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

Challenges

- Evaluation measures such as accuracy is not well-suited for imbalanced class

- Detecting the rare class is like finding needle in a haystack

Confusion Matrix

- Confusion Matrix:

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

a: TP (true positive)
b: FN (false negative)
c: FP (false positive)
d: TN (true negative)

Accuracy

		PREDICTED CLASS	
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	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 0 examples = 9990
  – Number of Class 1 examples = 10

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 the model does not detect any class YES example
  – Detecting the rare class is usually more interesting (e.g., frauds, intrusions, defects, etc)

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

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

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

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

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

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

Alternative Measures

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

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

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

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

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

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

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

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

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

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

Alternative Measures

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

$\text{Precision (p)} = 0.8$
$\text{Recall (r)} = 0.8$
$\text{F - measure (F)} = 0.8$
$\text{Accuracy} = 0.8$

Alternative Measures

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

$\text{Precision (p)} = 0.8$
$\text{Recall (r)} = 0.8$
$\text{F - measure (F)} = 0.8$
$\text{Accuracy} = 0.8$

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

$\text{Precision (p)} =\sim 0.04$
$\text{Recall (r)} = 0.8$
$\text{F - measure (F)} =\sim 0.08$
$\text{Accuracy} =\sim 0.8$

Measures of Classification Performance

	PREDICTED CLASS		
		Yes	No
ACTUAL CLASS	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).

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

$$ErrorRate = 1 - accuracy$$

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

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

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

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

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

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

False Positive Rate (FPR)

- False positives also referred to as "Type I errors," a phrase borrowed from the field of medical research.
- False positive rate is also known as fallout or probability of false alarm.
- FPR is given by:

$$\frac{FP}{N} = \frac{FP}{FP+TN}$$

False Negative Rate (FNR)

- The frequency with which the algorithm fails to point out when a prediction 0(False) actually occurs.
- Aka "Type II error."
- False negative rates change in an inverse proportion to false positive rates.
- *False Negative Rate= Σfalse negatives/ Σnegative test outcome*

Alternative Measures

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

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

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

Precision (p) =~ 0.04
TPR = Recall (r) = 0.8
FPR = 0.2
F - measure (F) =~ 0.08
Accuracy =~ 0.8

Alternative Measures

	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	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

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

Precision (p) =~ 0.04
TPR = Recall (r) = 0.8
FPR = 0.2
F - measure (F) =~ 0.08
Accuracy =~ 0.8

Alternative Measures

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

Precision (p) = 0.5
TPR = Recall (r) = 0.2
FPR = 0.2

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

Precision (p) = 0.5
TPR = Recall (r) = 0.5
FPR = 0.5

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

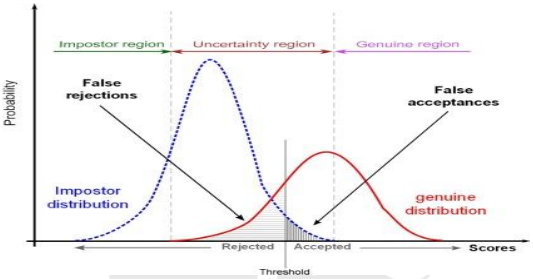
Precision (p) = 0.5
TPR = Recall (r) = 0.8
FPR = 0.8

- **In several classifying domains, on top of selecting the appropriate discriminant function, practitioners also modify the corresponding threshold in order to better suit an independent cost function.**
- **Use scores (also called weights) to express the similarity between a test pattern and a training set. The higher the score is, the higher is the similarity between them.**
- ***Ex. Intrusion Detection System***
- **In theory, normal packet scores should always be higher than the scores of malicious packets. If this would be true, a single threshold, that separates the two groups of scores, could be used to differ between them. But this assumption does not hold in real scenarios.**
- **Choice of threshold critical. If taken too small the score of malicious packet can be larger than selected threshold and hence it will be classified as normal. If too high some normal packets can also be classified as malicious.**
- **If you choose the threshold somewhere between those two points, both false rejections and false acceptances occur.**

Performance Evaluation

- FAR: False Acceptance Rate
  - Those falsely entered the system but actually were imposters, divided by total number of imposters.
- FRR: False Rejection Rate
  - Those could not enter the system but actually were genuines, divided by total number of genuines.
- TPR: True Positive Rate
  - Those who were verified as genuines and actually were genuines, divided by total number of genuines.
- TNR: True Negative Rate
  - Number of Imposters verified as Imposters, divided by total number of imposters.

