Classifier 성능 평가

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
  - In particular, 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 e.g. a three class problem with the classes Sitting, Standing, and Walking (let us refer to them as A, B, and C for ease of notation).
- A classifier may result in the following confusion matrix when tested on independent data.

		Predicted class		
		Sitting (A)	Standing (B)	Walking (C)
Known class (class label in data)	Sitting (A)	25	5	2
	Standing (B)	3	32	4
	Walking (C)	1	0	15

Let’s Simplify with Binary Classifier

- Now we want to classy Sitting (A) vs. Standing (B)
- The confusion metrics will look as below
  - TP: True Positive, TN: True Negative
  - FP: False Positive, FN: False Negative

		Estimated class	
		A	B
Actual class	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
- FP Rate (specificity) = FP / N

		Estimated class		
		A	B	
Actual class	A	*TP*	*FN*	*P*
	B	*FP*	*TN*	*N*

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.

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

ROC (Receiver-Operating Characteristic)

True	Predicted	
	pos	neg
pos	40	60
neg	30	70

Classifier 1
TPr = 0.4
FPr = 0.3

True	Predicted	
	pos	neg
pos	70	30
neg	50	50

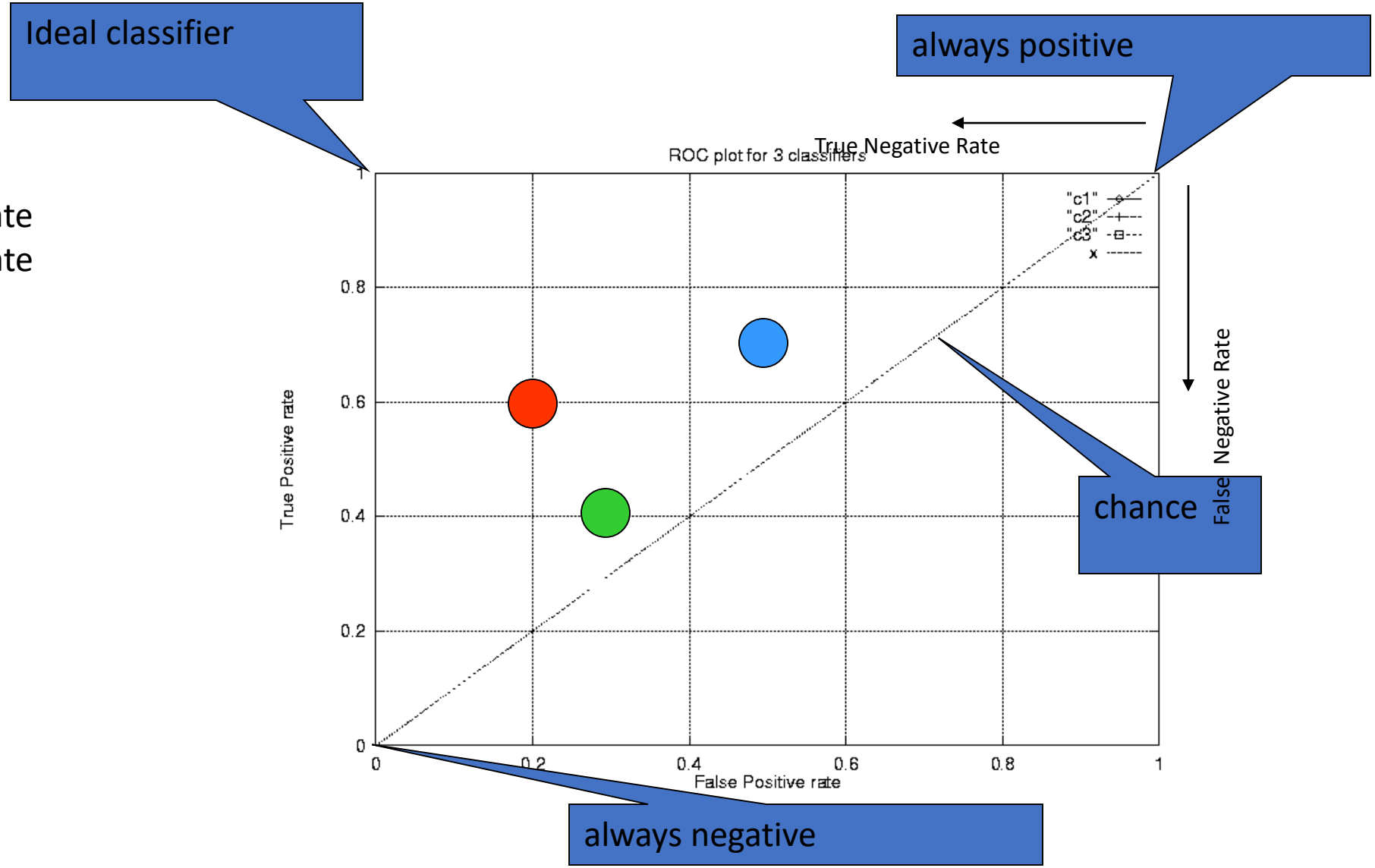
Classifier 2
TPr = 0.7
FPr = 0.5

True	Predicted	
	pos	neg
pos	60	40
neg	20	80

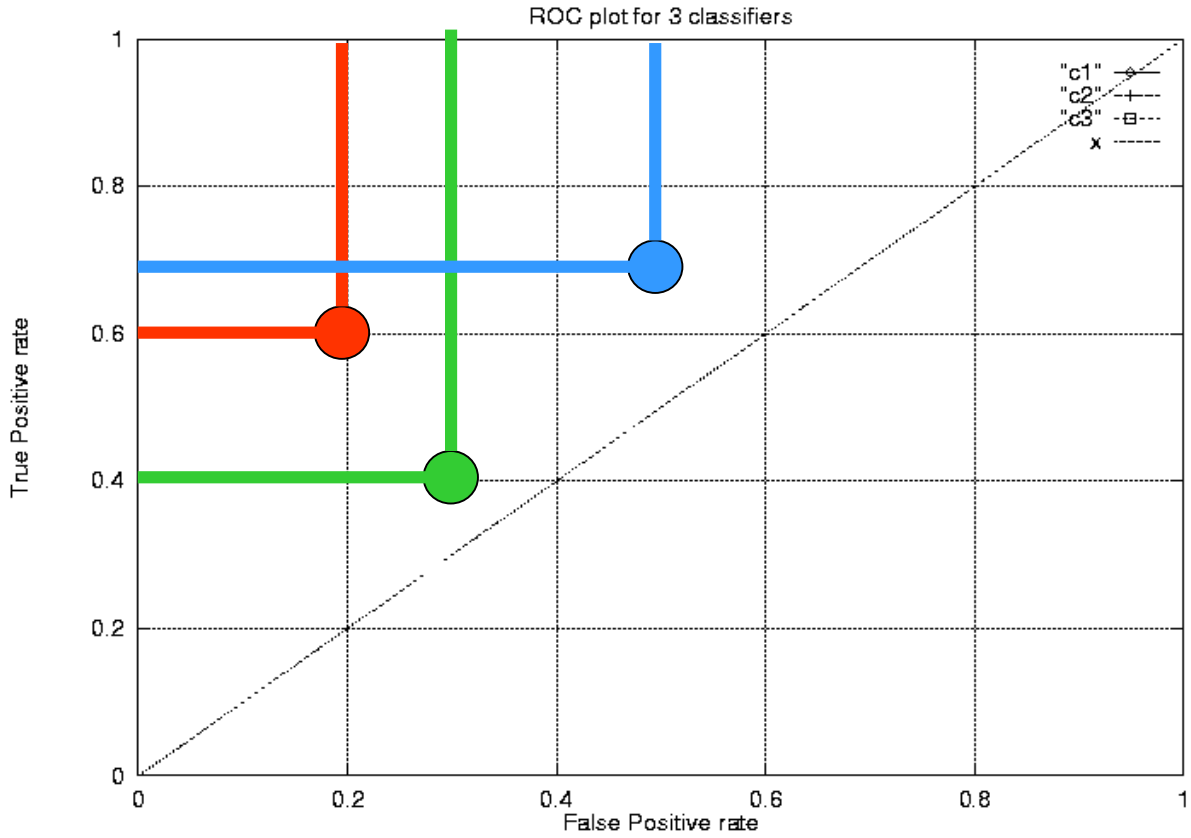
Classifier 3
TPr = 0.6
FPr = 0.2

ROC Space

Y-axis: TP Rate
X-axis: FP Rate



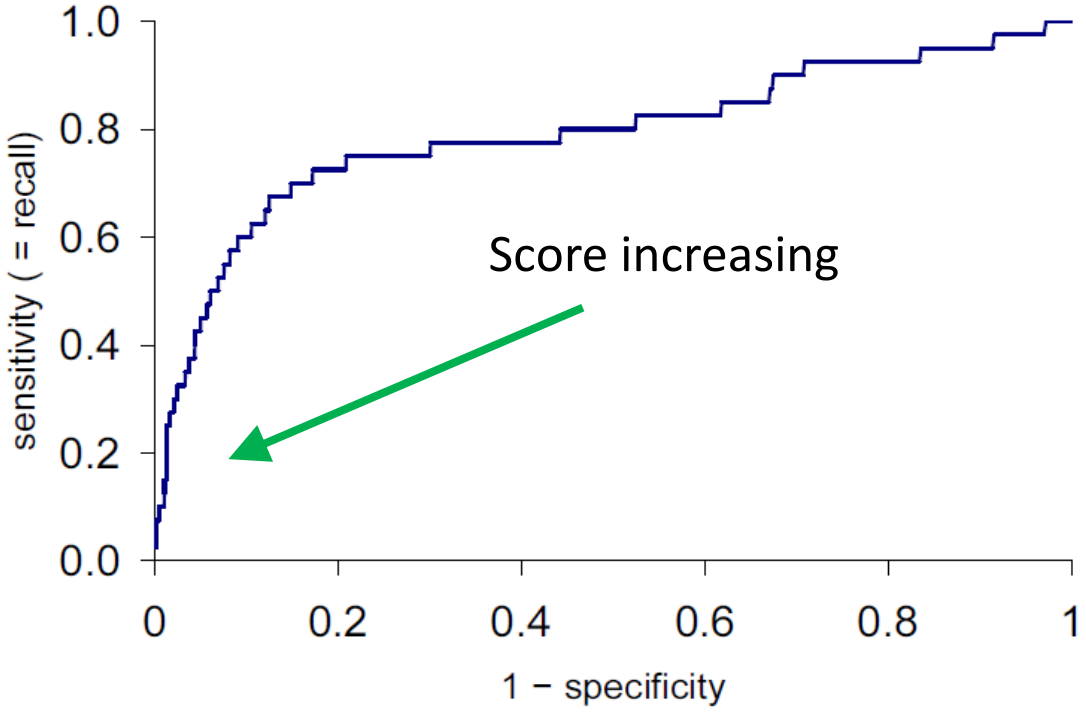
Dominance in the ROC Space



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).