Performance Evaluation

- Confusion Matrix:

<table>
<thead>
<tr>
<th>Actual</th>
<th>Detected</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>A: True Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>Negative</td>
<td>C: False Positive</td>
</tr>
</tbody>
</table>

- 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:

\[
FPR = \frac{C}{C + D}
\]

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

\[
TNR = \frac{D}{C + D}
\]

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

\[
FNR = \frac{B}{A + B}
\]

• Precision:
  o 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:
  o The F-Measure computes some average of the information retrieval precision and recall metrics.
  o 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
  o F-measure is computed using the harmonic mean:

Given n points, \( x_1, x_2, \ldots, x_n \), the harmonic mean is:

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

o So, the harmonic mean of Precision and Recall:

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

o 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:
\[ F_{ij} = \frac{2 \times \text{Recall}(i, j) \times \text{Precision}(i, j)}{\text{Precision}(i, j) + \text{Recall}(i, j)} \]

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

  \[
  TPR = \frac{\text{positives correctly classified}}{\text{total positives}}
  \]

  \[
  FPR = \frac{\text{negatives incorrectly classified}}{\text{total negatives}}
  \]

  o TPR is plotted along the y axis
  o 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.)
  o 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.
  o $P(+|A)$ 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 $P(+|A)$
  o Count the number of TP, FP, TN, FN at each threshold
  o TP rate, TPR = TP/(TP+FN)
  o FP rate, FPR = FP/(FP + TN)
<table>
<thead>
<tr>
<th>Class</th>
<th>+</th>
<th>-</th>
<th>+</th>
<th>-</th>
<th>+</th>
<th>-</th>
<th>+</th>
<th>-</th>
<th>+</th>
<th>+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.25</td>
<td>0.43</td>
<td>0.53</td>
<td>0.76</td>
<td>0.85</td>
<td>0.85</td>
<td>0.85</td>
<td>0.87</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>TP</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>FP</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TN</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>FN</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>TPR</td>
<td>1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.4</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>FPR</td>
<td>1</td>
<td>1</td>
<td>0.8</td>
<td>0.8</td>
<td>0.6</td>
<td>0.4</td>
<td>0.2</td>
<td>0.2</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
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:

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

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:

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