

Model Evaluation Metric

서울대학교 컴퓨터공학부 이영기



Human-Centered
Computer Systems Lab



SEOUL NATIONAL UNIVERSITY

Evaluating Classifier Performance

- Let's assume we want to evaluate the performance of activity recognizer
- We have labelled data
 - Accelerometer data for different activities (sitting, standing, etc.)
 - We know ground truth which activity the data is mapped to
- How do we measure the performance of the classifier?

Importance of Evaluation

- Determine if it is sufficient for the target application
 - It is important to estimate the accuracy of the classifiers for mission-critical applications, like medical applications
- Optimize the classifier
 - Designing a classifier usually does not end in a single cycle and need to go through a multiple rounds of optimizations

Performance Measures

- We need to know which metrics to use to evaluate the classifier performance
- There are many widely used metrics
 - Confusion Matrix, Accuracy, Precision, Recall, F-measure, ROC curve, etc.

Confusion Matrix

- Consider a binary classification problem that distinguish "Sitting" versus "Not Sitting" (let us refer to them as A and B for ease of notation)
- A classifier may result in the following confusion matrix when tested on independent data

		Predicted class	
		Sitting (A)	Not Sitting (B)
Known class (class label in data)	Sitting (A)	25	7
	Not Sitting (B)	4	51

Let's Simplify with Binary Classifier

- Now we want to classify Sitting (A) vs. Not Standing (B)
- The confusion metrics will look as below
 - TP: True Positive, TN: True Negative
 - FP: False Positive, FN: False Negative

		<i>Estimated class</i>		P	N	
		A	B			
<i>Actual class</i>	A	TP	FN			
	B	FP	TN			

Common Metrics

- Accuracy = $(TP+TN) / (P+N)$
- Error = $(FP+FN) / (P+N)$
- Precision = $TP / (TP+FP)$
- Recall = TP / P
- TP Rate (sensitivity) = TP / P
- TN Rate (specificity) = TN / N

		<i>Estimated class</i>		
		A	B	
<i>Actual class</i>	A	<i>TP</i>	<i>FN</i>	<i>P</i>
	B	<i>FP</i>	<i>TN</i>	<i>N</i>

Be careful of “Accuracy”

- The simplest measure of performance would be the fraction of items that are correctly classified, or the “accuracy” which is:

$$\frac{tp + tn}{tp + tn + fp + fn}$$

- But this measure is dominated by the larger set (of positives or negatives) and favors trivial classifiers
- e.g. if 5% of items are truly positive, then a classifier that always says “negative” is 95% accurate

F-measure

- It is often preferred to combine precision and recall into a single number referred to as the F-measure
- It can be interpreted as a weighted average of the precision and recall, where the measure reaches its best value at 1 and worst score at 0

$$F = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}} \right)}$$

ROC (Receiver-Operating Characteristic)