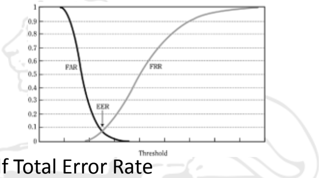
Selection of Threshold



Based on the level of security threshold can be decided. Like, biometric in Nuclear Plant. Not even a single imposter is tolerable. Like, biometric in Class room, let few imposter come into the class but genuine should not be rejected.

Performance Evaluation

- EER: Equal Error Rate
  - Selecting the threshold where FAR is equal to FRR.

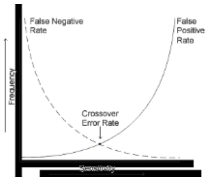


- HTER: Half Total Error Rate
  - Selecting the system where HTER is minimum.
  - HTER= (FAR+FRR)/2

Crossover Error Rate (CER)/Equal Error Rate(EER)

•As the sensitivity of systems may cause the false positive/negative rates to vary, it's critical to have some common measure that may be applied across the board.
•Moreover, False Positive Rate and False negative Rate are interdependent, hence, it is more meaningful to plot them against each other and use a measure that could incorporate both the ideas.
•The CER for a system is determined by adjusting the system's sensitivity until the false positive rate and the false negative rate are equal, as shown in the figure below.
•It gives the measure of system's accuracy, a system with lesser CER is more accurate.

•If you're interested in achieving a balance between false positives and false negatives, you may then simply select the system with the lowest CER.



Lets consider an IDS that classifies normal and malicious packets using classification techniques of machine learning. The test data consists of both malicious and normal. The scores for both malicious (Fig1) and normal packets(Fig2) would be somehow distributed around a certain mean score. A gaussian normal distribution is chosen in this example.

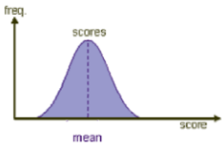


Fig 1

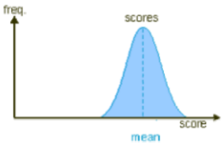
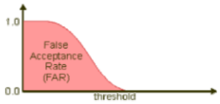
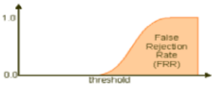


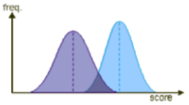
Fig 2

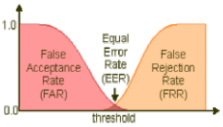


FAR value is one, if all malicious packets are falsely accepted and zero, if none of them are accepted.



FRR value is one, if all malicious packets are rejected and zero, if none of the normal packets are accepted.





The choice of the threshold value becomes a problem if the distributions of the normal and malicious packet scores overlap which they do in real world scenarios as the intrude tries to make a malicious packet look as similar to the normal ones.

ROC (Receiver Operating Characteristic)

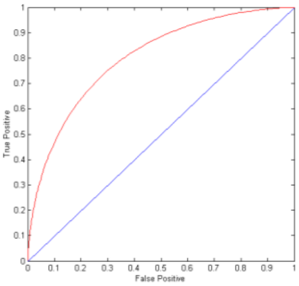
- A graphical approach for displaying trade-off between detection rate and false alarm rate
- Developed in 1950s for signal detection theory to analyze noisy signals
- ROC curve plots TPR against FPR
  - Performance of a model represented as a point in an ROC curve
  - Changing the threshold parameter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

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)

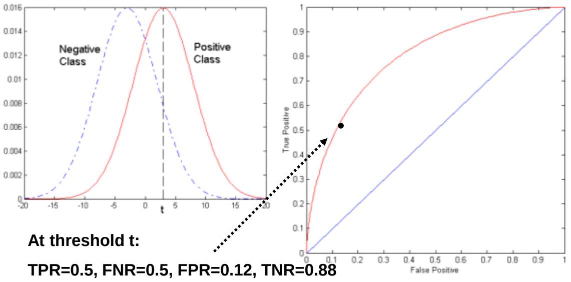
- To draw ROC curve, classifier must produce continuous-valued output
  - Outputs are used to rank test records, from the most likely positive class record to the least likely positive class record

