Data Mining

Classification VI - Evaluating Classifier Performance (Part B)

Dr. Jason T.L. Wang, Professor

Department of Computer Science

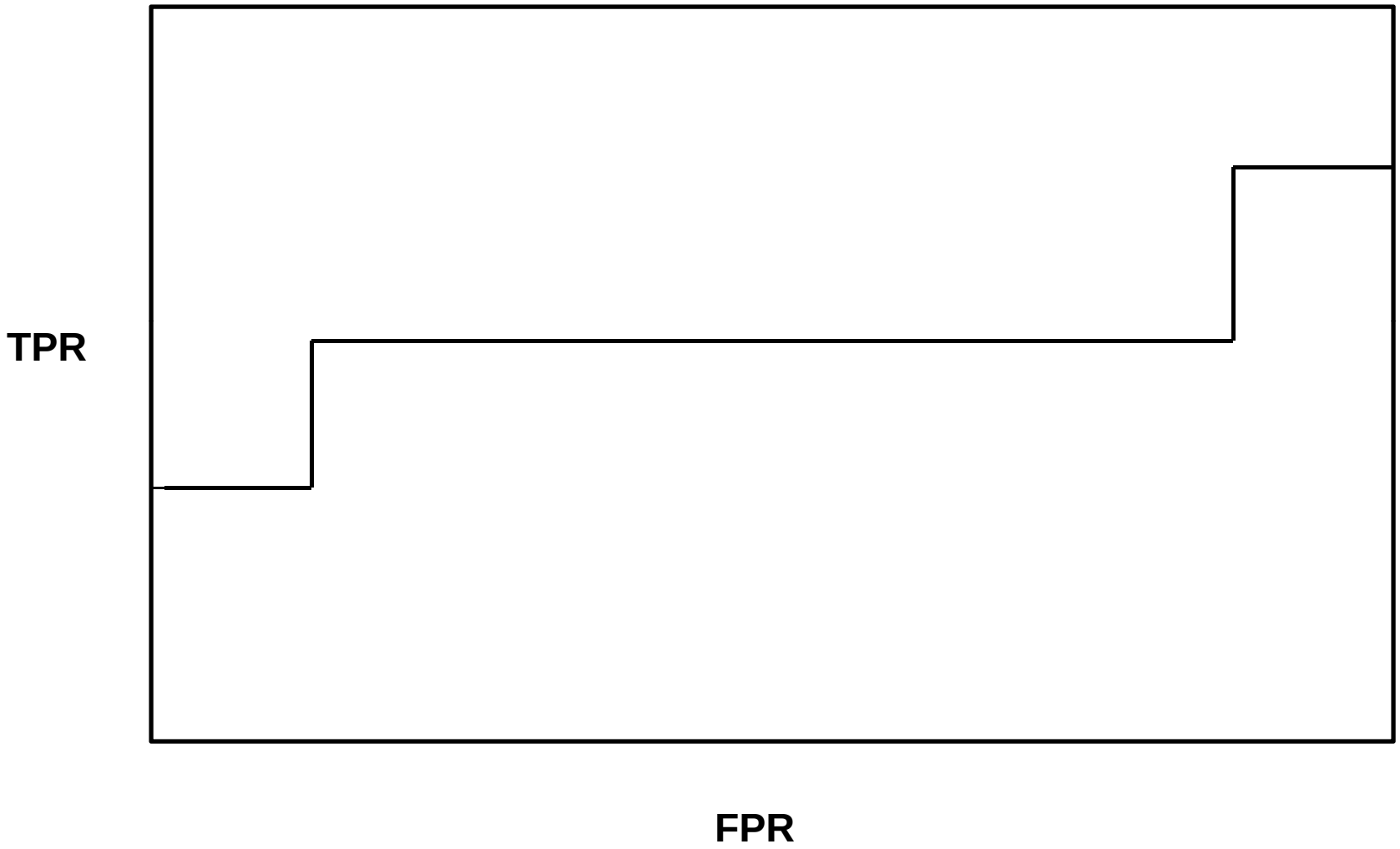
New Jersey Institute of Technology

Where am I?

- Part A presents the definitions and terms (TP, FP, TN, FN, TPR, FPR, etc.) and 10-fold cross validation.

- Part B shows how to generate an ROC curve and calculate AUC.

The Receiver Operating Characteristic (ROC) Curve



Area under ROC curve (AUC)

- Classifier A is better than classifier B if A's AUC is larger than B's AUC.
- A perfect classifier has an AUC of 1.
- A classifier making random guesses has an AUC of 0.5.
- AUC can be calculated by summing up the areas of the small rectangles underneath ROC.

Generating an ROC curve

1. Sort the test data records in increasing order of their output values defined for the positive class that are produced by the classifier.
2. Select the lowest ranked test record (with lowest output value). Assign the selected record and those ranked above it to the positive class (i.e. predict all the test records as positive). Calculate TPR and FPR.

