CISC 271   Class 23

Classification – Assessment With ROC Curve

Text Correspondence: Fawcett, 2006 [4]

Main Concepts:

- *ROC – Receiver Operator Characteristic*
- *ROC curve for scores of data*
- *AUC – Area Under Curve*

Sample Problem, Machine Inference: How good is a classification algorithm?

23.1 Receiver Operator Characteristic – ROC

One common way of visually presenting a classifier's performance is to use the Receiver Operator Characteristic, or ROC. Historically, this came from attempts during wartime to evaluate the performance of humans who interpreted radar signals. An example of a radar operator, from World War II in Britain, is show in Figure 23.1. The terminology arose from:

Receiver: the radar device that sensed and displayed electromagnetic readings
Operator: the human who interpreted the receiver display
Characteristic: an evaluation of the human operator's performance

Figure 23.1: A human operated, or interpreted, radar during its early wartime use. Here, a member of the Women's Auxiliary Air Force is observing the radar returns on a cathode ray tube.

The ROC, for a trained classifier on a single data set, is a 2D vector. The first entry is the false positive rate, or FPR, which is sometimes represented as “1 – specificity”. The second entry is the true positive rate, or TPR, which is also the sensitivity. The ROC is usually presented graphically, with the ROC vector being plotted in a 2D square that is bounded by $(0, 0)$ and $(1, 1)$. An example of five classifiers that are evaluated on a single data set is shown in Figure 23.2.

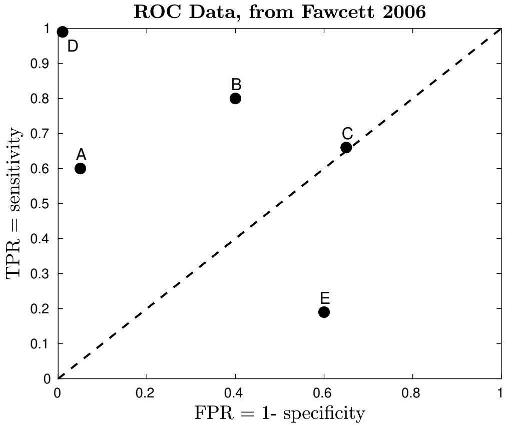


Figure 23.2: The receiver operator characteristic (ROC) points for five classifiers on a single data set, shown as solid circles; each circle is the ratio of the true positive rate to the false positive rate, or TPR/FPR. The horizontal axis is the false positive rate (FPR), which is 1 minus the specificity. The vertical axis is the true positive rate (TPR), which is the sensitivity. The dashed line is the ROC curve for the null hypothesis, which is that the populations are equal. In this plot, one classifier has exceptional performance (D); two have good performance (A,B); one performs at approximately the level of chance (C); and one performs much worse than chance (E).

The crucial value for investigating the ROC is a *score*, which is a non-binary output value of a classifier. In this course we have seen scores provided by various methods, such as principal components analysis (PCA) and linear discriminant analysis (LDA). From the class that explored confusion matrices, we can understand these mathematical relationships for binary classification:

- A score threshold produces a binary classification
- Labels and predicted classes produce a relative confusion matrix
- A relative confusion matrix can be represented by two degrees of freedom, for example, as (TPR,FPR)

It is *very* important to recognize that these relationships depend on the relative values, or rates; a standard confusion matrix needs to be transformed to a relative confusion matrix to understand that the entire matrix can be represented as a 2D point.

Conventional sources that explore confusion matrices do not always use our simple relationships. A simple example, from an article by Tom Fawcett [Fawcett, 2006], is for 20 scalar data points with associated scores; these values are in the extra notes for this class. Suppose that we say that a point is in Class +1 if and only if the score is above a threshold value θ, i.e., point x_i is in Class +1 if and only if $r_i \geq \theta$. As we vary the classification threshold – which, for these data, is varying $0 \leq \theta \leq 1$ – then for each threshold value we can compute a relative confusion matrix. The TPR and TNR for each threshold are an ROC vector. These vectors are plotted in Figure 23.3.

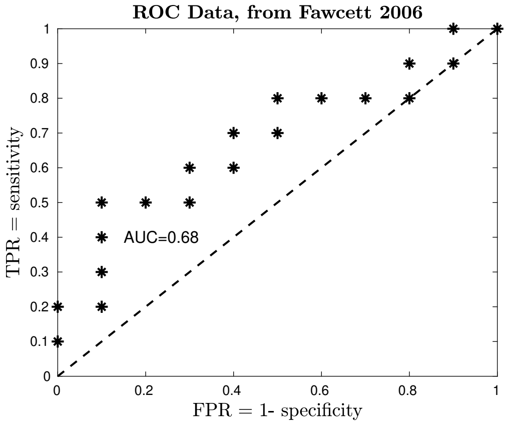


Figure 23.3: The receiver operator characteristic (ROC) points for a small set of discrete data sample populations, shown as solid asterisks.

