Giang Duong
Student ID: 014533857
Homework 3: 1, 4, 5, 6, 7, 8, 10. In CHAPTER 8 DATA ANALYSIS.
Professor Mark Stamp CS 271 Fall 2020

1.The topic of n-fold cross validation is discussed in Section 8.2.

a) Show that when n-fold cross validation is used, we obtain exactly n match scores, where n is the size of the match set.

Given S is the match set and U0, U1,… U (n-1) are a partition of S which each Ui has equal size of elements. Note, there is no duplicate in Ui and there is no duplicate element to others Ui.

Repeatedly, each time we take one Ui test set while trainning including others set. Call O.

We will train a model which is used to compute scores on the set O. After the loop, we will have Nth match scores.

b) Assuming that the nomatch set is of size n and we do n-fold cross validation, how many total nomatch scores are computed?

If we have the nomathc set is the size of N then do the n-fold cross validation. We will have N^2 total nomatch scores.

c) In practice, it's common to use either n = 5 or n = 10. What is the advantage of 5-fold cross validation over 10-fold? What is the advantage of 10-fold cross validation over 5-fold?

With the n =5, the model will run faster than model with n = 10.

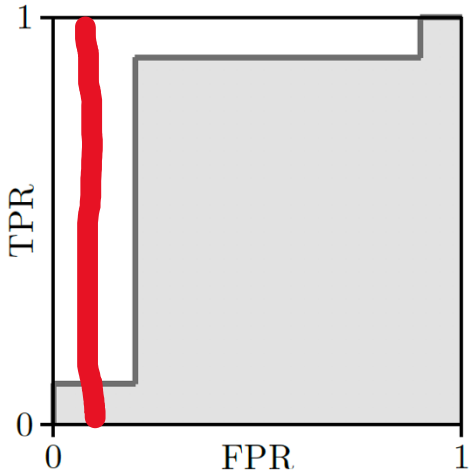
The variance in the data of the model 5 would be less and data will have more bias.

With the n = 10, the model will be more realiable than the model with n = 5

The variance in the data of the model 10 would be high while the bias in the dataset will be lower.

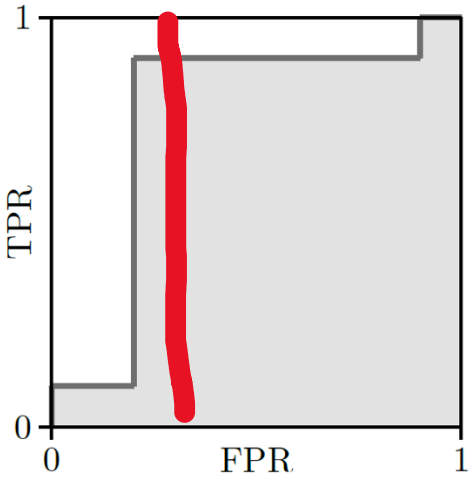
4. This problem deals with the partial AUC, which we denote as AUC, where $0 < \square < 1$. AUC == area under the curve.

a) Determine AUC 0.1 for the example in Figure 8.6.



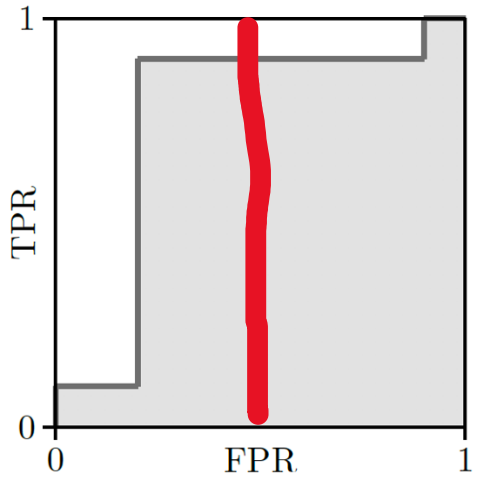
AUC (0.1) = 0.01 / 0.1 = 0.1

b) Determine AUC 0.3 for the example in Figure 8.6.



AUC (0.3) = 0.11/ 0.3 = 11/30 = 0.366667

c) Determine AUC 0.5 for the example in Figure 8.6.



AUG (0.5) = 0.29 /0.5 = 0.58

5. Repeat the analysis in Section 8.5, assuming that we have selected a threshold for which TPR = 0.90 and FPR = 0.0001.

Given TPR = 0.90 and FPR = 0.0001 and scan 100,000 samples. Suppose that only about 1 out of each 1000 sample tested are malware.

Since TPR = 0.90, we detect 90 malware samples and misclasssified 10 benigns.

We tested 99900 negative samples with FPR = 0.0001, we have 99890 negative sample and about 10 malwares.

== Probability detect benigns = (99890 / 99900) > 0.99989989999

== Probability detect malware = (90 / (90 +10)) = 0.9

6. Suppose that for a given experiment, TP = 98, FN = 2, TN = 8500, and FP = 1500.

a) Compute the accuracy.

