Performance Evaluation

- Confusion Matrix:

		Detected	
		Positive	**Negative**
Actual	**Positive**	A: True Positive	B: False Negative
	Negative	C: False Positive	D: True Negative

- Recall or Sensitivity or True Positive Rate (TPR):
  - It is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$\text{Recall} = \frac{A}{A+B}$$

- Accuracy (AC):
  - *AC*: is the proportion of the total number of predictions that were correct.
  - It is determined using the equation:

$$\text{Accuracy} = \frac{A+D}{A+B+C+D}$$

  - Error rate (misclassification rate)= 1 – AC

- The false positive rate (FPR) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$\text{FPR} = \frac{C}{C + D}$$

- The true negative rate (TNR) or Specificity:
  - It is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

$$\text{TNR} = \frac{D}{C + D}$$

- The false negative rate (FNR):
  - It is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

$$\text{FNR} = \frac{B}{A + B}$$

- Precision:
  - P is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$\text{Precision} = \frac{A}{A + C}$$

- F-measure:
  - The F-Measure computes some average of the information retrieval precision and recall metrics.
  - Why F-measure?
    - An arithmetic mean does not capture the fact that a (50%, 50%) system is often considered better than an (80%, 20%) system
  - F-measure is computed using the harmonic mean:

    Given n points, x_1, x_2, …, x_n, the harmonic mean is:

$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$$

  - So, the harmonic mean of Precision and Recall:

$$\frac{1}{F} = \frac{1}{2}\left(\frac{1}{R} + \frac{1}{P}\right) = \frac{P+R}{2PR}$$

  - The computation of F-measure:
    - Each cluster is considered as if it were the result of a query and each class as if it were the desired set of documents for a query
    - We then calculate the recall and precision of that cluster for each given class.
    - The F-measure of cluster *j* and class *i* is defined as follows:

$$F_{ij} = \frac{2 * \text{Recall}(i, j) * \text{Precision}(i, j)}{\text{Precision}(i, j) + \text{Recall}(i, j)}$$

- The F-measure of a given clustering algorithm is then computed as follows:

$$F - \text{measure} = \sum \frac{n_i}{n} \max(\{F_{ij}\})$$

Where *n* is the number of documents in the collection and n_i is the number of documents in cluster i.

- Note that the computed values are between 0 and 1 and a larger F-Measure value indicates a higher classification/clustering quality.

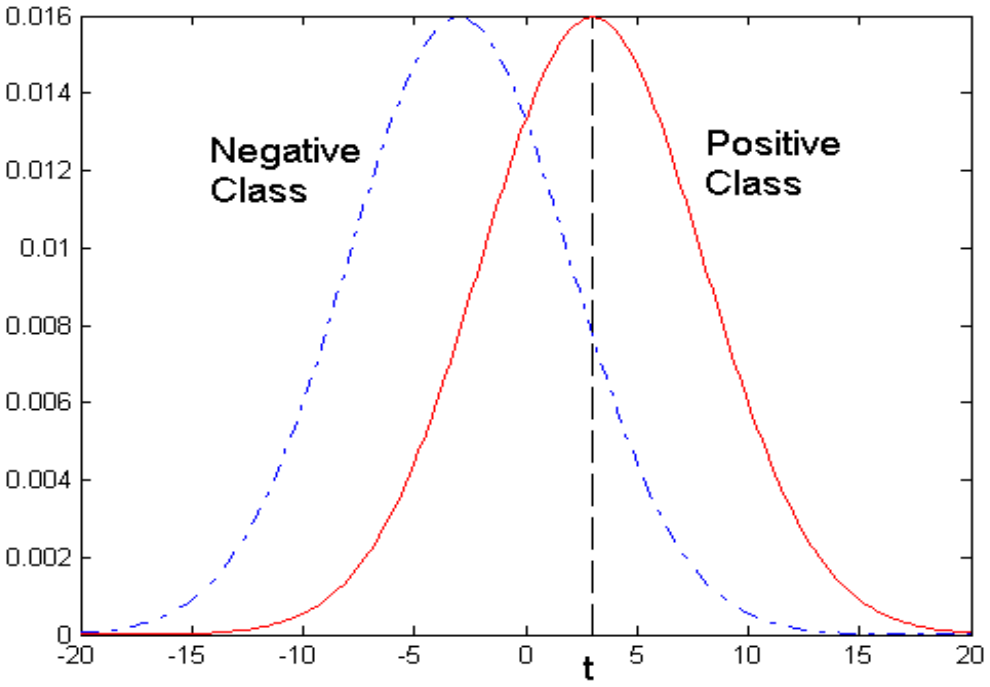
- Receiver Operating Characteristic (ROC) Curve:
  - It is a graphical approach for displaying the tradeoff between true positive rate (TPR) and false positive rate (FPR) of a classifier:

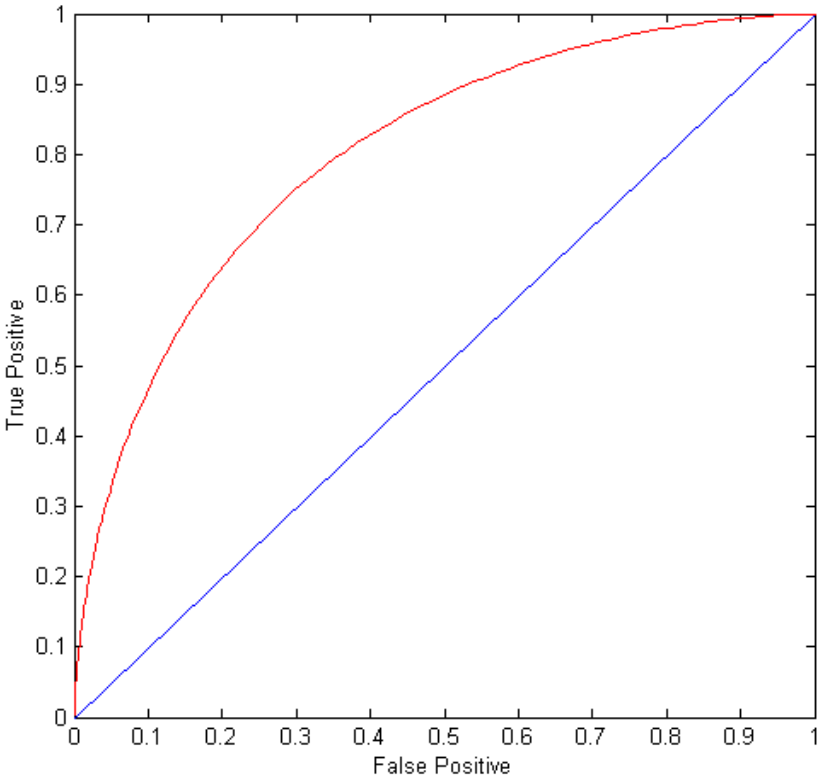
    TPR = positives correctly classified/total positives

    FPR = negatives incorrectly classified/total negatives

  - TPR is plotted along the y axis
  - FPR is plotted along the x axis

- Performance of each classifier represented as a point on the ROC curve





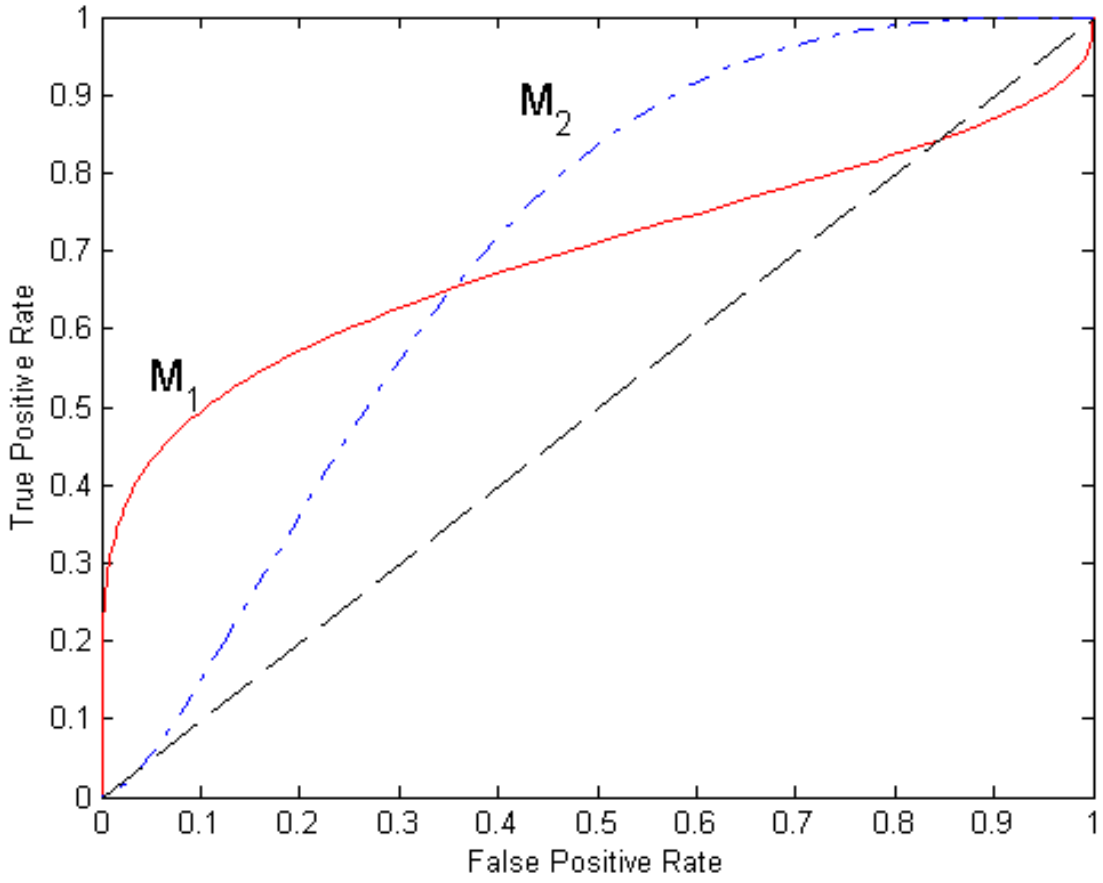
- Important Points: (TP,FP)
    - (0,0): declare everything to be negative class
    - (1,1): declare everything to be positive class
    - (1,0): ideal

