

Class Imbalance Problem

- Lots of classification problems where the classes are skewed (more records from one class than another)
 - Credit card fraud
 - Intrusion detection
 - Defective products in manufacturing assembly line
 - COVID-19 test results on a random sample

- **Key Challenge:**
 - Evaluation measures such as accuracy are not well-suited for imbalanced class

Confusion Matrix

□ Confusion Matrix:

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
Class=Yes	a	b
	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Accuracy

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
Class=Yes	a (TP)	b (FN)
	c (FP)	d (TN)

□ Most widely-used metric:

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + TN + FP + FN}$$

Problem with Accuracy

- Consider a 2-class problem
 - Number of Class NO examples = 990
 - Number of Class YES examples = 10
- If a model predicts everything to be class NO, accuracy is $990/1000 = 99\%$
 - This is misleading because this trivial model does not detect any class YES example
 - Detecting the rare class is usually more interesting (e.g., frauds, intrusions, defects, etc)

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
	Class=Yes	Class=No
	0	10
	0	990

Data Mining, 2nd Edition

Which model is better?

A

	PREDICTED		
		Class=Yes	Class=No
ACTUAL	Class=Yes	0	10
	Class=No	0	990

Accuracy: 99%

B

	PREDICTED		
		Class=Yes	Class=No
ACTUAL	Class=Yes	10	0
	Class=No	500	490

Accuracy: 50%

Which model is better?

A

ACTUAL	PREDICTED		
		Class=Yes	Class=No
	Class=Yes	5	5
	Class=No	0	990

B

ACTUAL	PREDICTED		
		Class=Yes	Class=No
	Class=Yes	10	0
	Class=No	500	490

Alternative Measures

	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

$$\text{Precision (p)} = \frac{a}{a + c}$$

$$\text{Recall (r)} = \frac{a}{a + b}$$

$$\text{F - measure (F)} = \frac{2rp}{r + p} = \frac{2a}{2a + b + c}$$

Alternative Measures

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
	Class=Yes	Class=No
	10	0
	10	980

$$\text{Precision (p)} = \frac{10}{10+10} = 0.5$$

$$\text{Recall (r)} = \frac{10}{10+0} = 1$$

$$\text{F - measure (F)} = \frac{2 * 1 * 0.5}{1 + 0.5} = 0.62$$

$$\text{Accuracy} = \frac{990}{1000} = 0.99$$

Alternative Measures

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
Class=Yes	10	0
Class=No	10	980

$$\text{Precision (p)} = \frac{10}{10+10} = 0.5$$

$$\text{Recall (r)} = \frac{10}{10+0} = 1$$

$$\text{F - measure (F)} = \frac{2 * 1 * 0.5}{1 + 0.5} = 0.62$$

$$\text{Accuracy} = \frac{990}{1000} = 0.99$$

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
Class=Yes	1	9
Class=No	0	990

$$\text{Precision (p)} = \frac{1}{1+0} = 1$$

$$\text{Recall (r)} = \frac{1}{1+9} = 0.1$$

$$\text{F - measure (F)} = \frac{2 * 0.1 * 1}{1 + 0.1} = 0.18$$

$$\text{Accuracy} = \frac{991}{1000} = 0.991$$

Which of these classifiers is better?

A

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
Class=Yes	40	10
Class=No	10	40

Precision (p) = 0.8

Recall (r) = 0.8

F - measure (F) = 0.8

Accuracy = 0.8

B

ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
Class=Yes	40	10
Class=No	1000	4000

Precision (p) = ~ 0.04

Recall (r) = 0.8

F - measure (F) = ~ 0.08

Accuracy = ~ 0.8

Measures of Classification Performance

	PREDICTED CLASS		
ACTUAL CLASS		Yes	No
	Yes	TP	FN
	No	FP	TN

α is the probability that we reject the null hypothesis when it is true. This is a Type I error or a false positive (FP).

β is the probability that we accept the null hypothesis when it is false. This is a Type II error or a false negative (FN).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

$$ErrorRate = 1 - accuracy$$

$$Precision = \text{Positive Predictive Value} = \frac{TP}{TP + FP}$$

$$Recall = \text{Sensitivity} = TP \text{ Rate} = \frac{TP}{TP + FN}$$

$$Specificity = TN \text{ Rate} = \frac{TN}{TN + FP}$$

$$FP \text{ Rate} = \alpha = \frac{FP}{TN + FP} = 1 - specificity$$

$$FN \text{ Rate} = \beta = \frac{FN}{FN + TP} = 1 - sensitivity$$

$$Power = sensitivity = 1 - \beta$$

Alternative Measures

A	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	40	10
	Class=No	10	40

Precision (p) = 0.8
TPR = Recall (r) = 0.8
FPR = 0.2
F-measure (F) = 0.8
Accuracy = 0.8

$$\frac{\text{TPR}}{\text{FPR}} = 4$$

B	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	40	10
	Class=No	1000	4000

Precision (p) = 0.038
TPR = Recall (r) = 0.8
FPR = 0.2
F-measure (F) = 0.07
Accuracy = 0.8

$$\frac{\text{TPR}}{\text{FPR}} = 4$$

Which of these classifiers is better?