23.2 Example: Two Variants of a Virus

The ROC curve can also be explored by using a continuous data set, or a data set that contains a large number of vectors. A simple example is to suppose that there are two clinical populations that have a definitive diagnosis: they have one of two variants of a virus that is widely infecting the population. Persons infected with Variant 1 of the virus are in Class +1, and those with Variant 2 of the virus are in Class -1. Suppose that there is a quantitative test that determines whether a person is infected with the virus but the test cannot distinguish between the variants; we can use the number of days after the subject presents symptoms as a hyper-parameter, which here is a real number that is bounded as $0.1 \leq t \leq 20$.

An example of the normalized probability of two such populations is shown in Figure 23.4. The red curve corresponds to the Class -1 population who have Variant 2 of the virus, which has a peak incidence at around 8 days. The blue curve corresponds to the Class +1 population who have Variant 1 of the virus, which has a peak incidence at around 4 days. The scores of the Class -1 population tend to be lower than the scores of the Class +1 population but, for any value t, there are some experimental subjects in Class -1 with that score and some experimental subjects in Class +1 with that score.

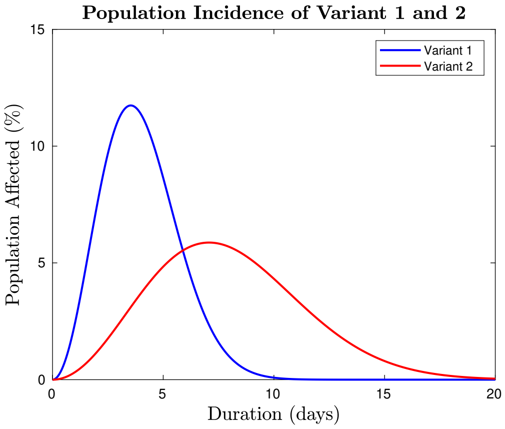
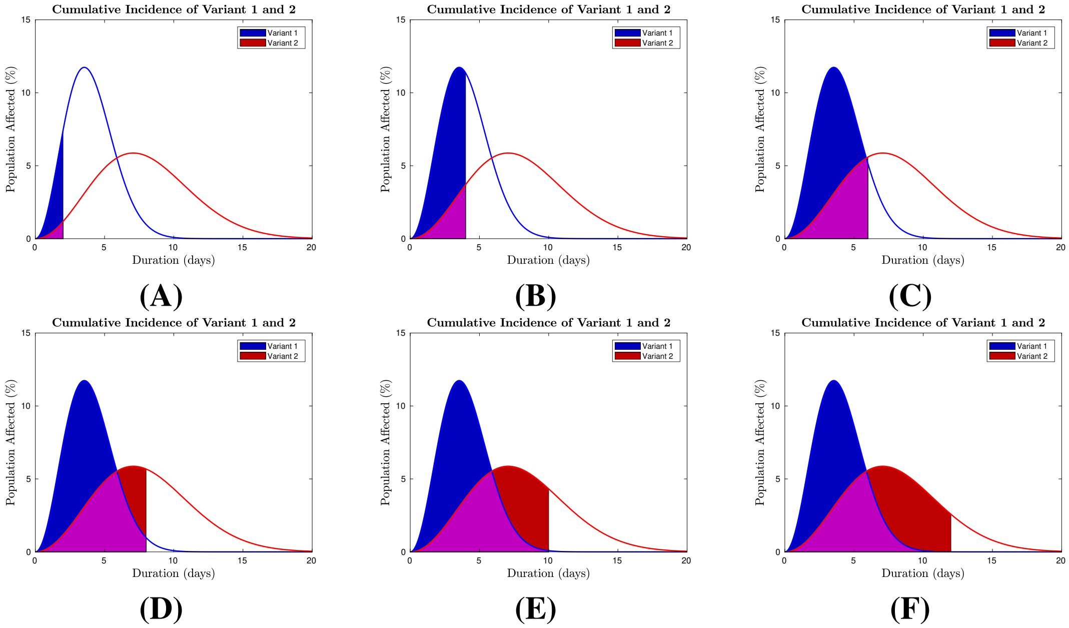


Figure 23.4: Probability density functions for two sample populations who are infected with two variants of a virus. The horizontal axis is a scalar argument, such as an experimental reading. The red curve is the probability density function of the experimental subjects with label -1. The blue curve is the probability density function of the experimental subjects with label +1.