- Diagonal line:
    - Random guessing

- Area Under Curve (AUC):
    - It provides which model is better on the average.
    - Ideal Model: area = 1

- If the model is simply performs random guessing, then its area under the curve would equal 0.5.
- A model that is better than another would have a larger area.

Example:



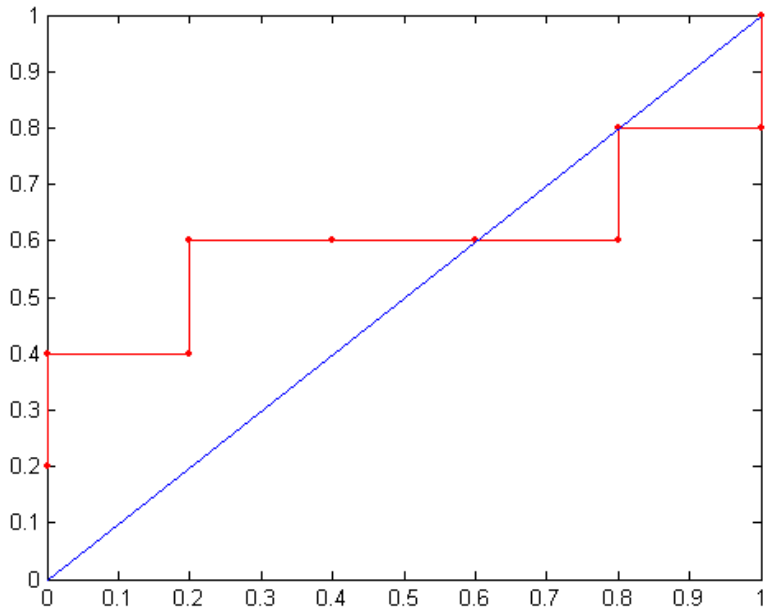
- No model consistently outperform the other
  - M1 is better for small FPR
  - M2 is better for large FPR

- Example: (Kumar et al.)
  - Compute P(+|A) which is a numeric value that represents the degree to which an instance is a member of a class. In other words, it is the probability or ranking of the predicted class of each data point.
  - P(+|A) is the posterior probability as defined in Bayesian classifier.

Instance	P(+|A)	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 classifier that produces posterior probability for each test instance P(+|A)
- Sort the instances according to P(+|A) in decreasing order
- Apply threshold at each unique value of P(+|A)
  - Count the number of TP, FP, TN, FN at each threshold
  - TP rate, TPR = TP/(TP+FN)
  - FP rate, FPR = FP/(FP + TN)

Class	+	-	+	-	-	-	+	-	+	+	
	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



Clustering Only

- Intra-Cluster Similarity (ICS):
  - It looks at the similarity of all the data points in a cluster to their cluster centroid.
  - It is calculated as arithmetic mean of all of the data point-centroid similarities.
  - Given a set of k clusters, ICS is defined as follows:

$$ICS = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{|C_i|}\sum_{d_j \in C_i} sim(d_j, c_i)$$

    Where c_i is the centroid of cluster C_i.

  - A good clustering algorithm maximizes intra-cluster similarity.

- Centroid Similarity (CS):
  - It computes the similarity between the centroids of all clusters.
  - Given a set of k clusters, CS is defined as follows:

$$CS = \sum_{i=1}^{k}\sum_{j=1}^{k} sim(c_i, c_j)$$