Accuracy = (TP + TN) / (P +N)

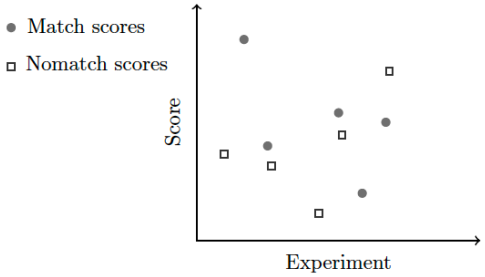
== (98 + 8500) /(98 + 2 +8500 + 1500) = 98.84%

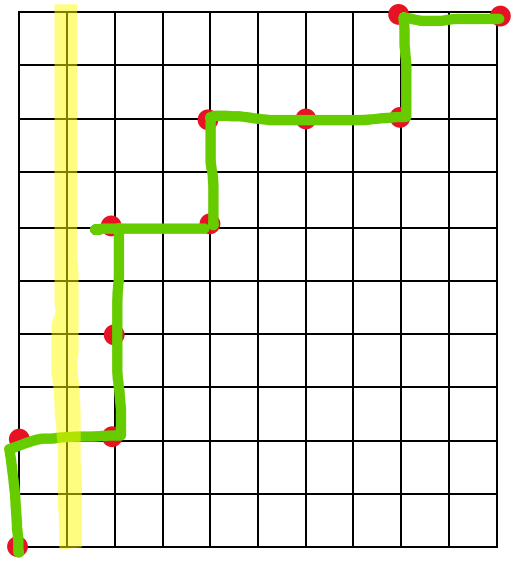
b) Compute the balanced accuracy.

Balanced acurracy = (TPR + TNR)/ 2 = 0.5(TP/P) + 0.5(TN/N)

== 0.5 (98 / (98 +2)) + 0.5 (8500 / (8500 + 1500) = 91.5%

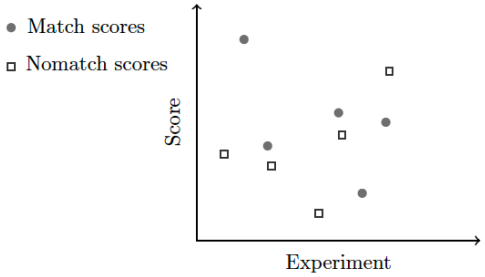
7. For the following scatterplot, draw the corresponding ROC curve, com-pute the area under the ROC curve (AUC), and compute the partial area under the curve, AUC0.1.

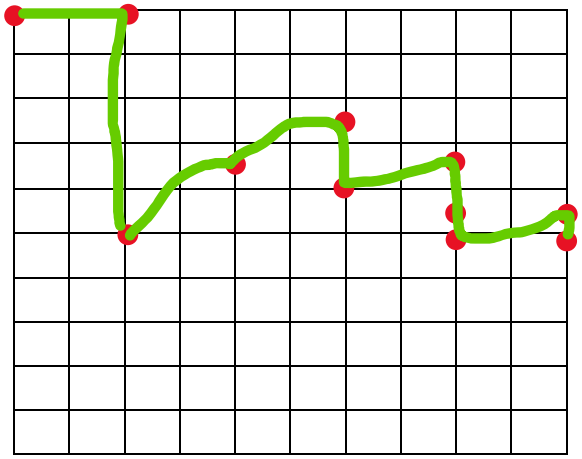




AUC (0.1) = 0.02 / 0.1 = 0.2

8. Using the same scatterplot as in Problem 7, draw the PR curve and compute the area under the PR curve, AUC-PR, using linear interpo-lation between points on the curve.





10. Suppose that we have developed a test for a specific type of malware. Further, suppose that for the threshold that we have chosen, we find that TPR = 0.95 and FPR = 0.01. Further still, suppose that we expect 1 out of every 1000 files tested will be malware of this specific type. Furthermost, assume that we test 200,000 files.

TPR = 0.95 and assumption we have 1 out of 1000 and we test 200,000 files ➔ we detected 190 malwares.
benigns file = 199800 files and FPR = 0.01 ➔ We have 199800* (1-0.01) = 197802 benigns and 1998 malwares.
== ∑ malware = 190 + 1998 = 2188
==∑ benign = 10 + 197802 = 197812

a) What fraction of the files classified as benign are actually benign?
197802 / (197802 +10) = 0.9999494469

b) What fraction of the files classified as malware are actually malware?
190 / (190 + 1998) = 0.08683729433

c) We could improve on these results by either changing the threshold to reduce the false positive rate, or by performing a secondary test on each sample that is classified as malware. Discuss the pros and cons of both of these approaches.
If you reduce the threhold of FPR, we may face with the number of false negatives. But we do not need to perform another computation. Testing is costly.

If we perfome a secondary test on each samle that that is classified as malware, this will give us the sample sets with a few malware classified as benigns and a few benigns sample classified as malwares. That will give us a balance. The cons for the secondary test is optimal threhold will also rely on the cost of these tests.

REFERENCES

1. Data Analysis, Introduction to Machine Learning with Applications in Information Security by Mark Stamp.