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:

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

- 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.

- Cohen's Kapa Measure:
  - Some studies involve the need for some degree of subjective interpretation by observers. For example:
    - Doctors' MRI reading
    - Observing animals' behavior
  - **Expected Frequency (EF)**: Agreements between observers may occur by chance
  - The kappa score considers that two or more observers may agree or disagree just by chance. Hence:
    - A kappa of 1 indicates perfect agreement
    - A kappa of 0 indicates agreement equivalent to chance
    - A Kappa score greater than 0.6 can be considered as substantial
  - Example:

G: Good -- N: No change -- W: Worst															
Animals	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Observer A	G	N	G	N	N	G	W	N	G	W	W	G	N	W	N
Observer B	G	N	W	G	N	G	W	N	G	N	W	W	N	W	W

- **Step 1: Contingency Table**

		Observer A		
		G	N	W
Observer B	G	3	1	0
	N	0	4	1
	W	2	1	3

- **Step 2: Compute the overall totals for rows and columns**

		Observer A			
		G	N	W	Total
Observer B	G	3	1	0	4
	N	0	4	1	5
	W	2	1	3	6
	Total	5	6	4	15

- **Step 3: Compute the total number of agreements**:

		Observer A			
		G	N	W	Total
Observer B	G	3	1	0	4
	N	0	4	1	5
	W	2	1	3	6
	Total	5	6	4	15

Total number of agreements: 3 + 4 + 3 = 10
The level of agreement = 10/15 = 0.66

- **Step 4: Compute the EF for the agreements**:

  - Compute the **EF** for each agreement (Diagonal):

$$\mathrm{EF(G)} = \frac{\text{Row Total} * \text{Column Total}}{\text{Overall Total}}$$

$$= \frac{5*4}{15} = \frac{20}{15} = \frac{4}{3} = 1.33$$

$$\mathrm{EF(N)} = \frac{6*5}{15} = 2$$

$$\mathrm{EF(W)} = \frac{4*6}{15} = \frac{24}{15} = 1.6$$

- Compute the sum of the **EFs**:

$$\sum \mathrm{EFs} = 1.33 + 2 + 1.6 = 4.93$$

- Compute Kappa:

$$\mathrm{Kappa} = \frac{\sum \mathrm{agreements} - \sum \mathrm{EFs}}{\mathrm{Total\ of\ Data\ points} - \sum \mathrm{EFs}}$$

$$\mathrm{Kappa} = \frac{10 - 4.93}{15 - 4.93} = \frac{5.07}{10.07} = 0.5$$

Kappa	Agreement
<0	Less Than Chance Agreement
0.0-0.2	Sight Agreement
0.2-0.4	Fair Agreement
0.4-0.6	Moderate Agreement
0.6-0.8	Substantial Agreement
0.8-0.99	Almost Perfect Agreement
1	Perfect Agreement

Source: Landis, J.R. and Koch, GG. (1977) ‘The Measurement of observer agreement for categorical data’. Biometrics, 33 159-74

- **Performance of Regression Model**
    - Evaluate the regression problem's accuracy.
    - **Mean Absolute Error or MAE**
        - It measures the error between the actual value and predicted value:

MAE = Predicted Value – Actual Value

        - Absolute difference means that if the result has a negative sign, it is ignored.
        - The lower the MAE score the better since we want to a smaller value between the predicted and actual values.
        - The closer MAE is to 0, the more accurate the model is
        - Note that MAE cannot be compared across different models and datasets.

