Lecture 07
Performance Analysis

Topics Covered in this Presentation

- Error Estimation
- Statistical Estimation
- Performance Estimation
- ROC Curve

Performance Estimation

Performance Estimation of a Classifier

- Predictive accuracy works fine, when the classes are balanced
  - That is, every class in the data set are equally important
- In fact, data sets with imbalanced class distributions are quite common in many real life applications
- When the classifier classified a test data set with imbalanced class distributions then, predictive accuracy on its own is not a reliable indicator of a classifier's effectiveness.

Effectiveness of Predictive Accuracy

- Example:
- Given a data set of stock markets, we want to classify them as “good” and “worst”.
- Suppose, in the data set, out of 100 entries, 98 belong to “good” class and only 2 are in “worst” class.
- With this data set, if classifier’s predictive accuracy is 0.98, a very high value!
- Here, there is a high chance that 2 “worst” stock markets may incorrectly be classified as “good”
- On the other hand, if the predictive accuracy is 0.02, then none of the stock markets may be classified as “good”

Imbalanced Datasets

- Thus, when the classifier classified a test data set with imbalanced class distributions, then predictive accuracy on its own is not a reliable indicator of a classifier's effectiveness.

- This necessitates an alternative metrics to judge the classifier.

- Before exploring them, we introduce the concept of **Confusion matrix.**

Confusion Matrix

- A confusion matrix for a two classes (+, -) is shown below.

	C1	C2
C1	True positive	False negative
C2	False positive	True negative

	+	-
+	++	+-
-	-+	--

- There are four quadrants in the confusion matrix, which are symbolized as below.
  - True Positive (TP: f++) : The number of instances that were positive (+) and correctly classified as positive (+v).
  - False Negative (FN: f+-): The number of instances that were positive (+) and incorrectly classified as negative (-). It is also known as **Type 2 Error**.
  - False Positive (FP: f-+): The number of instances that were negative (-) and incorrectly classified as (+). This also known as **Type 1 Error**.
  - True Negative (TN: f--): The number of instances that were negative (-) and correctly classified as (-).

Confusion Matrix

- Np = TP (f++) + FN (f+-)

  = is the total number of positive instances.

- Nn = FP(f-+) + Tn(f--)

  = is the total number of negative instances.

- $N = Np + Nn$

  = is the total number of instances.

- (TP + TN) denotes the number of correct classification
- (FP + FN) denotes the number of errors in classification.
- For a perfect classifier FP = FN = 0, that is, there would be **no Type 1 or Type 2 errors**.

Confusion Matrix: Example

A classifier is built on a dataset regarding Good and Worst classes of stock markets. The model is then tested with a test set of 10000 unseen instances. The result is shown in the form of a confusion matrix.

Class	Good	Worst	Total	Rate(%)
Good	6954	46	7000	99.34
Worst	412	2588	3000	86.27
Total	**7366**	**2634**	**10000**	**95.52**

Predictive accuracy?

Confusion Matrix for Multiclass Classifier

- Having m classes, confusion matrix is a table of size $m \times m$, where, element at (i, j) indicates the number of instances of class i but classified as class j.

- To have good accuracy for a classifier, ideally most diagonal entries should have large values with the rest of entries being close to zero.

- Confusion matrix may have additional rows or columns to provide total or recognition rates per class.

Confusion Matrix for Multiclass Classifier

Following table shows the confusion matrix of a classification problem with six classes labeled as C1, C2, C3, C4, C5 and C6.

Class	C1	C2	C3	C4	C5	C6
C1	52	10	7	0	0	1
C2	15	50	6	2	1	2
C3	5	6	6	0	0	0
C4	0	2	0	10	0	1
C5	0	1	0	0	7	1
C6	1	3	0	1	0	24

Predictive accuracy?

Confusion Matrix for Multiclass Classifier

- In case of multiclass classification, sometimes one class is important enough to be regarded as positive with all other classes combined together as negative.
- Thus a large confusion matrix of m*m can be concised into 2*2 matrix.

Example 11.6: *m*×*m* CM to 2×2 CM

- For example, the CM shown in Example 11.5 is transformed into a CM of size 2×2 considering the class C1 as the positive class and classes C2, C3, C4, C5 and C6 combined together as negative.