- 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

ROC Curve Example

- 1-dimensional data set containing 2 classes (positive and negative)

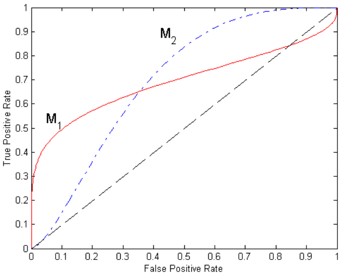
- Any points located at x > t is classified as positive



At threshold t:

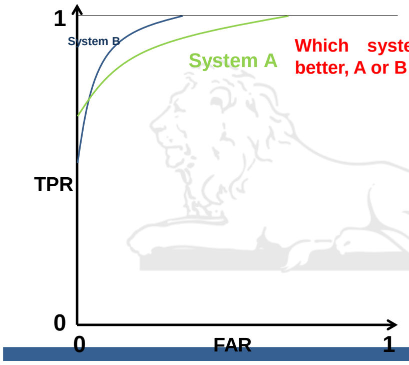
TPR=0.5, FNR=0.5, FPR=0.12, TNR=0.88

Using ROC for Model Comparison

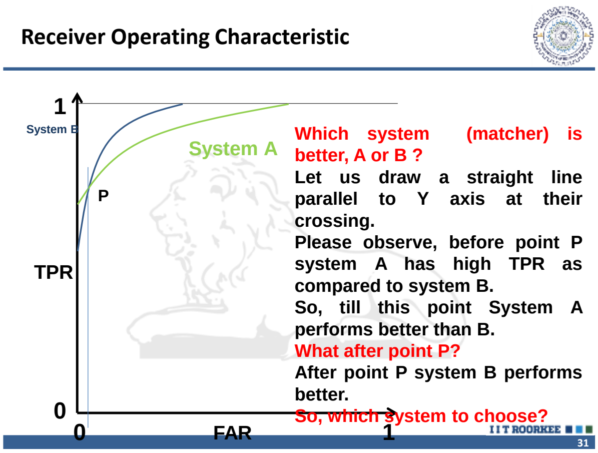


- No model consistently outperform the other
  - M_1 is better for small FPR
  - M_2 is better for large FPR
- Area Under the ROC curve
  - Ideal:
    - Area = 1
  - Random guess:
    - Area = 0.5

Receiver Operating Characteristic



Which system (matcher) is better, A or B ?



Receiver Operating Characteristic

Which system (matcher) is better, A or B ?

Let us draw a straight line parallel to Y axis at their crossing.

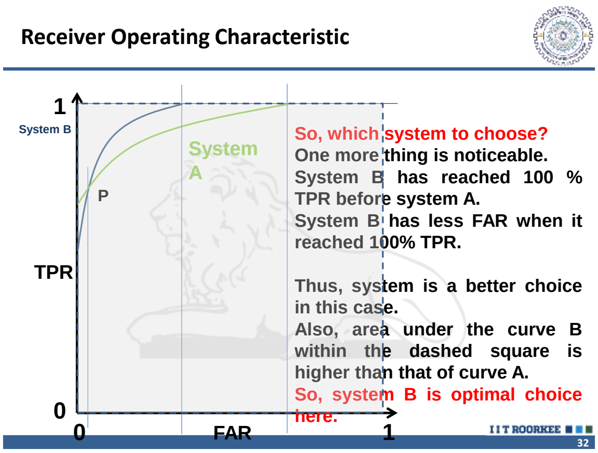
Please observe, before point P system A has high TPR as compared to system B.

So, till this point System A performs better than B.

What after point P?

After point P system B performs better.

So, which system to choose?



Receiver Operating Characteristic

So, which system to choose?

One more thing is noticeable.