    - **Mean Squared Error (MSE):**

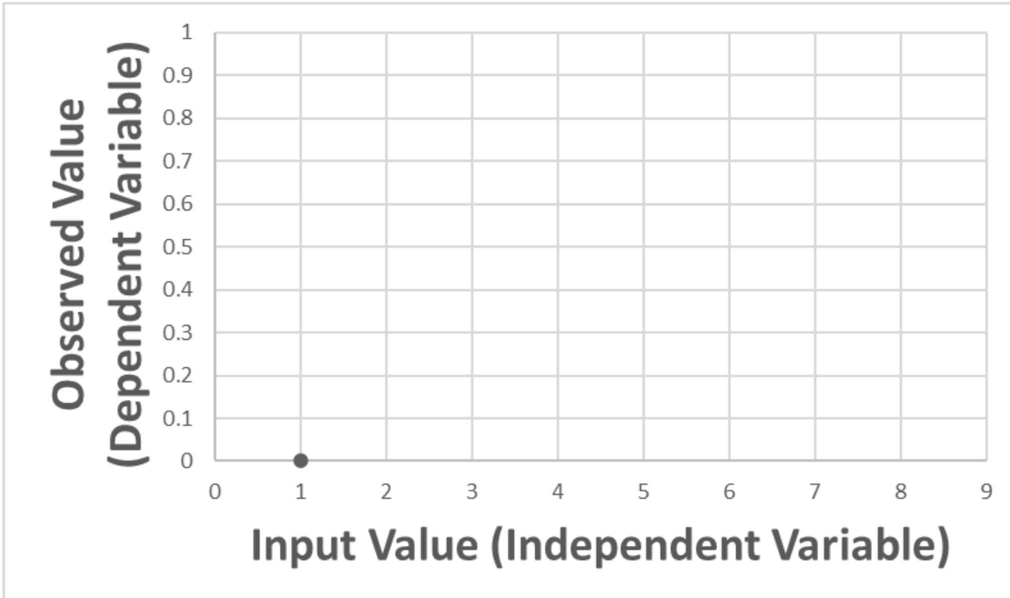
$$\text{MAE} = \frac{\sum_{i=1}^{N}(\text{Predicted Value} - \text{Actual value})^2}{\text{N}}$$

    - **Root Mean Square Error (RMSE):**
        - It measures the error of a model in predicting quantitative data.
        - It used to evaluate the accuracy of regression model

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\text{Predicted Value} - \text{Actual value})^2}{\text{N}}}$$

    - **The R-squared**
        - It is also called the **coefficient of determination**
        - It explains the degree to which the actual input explains the variation of predicted variables.
        - It provides information about the goodness of fit of a model.

- A higher R-squared indicates a better fit for the model.

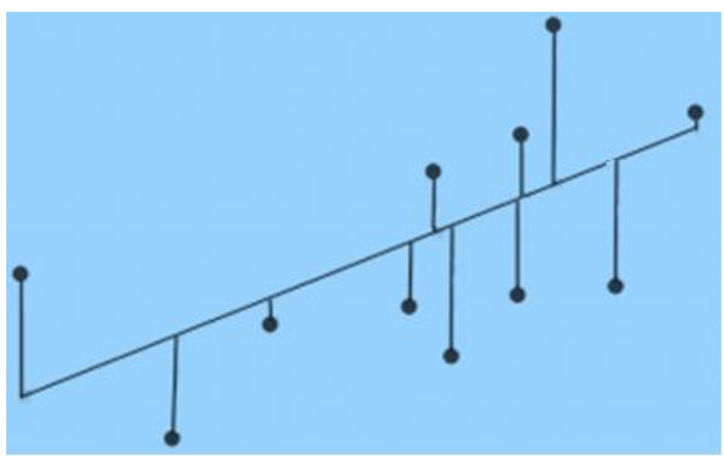


- The widely used equation is:

$$R^2 = \frac{1 - \text{Sum Squared Regression (SSR)}}{\text{Total Sum of Squares (SST)}}$$

SSR is also called the sum of residuals, which is the distance from regression line to each data point:

$$\text{SSR} = \sum (\text{Observed Value} - \text{Fitted Value})^2$$



To compute the Total Sum of Squares (SST), you need to first calculate the mean value of the observed values $\overline{\text{Observed Value}}$

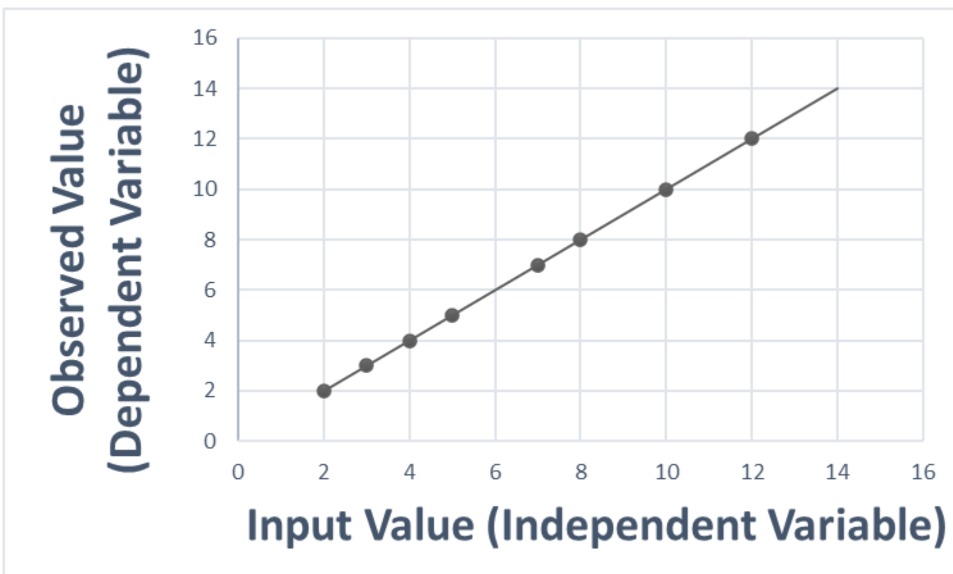
$$\text{SST} = \sum(\text{Observed Value} - \overline{\text{Observed Value}})^2$$