Class	+	-
+	52	18
-	21	123

How we can calculate the predictive accuracy of the classifier model in this case?

Are the predictive accuracy same as in the previous example?

Performance Evaluation Metrics

- We now define a number of metrics for the measurement of a classifier.
  - In our discussion, we shall make the assumptions that there are only two classes: + (positive) and – (negative)
  - Nevertheless, the metrics can easily be extended to multi-class classifiers (with some modifications)
- **True Positive Rate (TPR):** It is defined as the fraction of the positive examples predicted correctly by the classifier.

$$TPR = \frac{TP}{P} = \frac{TP}{TP+FN} = \frac{f_{++}}{f_{++}+f_{+-}}$$

  - This metrics is also known as ***Recall***, ***Sensitivity*** or ***Hit rate***.
- **False Positive Rate (FPR):** It is defined as the fraction of negative examples classified as positive class by the classifier.

$$FPR = \frac{FP}{N} = \frac{FP}{FP+TN} = \frac{f_{-+}}{f_{-+}+f_{--}}$$

  - This metric is also known as ***False Alarm Rate***.

Performance Evaluation Metrics

- **False Negative Rate (FNR):** It is defined as the fraction of positive examples classified as a negative class by the classifier.

$$FNR = \frac{FN}{P} = \frac{FN}{TP + FN} = \frac{f_{+-}}{f_{++} + f_{+-}}$$

- **True Negative Rate (TNR):** It is defined as the fraction of negative examples classified correctly by the classifier

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = \frac{f_{--}}{f_{--} + f_{-+}}$$

  - This metric is also known as ***Specificity***.

Performance Evaluation Metrics

- **Positive Predictive Value (PPV):** It is defined as the fraction of the positive examples classified as positive that are really positive

$$PPV = \frac{TP}{TP + FP} = \frac{f_{++}}{f_{++} + f_{-+}}$$

  - It is also known as *Precision*.

- **F_1 Score (F_1):** Recall (r) and Precision (p) are two widely used metrics employed in analysis, where detection of one of the classes is considered more significant than the others.

  - It is defined in terms of (r or TPR) and (p or PPV) as follows.

$$F_1 = \frac{2r.p}{r + p} = \frac{2TP}{2TP + FP + FN}$$

$$= \frac{2f_{++}}{2f_{++} + f_{+-} + f_{-+}} = \frac{2}{\frac{1}{r} + \frac{1}{p}}$$

Note

- F_1 represents the harmonic mean between recall and precision
- High value of F_1 score ensures that both Precision and Recall are reasonably high.

Performance Evaluation Metrics

- More generally, F_β score can be used to determine the trade-off between Recall and Precision as

$$F_\beta = \frac{(\beta + 1)rp}{r + \beta p} = \frac{(\beta + 1)TP}{(\beta + 1)TP + \beta FN + FP}$$

- Both, Precision and Recall are special cases of F_β when $\beta = 0$ and $\beta = 1$, respectively.

$$F_\beta = \frac{TP}{TP + FP} = Precision$$

$$F_\alpha = \frac{TP}{TP + FN} = Recall$$

Performance Evaluation Metrics

- A more general metric that captures Recall, Precision as well as is defined in the following.

$$F_\omega = \frac{\omega_1 TP + \omega_4 TN}{\omega_1 TP + \omega_2 FP + \omega_3 FN + \omega_4 TN}$$

Metric				
Recall	1	1	0	1
Precision	1	0	1	0
			1	0

Metric	ω_1	ω_2	ω_3	ω_4
Recall	1	1	0	1
Precision	1	0	1	0
F_β	$\beta+1$	β	1	0

Note

- In fact, given *TPR*, *FPR*, *p* and *r*, we can derive all others measures.
- That is, these are the universal metrics.

Predictive Accuracy (ε)

- It is defined as the fraction of the number of examples that are correctly classified by the classifier to the total number of instances.

$$\varepsilon = \frac{TP + TN}{P + N}$$

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

$$= \frac{f_{++} + f_{--}}{f_{++} + f_{+-} + f_{\mp} + f_{--}}$$

- This accuracy is equivalent to F_w with $w_1 = w_2 = w_3 = w_4 = 1$.

