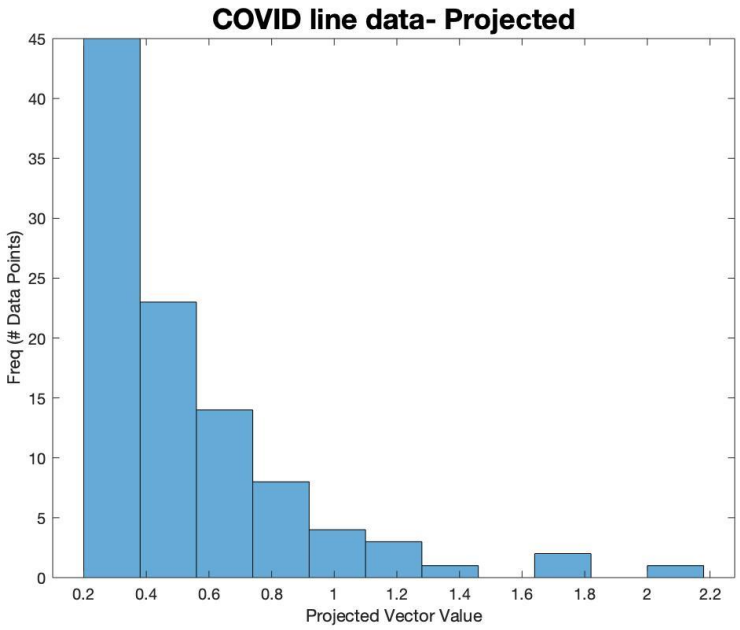

Figure 2. This figure depicts the spread of data points over u, which is the projection of X over the weighted vector w. The graph is right-skewed, and there is only one peak, showing that there wasn't enough of a variation of collected data to fit a classification model that uses a threshold value.