Then,

$$\text{R}^2 = \frac{1 - \sum(\text{Observed Value} - \text{Fitted Value})^2}{\sum(\text{Observed Value} - \overline{\text{Fitted Value}})^2}$$

- Interpretation of R2 value:

When R = 1:



- R-Squared vs. RMSE:
  - R-squared gives good indication on how well the model fit.
  - RMSE is better if you are interested in how your model will predict values for new data

- Receiver Operating Characteristic (ROC) Curve:

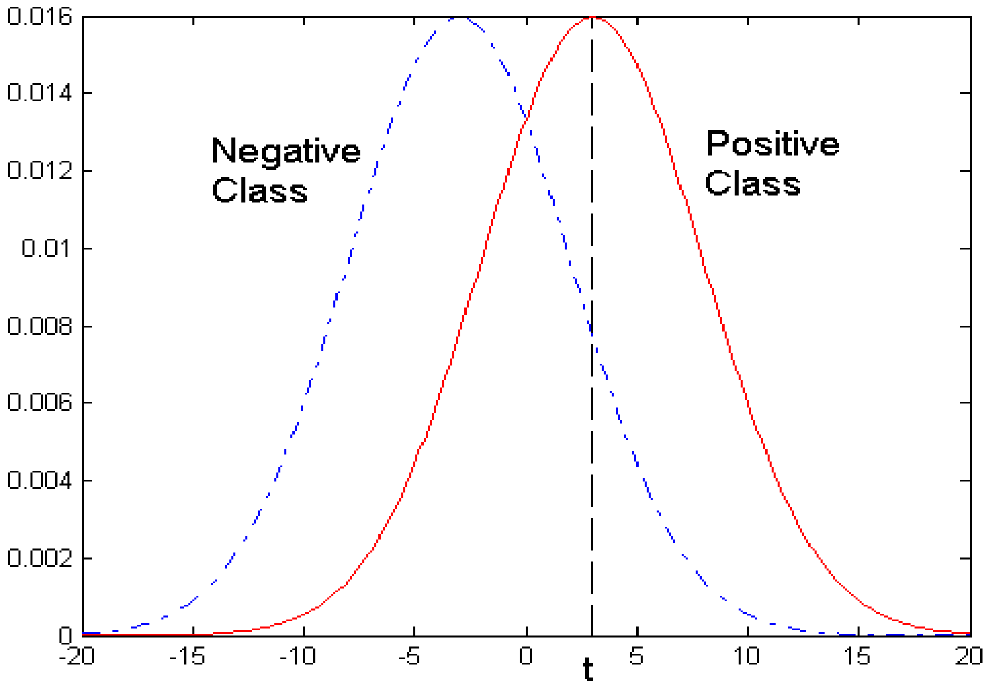
  - It is a graphical approach for displaying the tradeoff between true positive rate (TPR) and false positive rate (FPR) of a classifier:

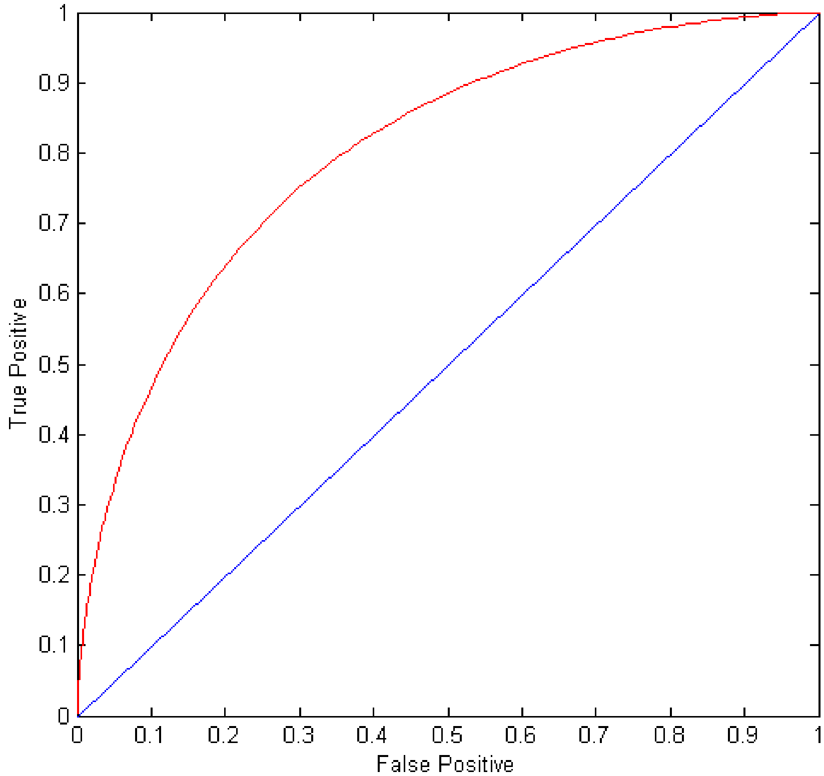
    TPR = positives correctly classified/total positives