Error Rate ($\bar{\varepsilon}$)

- The error rate $\bar{\varepsilon}$ is defined as the fraction of the examples that are incorrectly classified.

$$\bar{\varepsilon} = \frac{FP + FN}{P + N}$$

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

$$= \frac{f_{+-} + f_{-+}}{f_{++} + f_{+-} + f_{-+} + f_{--}}$$

Note

$\bar{\varepsilon} = 1 - \varepsilon$.

Accuracy, Sensitivity and Specificity

- Predictive accuracy (ε) can be expressed in terms of sensitivity and specificity.
- We can write

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

$$= \frac{TP + TN}{P + N}$$

$$\varepsilon = \frac{TP}{P} \times \frac{P}{P + N} + \frac{TN}{N} \times \frac{N}{P + N}$$

Thus,

$$\varepsilon = \text{Sensitivity} \times \frac{P}{P+N} + \text{Specificity} \times \frac{N}{P+N}$$

Analysis with Performance Measurement Metrics

- Based on the various performance metrics, we can characterize a classifier.
- We do it in terms of TPR, FPR, Precision and Recall and Accuracy
- **Case 1: Perfect Classifier**

When every instance is correctly classified, it is called the perfect classifier. In this case, $TP = P$, $TN = N$ and CM is

$$TPR = \frac{P}{P} = 1$$

$$FPR = \frac{0}{N} = 0$$

$$Precision = \frac{P}{P} = 1$$

$$F_1\ Score = \frac{2 \times 1}{1+1} = 1$$

$$Accuracy = \frac{P+N}{P+N} = 1$$

		Predicted Class	
		+	-
Actual class	+	P	0
	-	0	N

Analysis with Performance Measurement Metrics

- **Case 2: Worst Classifier**

When every instance is wrongly classified, it is called the worst classifier. In this case, $TP = 0$, $TN = 0$ and the CM is

$$TPR = \frac{0}{P} = 0$$

$$FPR = \frac{N}{N} = 1$$

$$Precision = \frac{0}{N} = 0$$

$F_1\ Score$ = Not applicable as $Recall + Precision = 0$

$$\text{Accuracy} = \frac{0}{P+N} = 0$$

		Predicted Class	
		+	-
Actual class	+	0	P
	-	N	0

Analysis with Performance Measurement Metrics

- **Case 3: Ultra-Liberal Classifier**

The classifier always predicts the + class correctly. Here, the False Negative (FN) and True Negative (TN) are zero. The CM is

$TPR = \quad = 1$

$FPR = \quad = 1$

$Precision =$

$F1\ Score =$

Accuracy = $\quad = 0$

		Predicted Class	
		+	-
Actual class	+	P	0
	-	N	0

Analysis with Performance Measurement Metrics

Case 4: Ultra-Conservative Classifier

This classifier always predicts the - class correctly. Here, the False Negative (FN) and True Negative (TN) are zero. The CM is

$$TPR = \frac{0}{P} = 0$$

$$FPR = \frac{0}{N} = 0$$

$$Precision = \text{Not applicable}$$

$$(\text{as } TP + FP = 0)$$

$$F_1\ Score = \text{Not applicable}$$

$$\text{Accuracy} = \frac{N}{P+N} = 0$$

		Predicted Class	
		+	-
Actual class	+	0	p
	-	0	N

Predictive Accuracy versus TPR and FPR

One strength of characterizing a classifier by its *TPR* and *FPR* is that they do not depend on the relative size of *P* and *N*.

- The same is also applicable for *FNR* and *TNR* and others measures from CM.

- In contrast, the *Predictive Accuracy*, *Precision*, *Error Rate*, *F1 Score*, etc. are affected by the relative size of *P* and *N*.

- *FPR*, *TPR*, *FNR* and *TNR* are calculated from the different rows of the CM.
  - On the other hand Predictive Accuracy, etc. are derived from the values in both rows.
- This suggests that *FPR*, *TPR*, *FNR* and TNR are more effective than *Predictive Accuracy*, etc.

ROC Curves

ROC Curves

- ROC is an abbreviation of Receiver Operating Characteristic come from the signal detection theory, developed during World War 2 for analysis of radar images.

