# Linear Data Analysis
# Classification - Assesment With ROC Curve

Cain Susko

Queen's University
School of Computing

March 16, 2022

# a Reciever Operator Charaterisitc

Reciever Operator Charaterisitc (ROC) is a curve for scoring data. We do this by finding the Area Under the Curve or (AUC).
An early use of the ROC included radar threat detection in wartime. The Reciever is the Operator of the radar; the Operator is the human who interpereted the display; and the Characteristic is the evaluation of human performance.

# b ROC And The Confusion Matrix

This section relates the confusion matrix to ROC. we have encountered an Absolute confusion matrix:

| | +1 | -1 |
|---|---|---|
| +1 | TP | FN |
| -1 | FP | TN |

but we can also make a **realtive** Confusion Matrix.

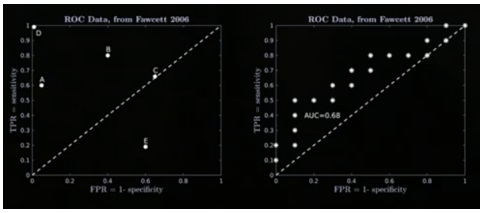
| | +1 | -1 |
|---|---|---|
| +1 | TPR | FNR |
| -1 | FPR | TNR |

which uses the rates that can be derived from the Absolute Confusion Matrix. Note: the sum of the first row should be 1 as well as the second row. In this relative Matrix, there are 4 entries and 2 contsants which implies that there are 2 degrees of freedom. This means that we can use 2 variables from the the relative confusion table in order to plot the entire entire table.

**ROC as 2D Measure** Thus, the ROC is a 2D measure thar uses (FPR, TPR). Once we plot this measure, we can change the threshold we're using $\Theta$ and plot these different values of $\Theta$. When one plots all values of $\Theta$, then any point on the curve plotted specifies a relative confusion matrix and any relative confusion matrix specifies a point on the ROC curve (for a given system)

### Example

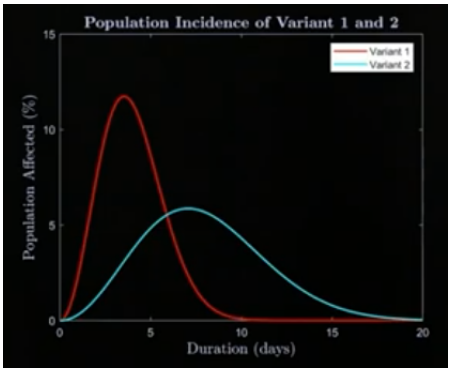
Consider the following plots where 5 Relative Confusion Matricies of Classifiers are plotted (left) and where 1 point is iterated over time to increase the threshold for classification $\Theta$




## c ROC Curve of Ficticious Virus

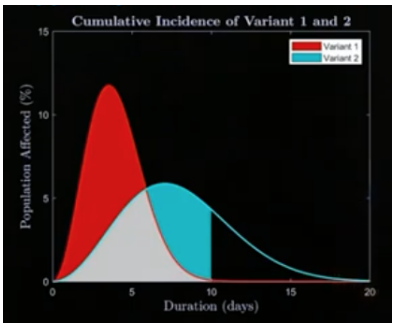
1. variant of a virus infects most people quite promptly

2. variant infects most people slowly and eventually



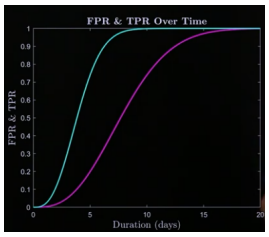

Say we want to be able to determine if someone has a variant while only testing for the second variant.

**Solution** The cumulative detection of the virus is equal to the area under the curve. The number of days after the onset of the virus is $\Theta$ where the time someone has had the virus determines which variant they have.

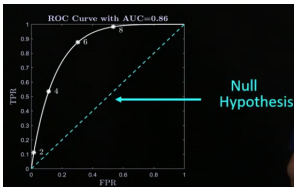



Note: red = variant 1, blue = variant 2
Observe that the area under the red curve but above the blue curve are true positives for variant 1 and the area under the blue & red curves represents the false positives for variant 1.

**Vary** $\Theta$ We shall vary the hyper-parameter $\Theta$ from $0:20$. For small $\Theta$: FPR $= 0$ and TPR $= 0$ where FPR is magenta and TPR is cyan.




we can now find the ROC Curve where the independent variable $\Theta$ has the dependent variables FPR and TPR. The area under the curve is usually $0.5 \leq AUC \leq 1$



The AUC is 0.86 which is not very good, this suggests that time is not ideal for diagnosing variant.

## ROC Summary

- Confusion matirx - absolute and relative methods
- Key measures - FPR, TPR
- $2 \times 2$ relative confusion matirx has 2 degrees of freedom (can be represented by 2 of the four entries)