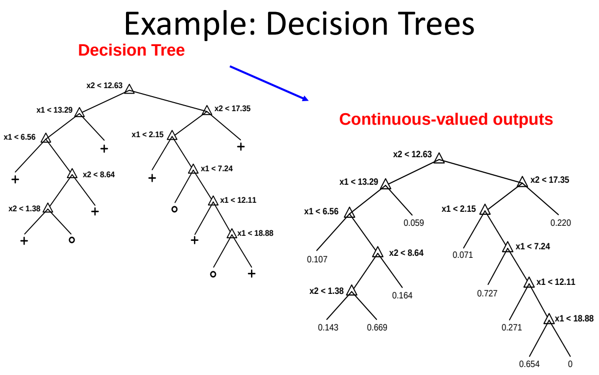
System B has reached 100 % TPR before system A.

System B has less FAR when it reached 100% TPR.

Thus, system is a better choice in this case.

Also, area under the curve B within the dashed square is higher than that of curve A.

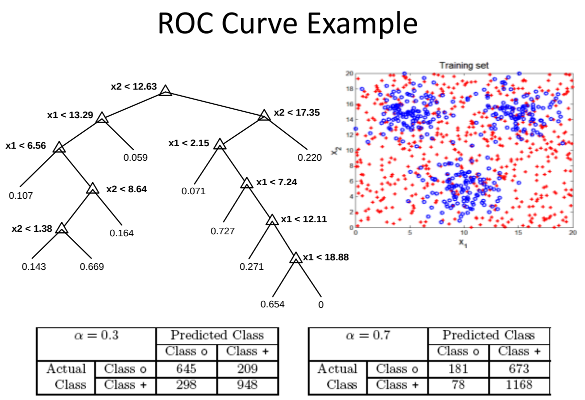
So, system B is optimal choice here.



Example: Decision Trees

Decision Tree

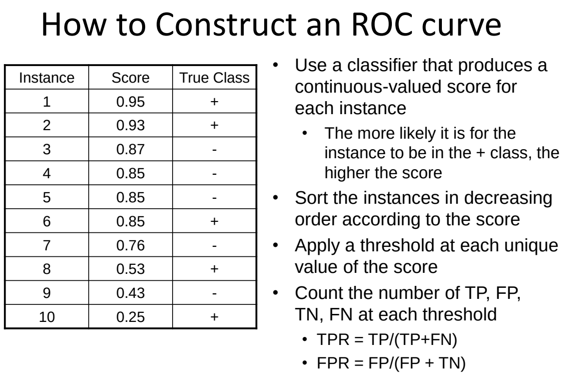
Continuous-valued outputs



ROC Curve Example

$\alpha = 0.3$		Predicted Class	
		Class o	Class +
Actual Class	Class o	645	209
	Class +	298	948

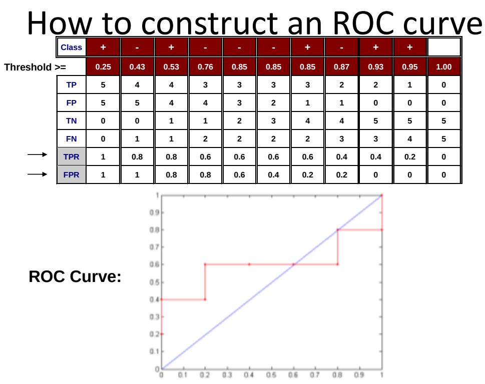
$\alpha = 0.7$		Predicted Class	
		Class o	Class +
Actual Class	Class o	181	673
	Class +	78	1168



How to Construct an ROC curve

Instance	Score	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

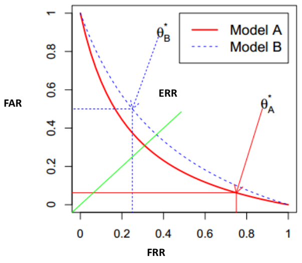
- Use a classifier that produces a continuous-valued score for each instance
  - The more likely it is for the instance to be in the + class, the higher the score
- Sort the instances in decreasing order according to the score
- Apply a threshold at each unique value of the score
- Count the number of TP, FP, TN, FN at each threshold
  - TPR = TP/(TP+FN)
  - FPR = FP/(FP + TN)