- In the context of classifier, ROC plot is a useful tool to study the behaviour of a classifier or comparing two or more classifiers.

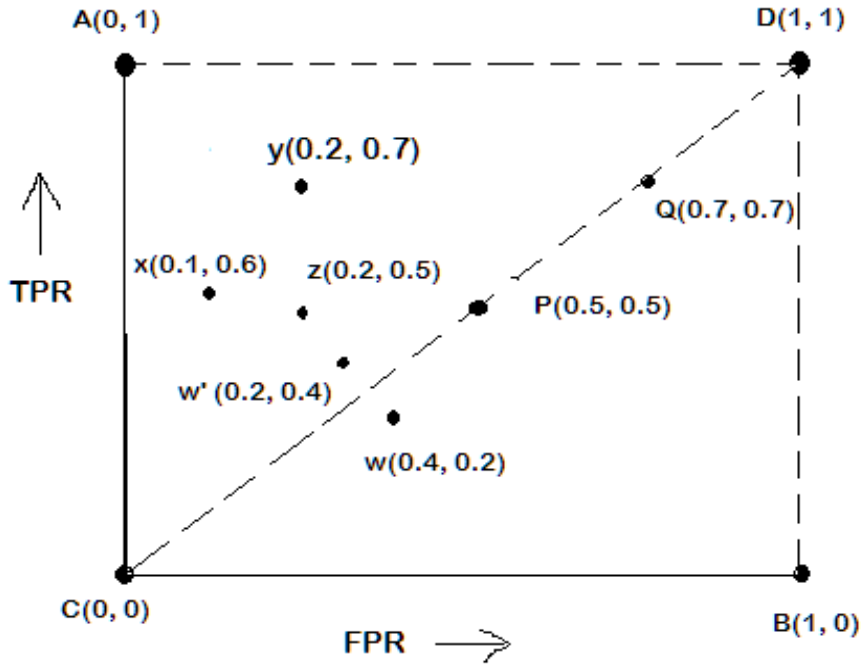
- A ROC plot is a two-dimensional graph, where, X-axis represents FP rate (FPR) and Y-axis represents TP rate (TPR).

- Since, the values of FPR and TPR varies from 0 to 1 both inclusive, the two axes thus from 0 to 1 only.

- Each point *(x, y)* on the plot indicating that the FPR has value *x* and the TPR value *y*.

ROC Plot

- A typical look of ROC plot with few points in it is shown in the following figure.