True	Predicted	
	pos	neg
pos	40	60
neg	30	70

True	Predicted	
	pos	neg
pos	70	30
neg	50	50

True	Predicted	
	pos	neg
pos	60	40
neg	20	80

Classifier 1

TPr = 0.4

FPr = 0.3

Classifier 2

TPr = 0.7

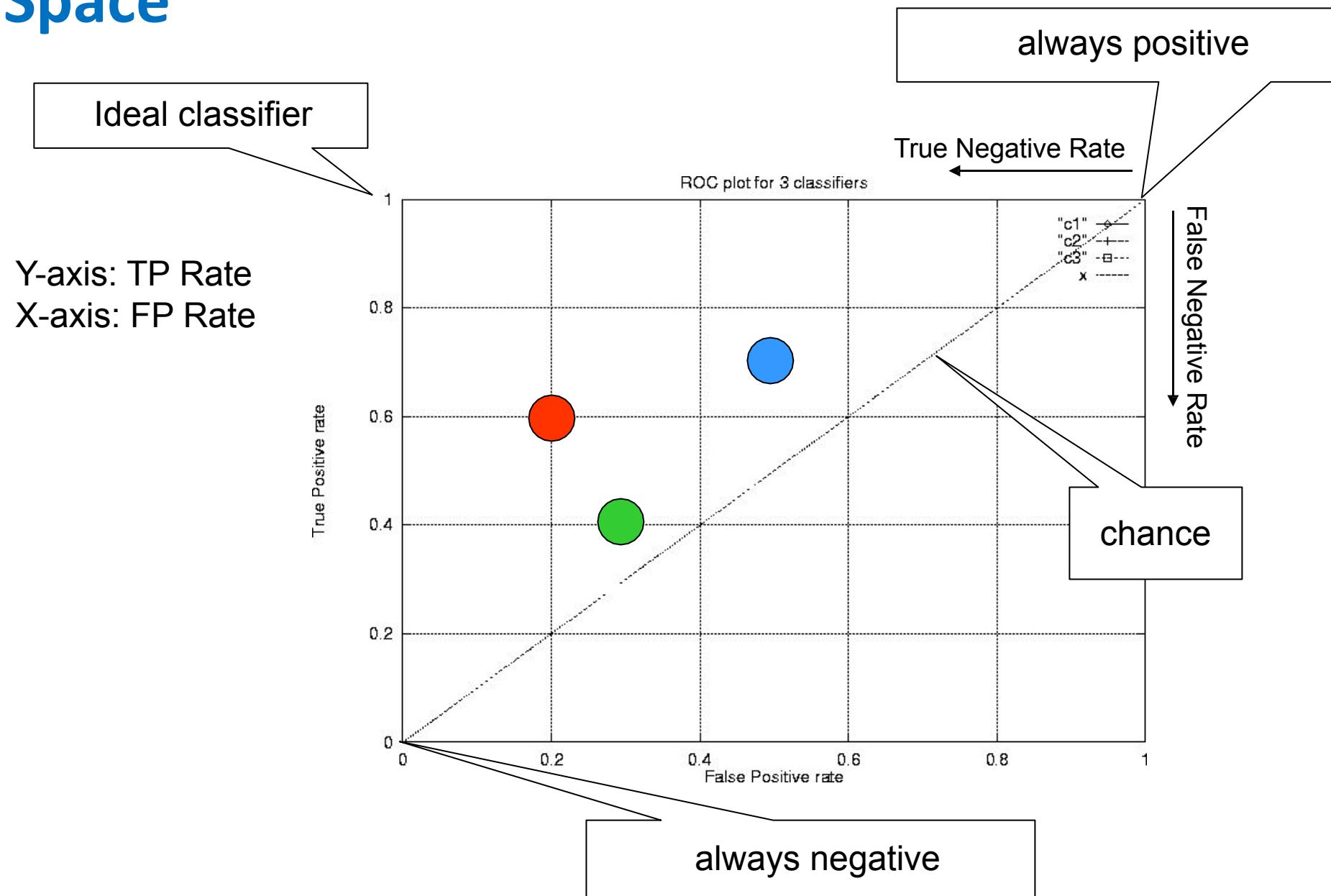
FPr = 0.5

Classifier 3

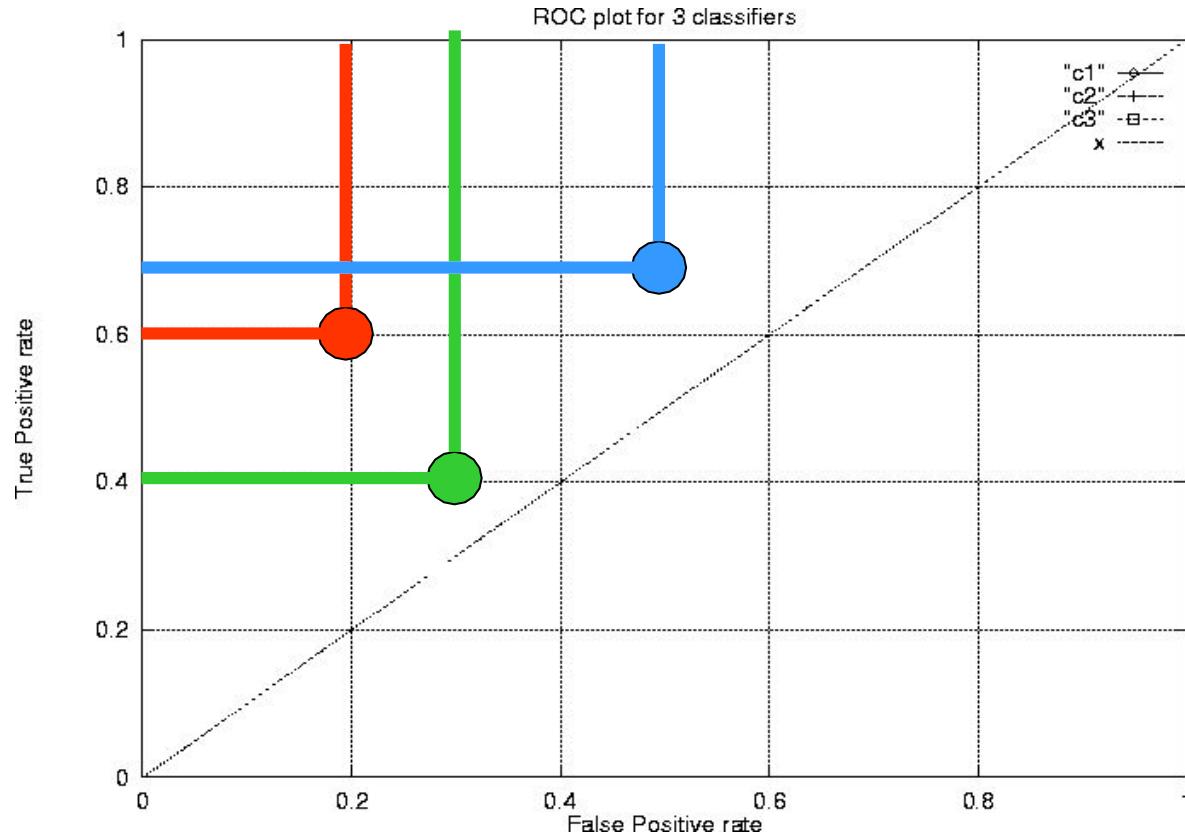
TPr = 0.6

FPr = 0.2

ROC Space



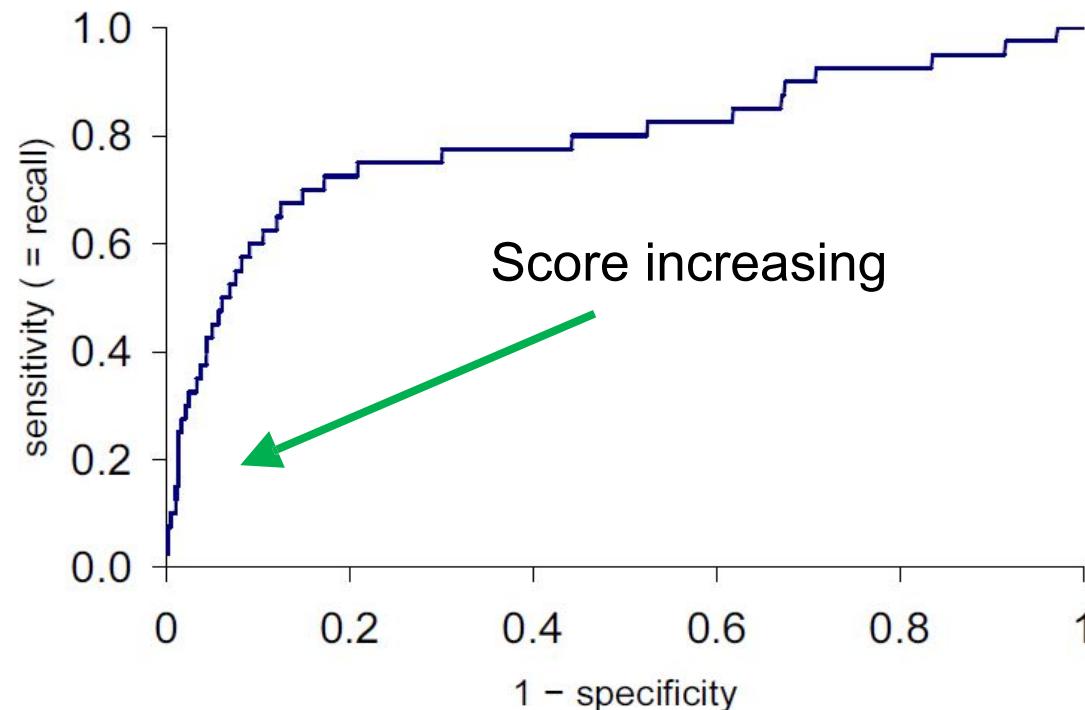
Dominance in the ROC Space



Classifier A dominates classifier B if and only if $TPr_A > TPr_B$ and $FPr_A < FPr_B$.

ROC Curve

- It is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied
- The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings



ROC AUC

- ROC AUC is the “Area Under the Curve” – a single number that captures the overall quality of the classifier. It should be between 0.5 (random classifier) and 1.0 (perfect)