    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

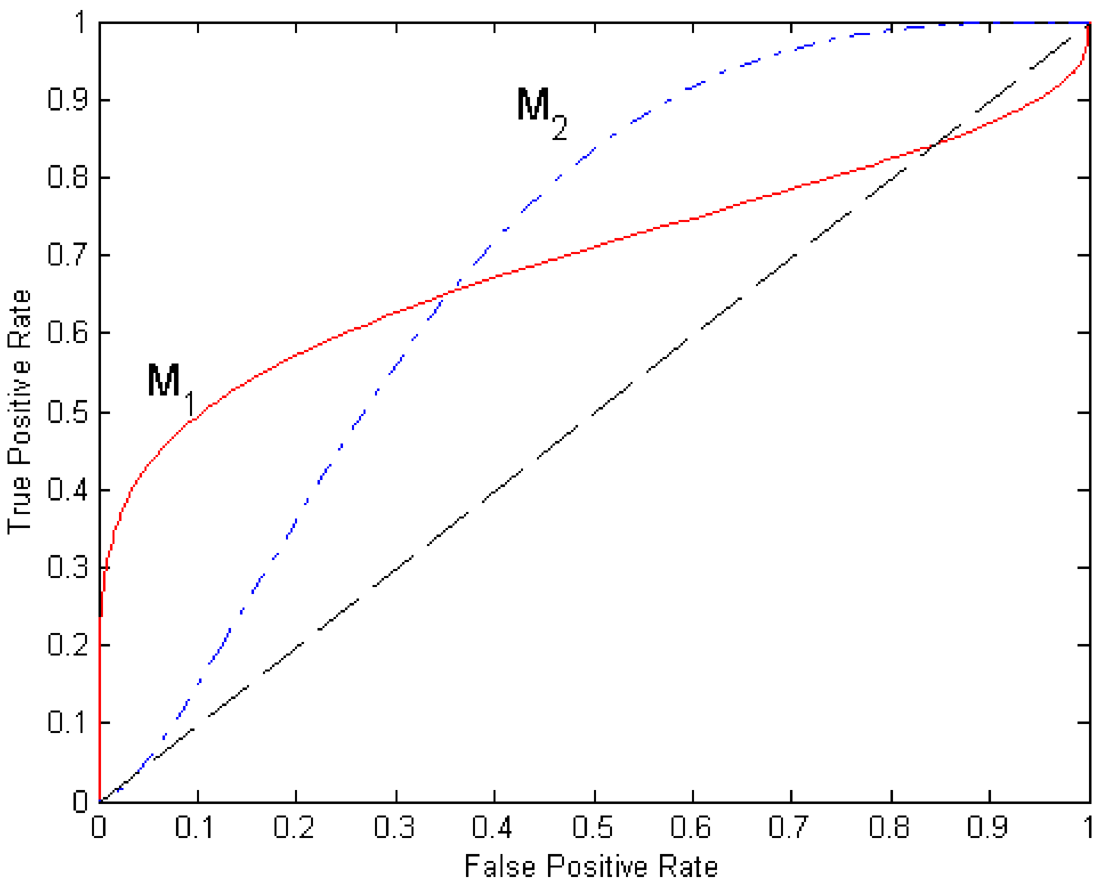




- Important Points: (TP,FP)
    - (0,0): declare everything to be negative class
    - (1,1): declare everything to be positive class
    - (1,0): ideal
- Diagonal line:
    - Random guessing
- Area Under Curve (AUC):
    - It provides which model is better on the average.
    - Ideal Model: area = 1

- 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

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:

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