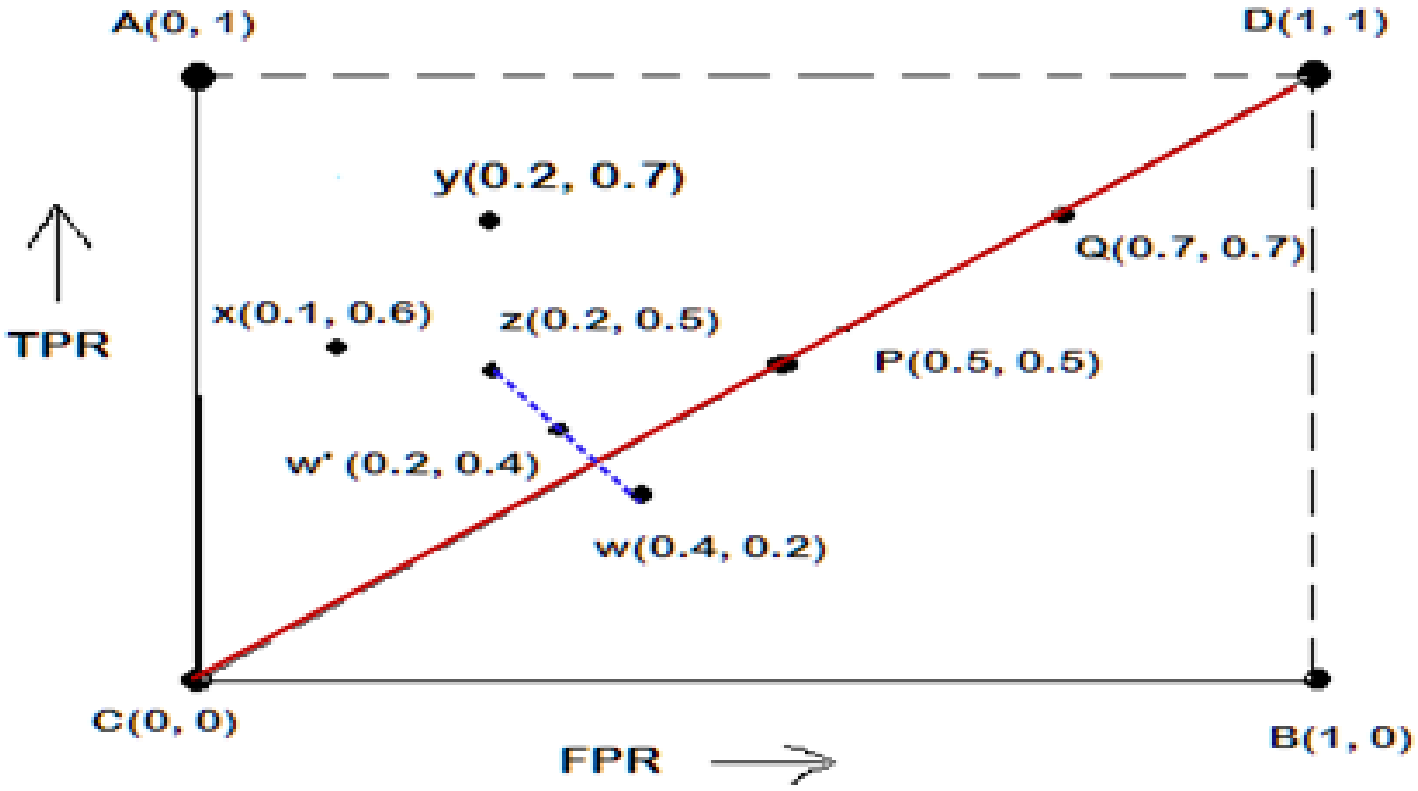


Note the four cornered points are the four extreme cases of classifiers

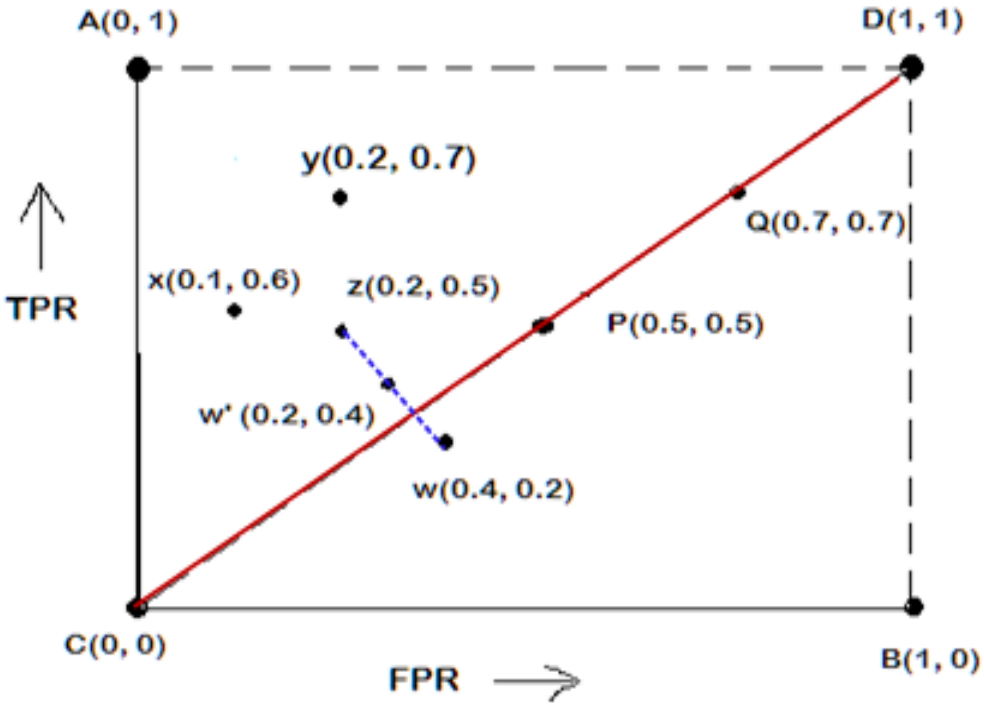
Identify the four extreme classifiers.