How to construct an ROC curve

	Class	+	-	+	-	-	-	+	-	+	+	
Threshold >=		0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
	TP	5	4	4	3	3	3	3	2	2	1	0
	FP	5	5	4	4	3	2	1	1	0	0	0
	TN	0	0	1	1	2	3	4	4	5	5	5
	FN	0	1	1	2	2	2	2	3	3	4	5
→	TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
→	FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

ROC Curve:

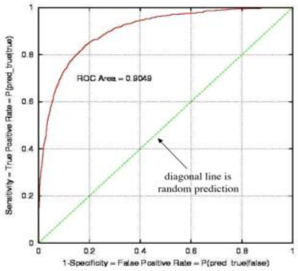


The two ROCs are shown in the above graph along with their respective EERs as θB* and θA* .
The line intersecting the two ROCs is the EER line for both the models.

Potential Risk

The best model appears to always be model A, since its curve is always below that of model B. Moreover, computing the BEP of models A and B yields the same conclusion. Remember that each point of the ROC corresponds to a particular setting of the threshold θ. However, in real applications, θ needs to be decided prior to seeing the test set. This is in general done using equation (1) , (2), using some development data (obviously different from the test set).
Hence, obtained EER by equation(2) on a development set, may not correspond to the EER on the test set. There are many reasons that could yield such mismatch, the simplest being that assuming the test and development sets to come from the same distribution but be of fixed (non-infinite) size, the estimate of (2) on one set is not guaranteed to be the same as the estimate on the other set.

ROCs assume that the training error will reflect the expected generalization error: this is true when the size of the data is huge, but false in the general case. Furthermore, real applications often suffer from an additional mismatch between training and test conditions which should be reflected in the procedure, something that ROCs don't consider.
To overcome this drawback in ROCs various other graphs like Expected Performance Curves are used.



Properties of ROC

- **Slope is non-increasing**
- **Each point on ROC represents different tradeoff (cost ratio) between false positives and false negatives**
- **Slope of line tangent to curve defines the cost ratio**
- **ROC Area represents performance averaged over all possible cost ratios**
- **If two ROC curves do not intersect, one method dominates the other**
- **If two ROC curves intersect, one method is better for some cost ratios, and other method is better for other cost ratios**

AUC-ROC Curve

<u>What is AUC - ROC Curve?</u>

AUC-ROC curve: Area Under the Receiver Operating Characteristic curve.
AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.
The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

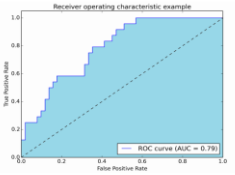
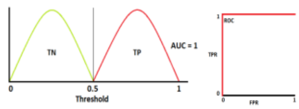


Fig: ROC curve with AUC section (blue border). The dashed line represent as random predictor (baseline model) with AUC=.5

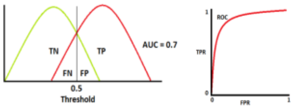
How to speculate the performance of the model?

An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means model has no class separation capacity.
Let's interpret above statements.
As we know, ROC is a curve of probability. So lets plot the distributions of those probabilities:
Note: Red distribution curve is of the positive class (patients with disease) and green distribution curve is of negative class(patients with no disease).

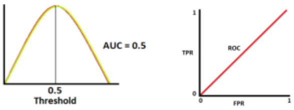
<u>Case1 :</u> **AUC=1 ,This is an ideal situation. When two curves don't overlap at all means model has an ideal measure of separability. It is perfectly able to distinguish between positive class and negative class.**



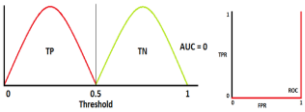
Case2: **AUC=.7, When two distributions overlap, we introduce type 1 and type 2 error. When AUC is 0.7, it means there is 70% chance that model will be able to distinguish between positive class and negative class.**



Case3: **AUC=.5, When AUC is approximately 0.5, model has no discrimination capacity to distinguish between positive class and negative class.-Random prediction**



Case4: **AUC=0, This is the worst situation. When AUC is approximately 0, model is actually reciprocating the classes. It means, model is predicting negative class as a positive class and vice versa.**



Relationship between AUC and model performance

ROC Area:

1.0: perfect prediction
0.9: excellent prediction
0.8: good prediction
0.7: mediocre prediction
0.6: poor prediction
0.5: random prediction
<0.5: something wrong!