Using numerical methods, we can estimate the various entries of the confusion matrix for any value of t. For example, in Figure 23.5, the true positive rate is the area under the blue curve between the value t and the maximum value, which has been normalized to be 1.

Let us suppose that we are most interested in the fast-infecting variant of the virus. For any given day, we will take the area under the Variant 1 curve to be the true positive population and the area under the red curve to be the false positive population. The true positive rate (TPR) and false positive rate (FPR) can be graphed as a function of the number of days, after presenting symptoms, that the diagnostic test was performed. Figure 23.6 plots these two rates, so we can see that true positive rate is generally greater than the false positive rate. But it may be difficult to draw an inference from these rate curves. We can, however, see that as t approaches 20, both the FPR and the TPR approach 1; this is because only the "tails" of the distributions are remaining.

If we plot the FPR as the horizontal axis, and the TPR as the vertical axis, we have the ROC curve that is shown in Figure 23.7. Integrating this curve, we find that the area under the curve – which is the AUC – is approximately 0.86. We can deduce that, for most values of t, the ratio TPR/FPR is substantially less than 1. The virus test that is provided to us in this example does not perform especially well at distinguishing the two variants of the virus, and we might recommend that a better diagnostic test be developed.



(A)

(B)

(C)

(D)

(E)

(F)

Figure 23.5: Cumulative incidences of two variants of a virus in a population of subjects. The blue area represents the percentage of persons who are infected with Variant 1 of the virus on a given number of days after presenting with symptoms. The red area represents the percentage of persons who are infected with Variant 2 of the virus on a given number of days after presenting with symptoms.

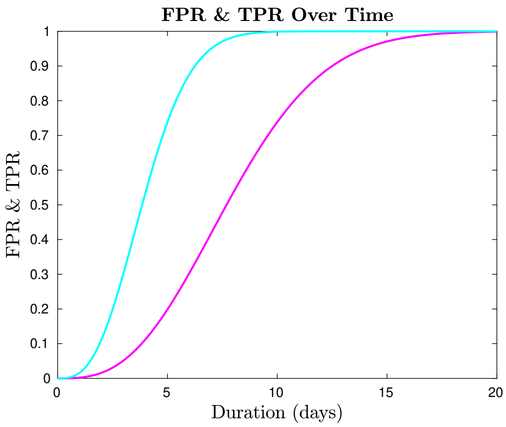


Figure 23.6: True positive rate (TPR) and negative positive rate (FPR) for two sample populations. The horizontal axis is the number of days, after presenting symptoms, that patients were tests for a virus. The cyan curve is the TPR for each value of the scalar argument; this is the area under the probability density function of the label +1 population. The magenta curve is the FPR for each value of the scalar argument; this is the area under the probability density function of the label -1 population.

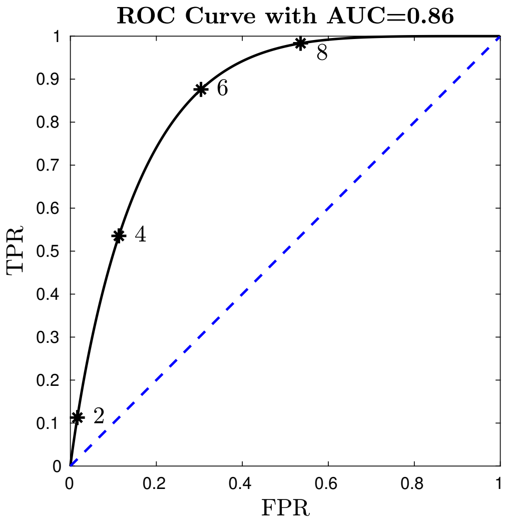


Figure 23.7: The receiver operator characteristic (ROC) curve for two sample populations, shown as a solid black line.

These are simple, sparse examples of the use of ROC curves and the AUC measure. Interested students may wish to consult the abundant literature available on the assessment of a classifier. The two journal articles cited below were used to develop this material, so these would be potential starting points for deeper understanding.

Extra Notes

MATLAB compatible version of Fawcett's data for ROC analysis

```
values=   [ .90 .80 .70 .60 .55 .54 .53 .52 .51 .505 ...
            .40 .39 .38 .37 .36 .35 .34 .33 .30 .1]';
labels=   [+1 +1 -1 +1 +1 +1 -1 -1 +1 -1 ...
           +1 -1 +1 -1 -1 -1 +1 -1 +1 -1]';
```

End of Extra Notes