Interpretation of Different Points in ROC Plot

- Le us interpret the different points in the ROC plot.



Interpretation of Different Points in ROC Plot

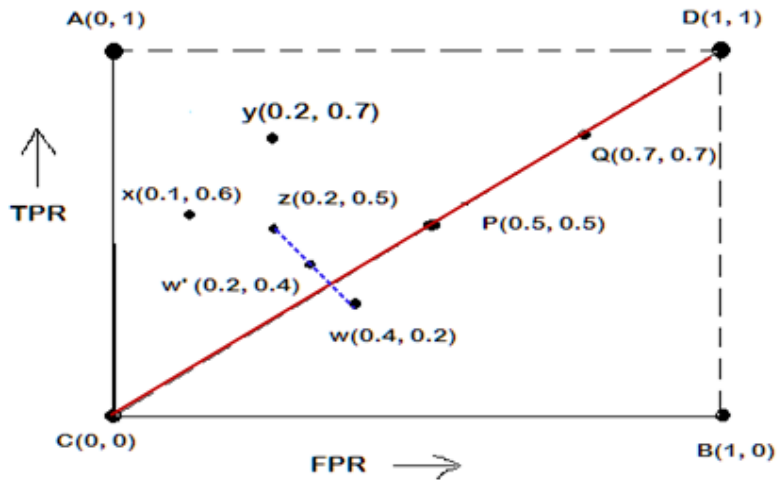


- The diagonal line joining point C(0,0) and D(1,1) corresponds to random guessing.
- Random guessing means that a record is classified as positive (or negative) with a certain probability

- Suppose, a test set contacting N+ positive and N- negative instances. Suppose, the classifier guesses any instances with probability p
- Thus, the random classifier is expected to correctly classify p.N+ of the positive instances and p.N- of the negative instances
- Hence, TPR = FPR = p
- Since TPR = FPR, the random classifier results reside on the main diagonals

Interpretation of Different Points in ROC Plot

- Let us interpret the different points in the ROC plot.

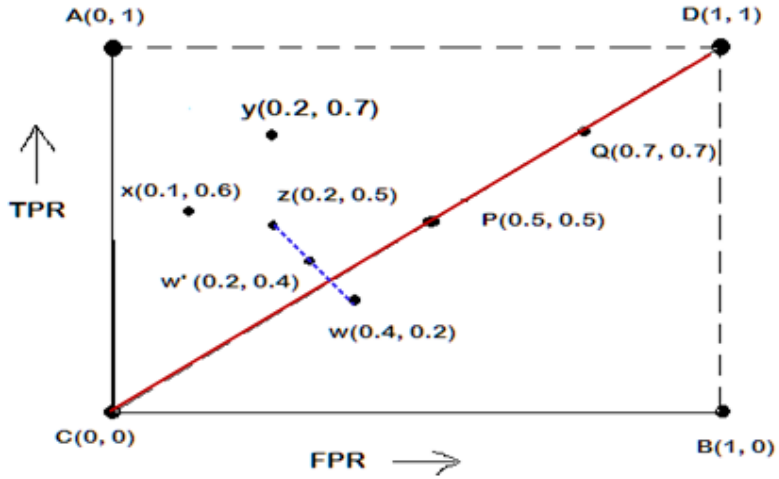


The points on the upper diagonal region

- All points, which reside on upper-diagonal region are corresponding to classifiers "good" as their TPR is as good as FPR (i.e., FPRs are lower than TPRs)
- Here, X is better than Z as X has higher TPR and lower FPR than Z.
- If we compare X and Y, neither classifier is superior to the other

Interpretation of Different Points in ROC Plot

- Let us interpret the different points in the ROC plot.