A	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	10	40
	Class=No	10	40

Precision (p) = 0.5

TPR = Recall (r) = 0.2

FPR = 0.2

F – measure = 0.28

B	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	25	25
	Class=No	25	25

Precision (p) = 0.5

TPR = Recall (r) = 0.5

FPR = 0.5

F – measure = 0.5

C	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	40	10
	Class=No	40	10

Precision (p) = 0.5

TPR = Recall (r) = 0.8

FPR = 0.8

F – measure = 0.61

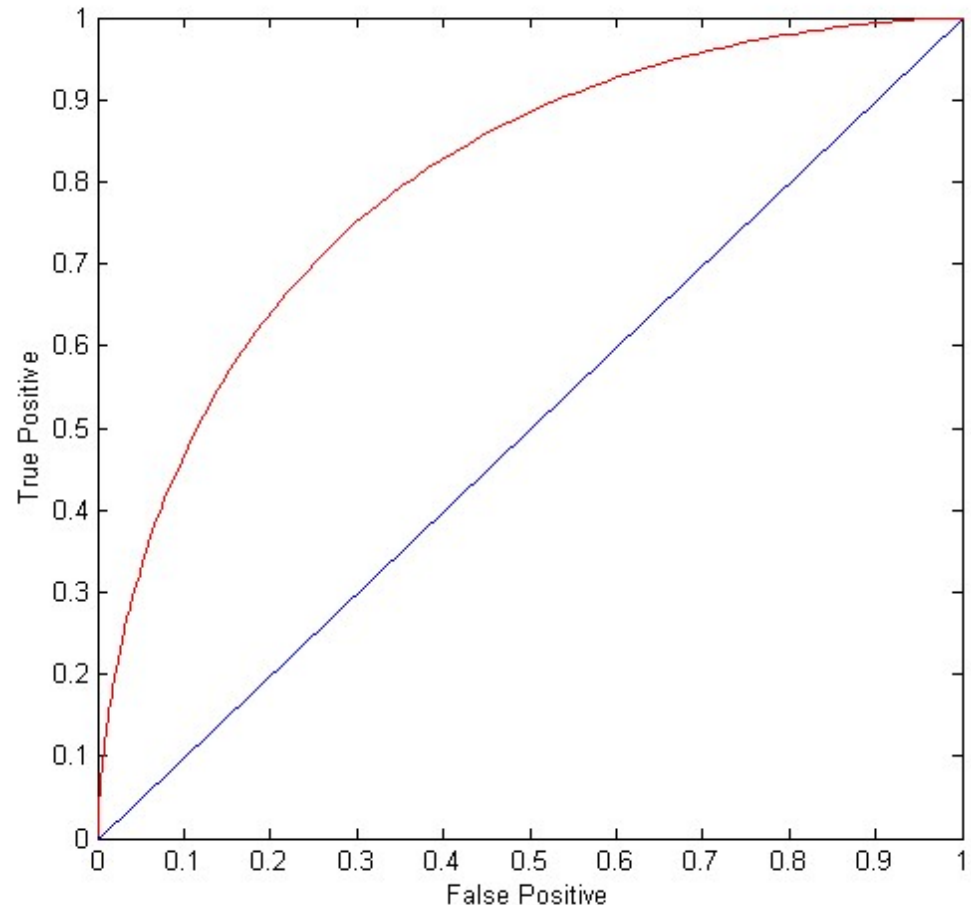
ROC (Receiver Operating Characteristic)

- A graphical approach for displaying trade-off between detection rate and false alarm rate
- Developed in 1950s for signal detection theory to analyze noisy signals
- ROC curve plots TPR against FPR
 - Performance of a model represented as a point in an ROC curve

ROC Curve

(TPR, FPR):

- (0,0): declare everything to be negative class
- (1,1): declare everything to be positive class
- (1,0): ideal
- Diagonal line:
 - Random guessing
 - Below diagonal line:
 - ◆ prediction is opposite of the true class



ROC (Receiver Operating Characteristic)

- To draw ROC curve, classifier must produce continuous-valued output
 - Outputs are used to rank test records, from the most likely positive class record to the least likely positive class record
 - By using different thresholds on this value, we can create different variations of the classifier with TPR/FPR tradeoffs
- Many classifiers produce only discrete outputs (i.e., predicted class)
 - How to get continuous-valued outputs?
 - ◆ Decision trees, rule-based classifiers, neural networks, Bayesian classifiers, k-nearest neighbors, SVM