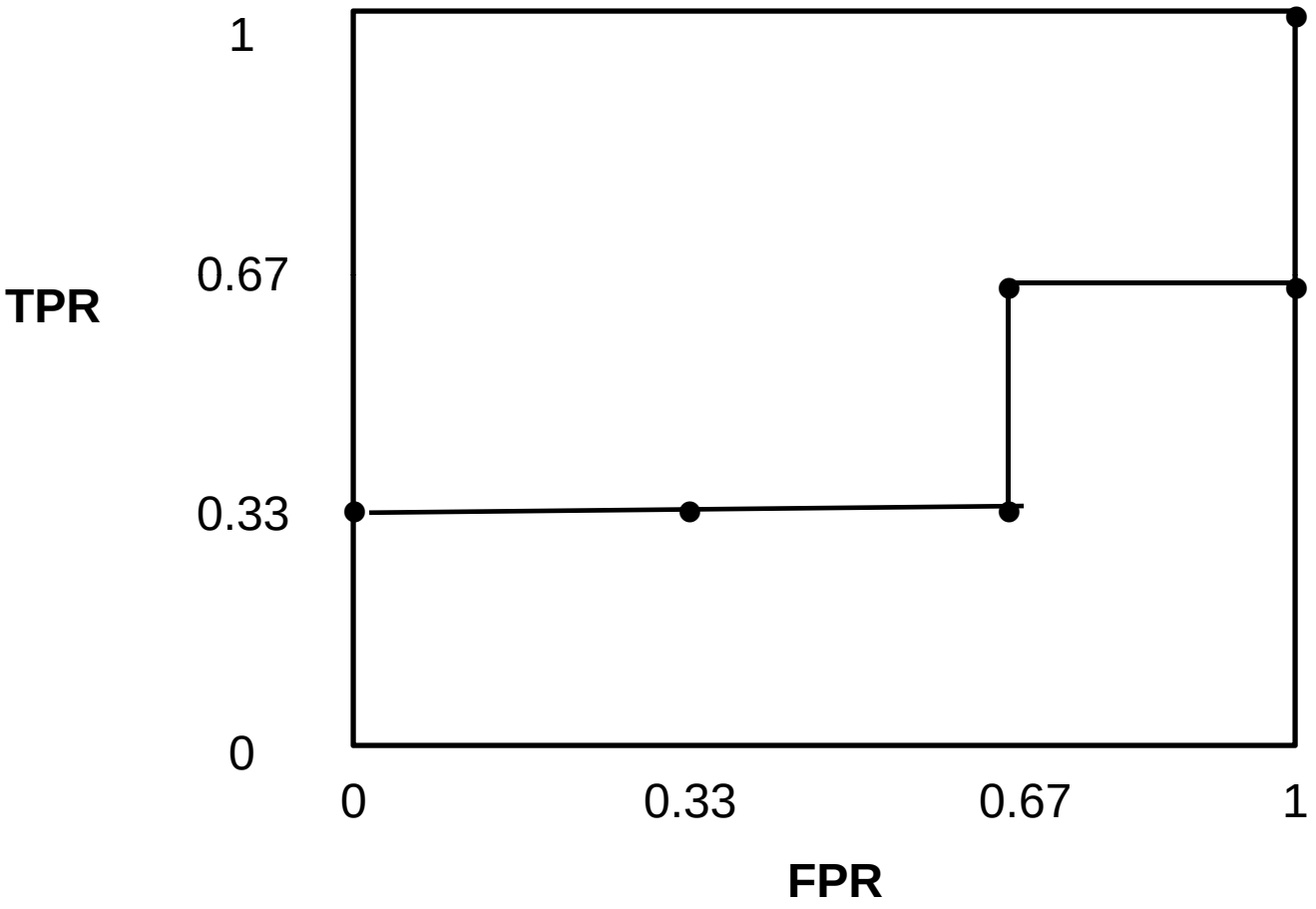
Generating an ROC curve

3. Select the next test record from the sorted list. Classify the selected record and those ranked above it as positive, and those below it as negative. Calculate TPR and FPR.
4. Repeat Step 3.
5. Plot TPR against FPR.

Example

Class Label	+	-	+	-	-	+	
	0.21	0.42	0.57	0.78	0.84	0.92	1.00
TP	3	2	2	1	1	1	0
FP	3	3	2	2	1	0	0
TN	0	0	1	1	2	3	3
FN	0	1	1	2	2	2	3
TPR	1	0.67	0.67	0.33	0.33	0.33	0
FPR	1	1	0.67	0.67	0.33	0	0

Example (ROC Curve)



End of

Evaluating Classifier Performance Module (Part B)