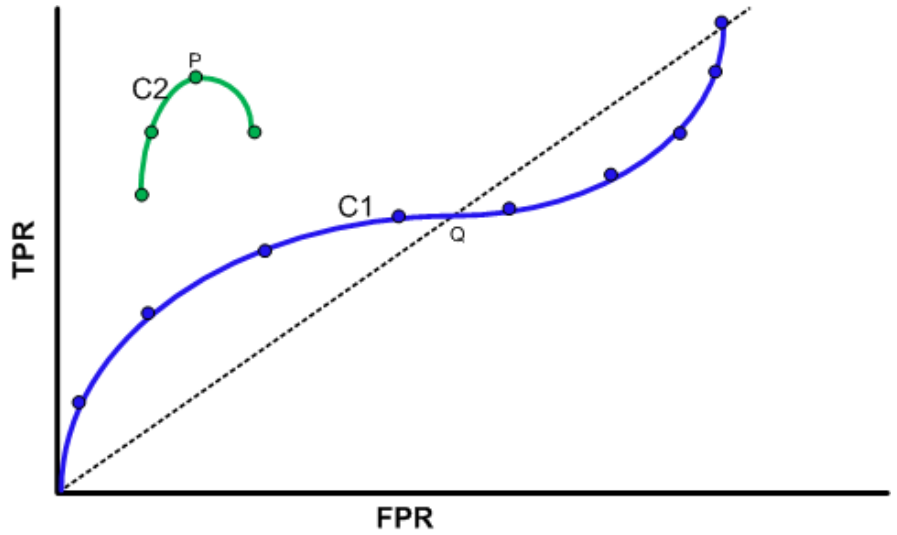


The points on the lower diagonal region

- The Lower-diagonal triangle corresponds to the classifiers that are worst than random classifiers
- Note: A classifier that is worst than random guessing, simply by reversing its prediction, we can get good results.
  - W'(0.2, 0.4) is the better version than W(0.4, 0.2), W' is a mirror reflection of W

Tuning a Classifier through ROC Plot

- Using ROC plot, we can compare two or more classifiers by their TPR and FPR values and this plot also depicts the trade-off between TPR and FPR of a classifier.



Examining ROC curves can give insights into the best way of tuning parameters of classifier.

- For example, in the curve C2, the result is degraded after the point P. Similarly for the observation C1, beyond Q the settings are not acceptable.

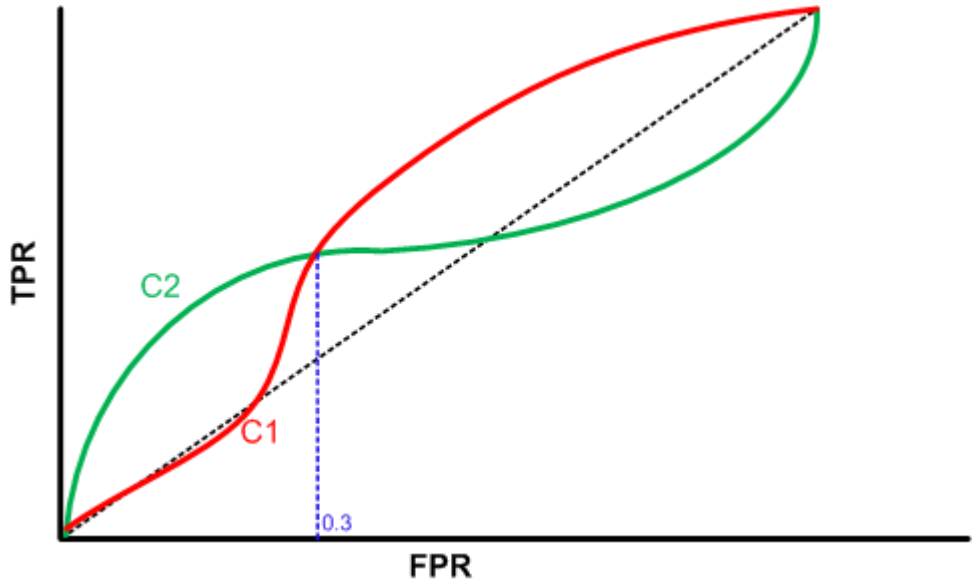
Comparing Classifiers trough ROC Plot

Two curves C1 and C2 are corresponding to the experiments to choose two classifiers with their parameters.

Here, C1 is better than C2 when FPR is less than 0.3.

