Name: Jason Zhou
Mentor: Dr. Dongjin Song
Status Report #: 11
Time Spent on Research This Week: 2.5
Cumulative Time Spent on Research: 76
Miles Traveled to/from Mentor This Week: 0
Cumulative Miles Traveled to/from Mentor: 0
==========

Monday, November 22nd, 2021:

Although my mentor and I usually have our weekly meeting on Mondays, my mentor canceled because UConn staff and students were on Thanksgiving Break.

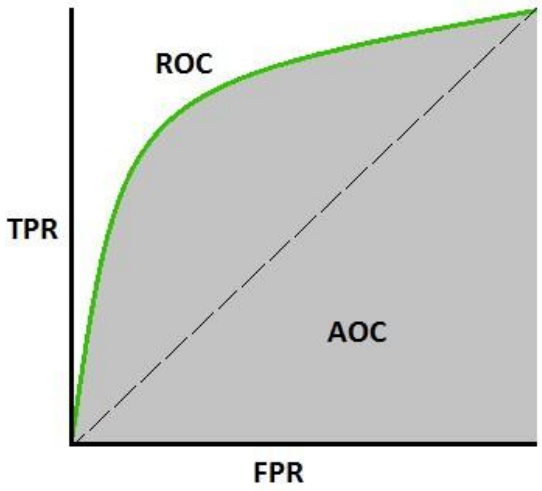
Monday, November 29th, 2021:

I had to cancel this week's Zoom meeting because of a family emergency.

Saturday, December 4th, 2021: (1.5 Hours)

Over the last two weeks, I have not had much time to dedicate to ARM due to wrapping up college applications and supplements for the December 1st deadlines. However, I made it a priority to spend time on ARM as I did not want two weeks to go by without any progress.

On this day, I mainly focused on learning the AUC-ROC curve. This was something I looked into previously but only had a basic understanding of it, so I took a deeper delve into the concept.



(A picture of the AUC-ROC curve, which is a graph that has True Positive Rate, TPR, as the y-axis and False Positive Rate, FPR, as the x-axis. The ROC is the green curve, while the AUC is the area under the green curve)

On a basic level, the AUC-ROC curve works to measure how well a machine learning model classifies items into categories. On a deeper level, it does this by measuring the number of True Positives and False Positives the model has. In the context of anomaly detection, a positive would be all data points that are considered anomalies, and negatives would be all data points that are considered normal. Thus, a true positive is an anomaly that was correctly predicted. A False positive would be defined as a data point that was predicted to be an anomaly but was actually normal. TPR is how many anomalies the model correctly predicts given the total amount of anomalies. FPR is the number of normal data points the model incorrectly classifies. Thus, naturally, one would want a model that has a high TPR and a low FPR.

Usually, the ROC curve is useful for determining the best threshold for a certain method. A threshold is a value or percentage at which the model will accept a data point as an anomaly or not. For example, if the datapoint has a 49% chance of being an anomaly, and the threshold to actually consider it as an anomaly is 50%, then because the percentage is below the threshold, the data point would be considered as normal.

A ROC curve represents the performance of a single method. Therefore, if I had more than one ROC curve, each curve would represent a different method. However, how would I determine which curve/method is the best one? Although one could compare the various TPR's and FPR's, the easier method is just to calculate the AUC or the area under each curve. The greater the area, the better the method is compared to the rest.

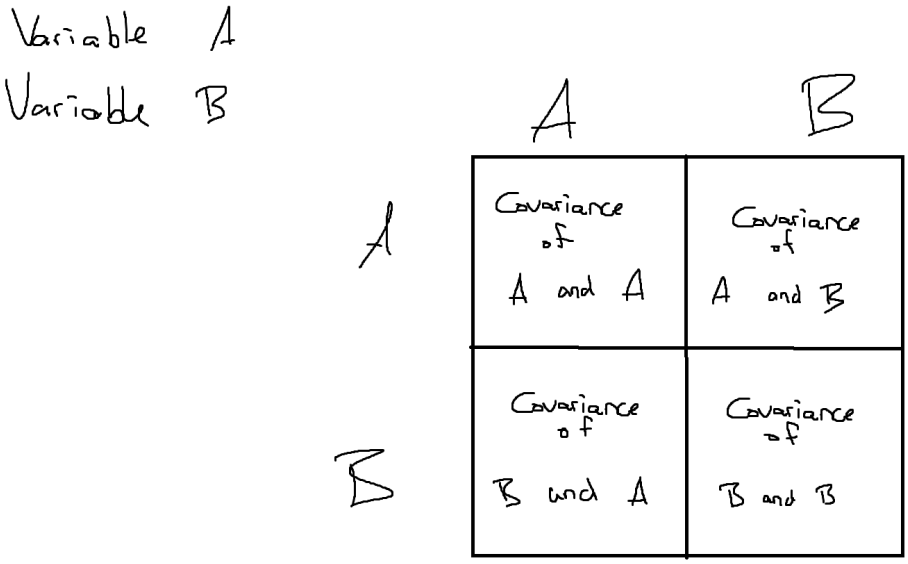
November, December 5th, 2021: (1 Hour)

For the final day of the week, I decided to look into a few concepts that I had seen within machine learning but did not quite understand. Specifically, I investigated variance, covariance, and covariance matrices.

Variance is how different a data point is from a given mean, which is similar to standard deviation. The only difference is that variance is simply the square of standard deviation.

Covariance is the relationship between two different variables. This is similar to correlation. The only difference is that Covariance can be any number from negative infinity to infinity, while correlation is a number from -1 to 1. Also, I believe that at a fundamental level, correlation is simply a function of covariance.

Finally, a covariance matrix is simply an array of information that stores values of covariance for two variables, which can be set up as a table like this:



(An example of a covariance matrix using variables A and B. These variables are usually a collection of data points/numbers. For reference, each box within the large square is called a cell)

The covariance for each block is calculated using this formula:

$$Cov(X,Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right]$$

An alternative expression of Cov(X, Y)

$$Cov(X,Y) = E\left[\{X - E(X)\}\{Y - E(Y)\}\right]$$

$$= E\left[X.\{Y - E(Y)\} - E(X).\{Y - E(Y)\}\right]$$

$$= E\left[X.Y - X.E(Y)\right] = E(X.Y) - E(X).E(Y)$$

(Formula to calculate the covariance for a cell in the matrix. E stands for the expected value, which is another way of saying the average. In other words, the equation E(X*Y) - E(X)(Y) means to take the average of the product of X and Y minus the average of X times the average of Y. Keep in mind that X and Y are usually a collection of numbers, so taking the mean of X, for example, would mean taking the average of all values stored inside X)

References

ritvikmath. (2019, September 20). *The covariance matrix : data science basics* [Video]. YouTube. https://www.youtube.com/watch?v=152tSYtiQbw

ritvikmath. (2020, November 9). *The ROC curve : data science concepts* [Video]. YouTube. https://www.youtube.com/watch?v=SHM_GgNI4fY&t=879s

Starmer, J. (2019, July 11). *ROC and AUC, clearly explained!* [Video]. YouTube. https://www.youtube.com/watch?v=4jRBRDbJemM