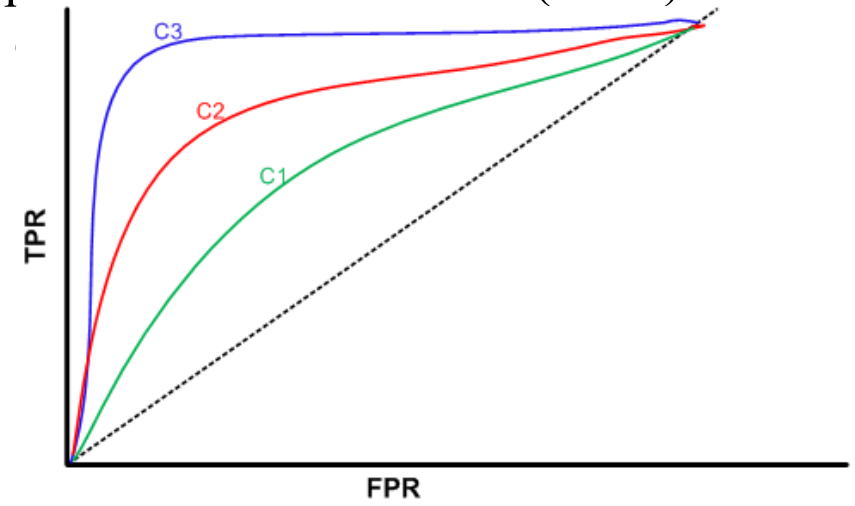
However, C2 is better, when FPR is greater than 0.3.

- Clearly, neither of these two classifiers dominates the other.



Comparing Classifiers trough ROC Plot

- We can use the concept of "area under curve" (AUC) as a better method to compare two or more



If a model is perfect, then its AUC = 1.

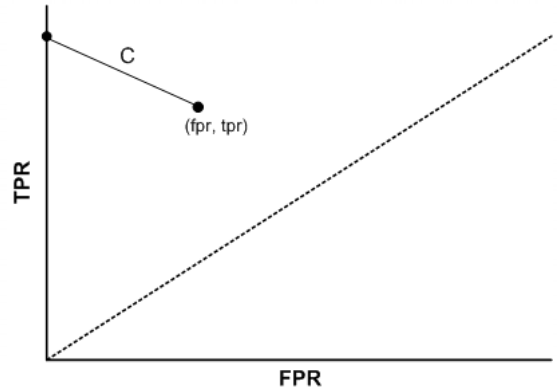
If a model simply performs random guessing, then its AUC = 0.5

A model that is strictly better than other, would have a larger value of AUC than the other.

Here, C3 is best, and C2 is better than C1 as AUC(C3)>AUC(C2)>AUC(C1).

A Quantitative Measure of a Classifier

- The concept of ROC plot can be extended to compare quantitatively using Euclidean distance measure.
- See the following figure for an explanation.



- Here, *C(fpr, tpr)* is a classifier and $\boldsymbol{\delta}$ denotes the Euclidean distance between the best classifier (0, 1) and C. That is,
  - $\boldsymbol{\delta} = \sqrt{fpr^2 + (1 - tpr)^2}$

A Quantitative Measure of a Classifier

- The smallest possible value of $\boldsymbol{\delta}$ is 0
- The largest possible values of $\boldsymbol{\delta}$ is $\sqrt{2}$ (when (fpr = 1 and tpr = 0).
- We could hypothesise that the smaller the value of $\boldsymbol{\delta}$, the better the classifier.
- $\boldsymbol{\delta}$ is a useful measure, but does not take into account the relative importance of true and false positive rates.
- We can specify the relative importance of making TPR as close to 1 and FPR as close 0 by a weight *w* between 0 to 1.
- We can define weighted $\boldsymbol{\delta}$ (denoted by $\boldsymbol{\delta}_w$) as

$$\boldsymbol{\delta}_w = \sqrt{(1-w)fpr^2 + w(1-tpr)^2}$$

Note

- If $w = 0$, it reduces to $\boldsymbol{\delta}_w$= fpr, i.e., FP Rate.
- If $w = 1$, it reduces to $\boldsymbol{\delta}_w = 1 - tpr$, i.e., we are only interested to maximizing TP Rate.