Information Retrieval and Text Mining

Assessment in Information Retrieval

Nuno Escudeiro (nfe@isep.ipp.pt)
Ricardo Almeida (ral@isep.ipp.pt)

Session outline

1. Assessment in IR
    - Precision, Recall
    - Precision/Recall curve
    - F_β Score
    - F_1 Score
    - Confusion Matrix
    - ROC curve, AUC

Learning outcomes

At the end of this session we will be able to:

- Compute and explain precision, recall and F_β score
- Represent and analyze Precision/Recall curves
- Represent and analyze ROC curves and AUC

1. Assessment in IR

A general overview of the IR assessment: precision, recall, F_β and F_1 score.

Graphically representing ROC curve, AUC, Precision/Recall curve

Precision and Recall

- Precision, Recall calculation:

  - **Precision :** number of retrieved docs, from all that were obtained, that are relevant to the user's information need

$$\textbf{Precision} = \frac{Number\ of\ Relevant\ Documents\ Retrieved}{Total\ Number\ of\ Documents\ Retrieved}$$

  - **Recall :** number of relevant docs in collection that are retrieved

$$\textbf{Recall} = \frac{Number\ of\ Relevant\ Documents\ Retrieved}{Total\ Number\ of\ Relevant\ Documents\ in\ the\ Corpus}$$

Precision and Recall

- Let us consider the following situation:
  - A group of human experts identified R1, the set of documents that are deemed relevant for a given query Q1: R1={d3, d5, d8, d23, d33, d48, d50, d66, d74, d92}
  - Using an automatic retrieval algorithm for the same query Q1 over the same corpus, the ranking obtained was the following:

1	**d74**	6	**d33**	11	d101
2	d21	7	d123	12	d90
3	**d48**	8	**d5**	13	**d92**
4	d24	9	d22	14	d100
5	d18	10	d12	15	d1

  - The retrieved documents that are relevant to Q1 are represented in bold.

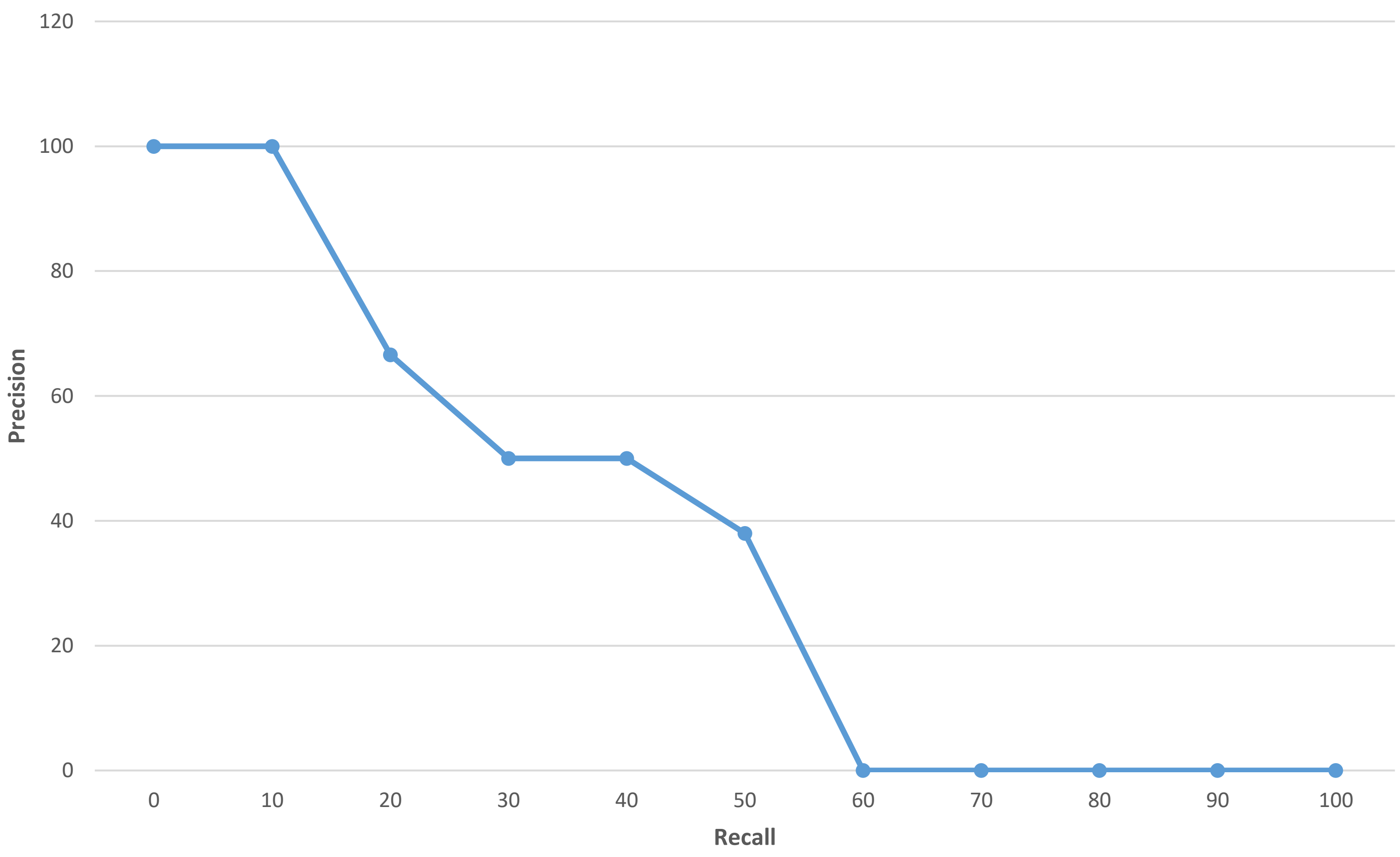
Precision and Recall

1	**d74**	6	**d33**	11	d101
2	d21	7	d123	12	d90
3	**d48**	8	**d5**	13	**d92**
4	d24	9	d22	14	d100
5	d18	10	d12	15	d1

- Since the first element of the result set obtained with the retrieval algorithm is relevant ($d74 \in R1$) the precision after assessing the first document retrieved is 1/1 (100%) and the corresponding recall is 1/10 (10%).
- The second relevant element that was retrieved is in the third position of the result set; at this stage, after assessing the top three retrieved documents, we have a precision of 2/3 (66,6%) and a recall of 2/10 (20%)
- The third relevant element is found at the sixth position in the result set; after assessing the first six results retrieved, we have a precision of 3/6 (50%) and a recall of 3/10 (30%)
- What about the precision and recall of the 4th and 5th relevant document?

Precision/Recall curve

R	P
0	100
10	100
20	66,6
30	50
40	50
50	38
60	0
70	0
80	0
90	0
100	0



1.2 Confusion Matrix

Performance of classification.

Confusion Matrix

- Confusion Matrix : is a table used to evaluate the performance of a classification model.
- It allows visualization of the performance of a model by comparing predicted class labels with true class labels.
- The matrix is particularly useful for assessing the accuracy of a model's predictions.

Confusion Matrix

- How to compute a Confusion Matrix:
  1. You need a test dataset or a validation dataset with expected outcome values
  2. Make a prediction for each row in your test dataset
  3. From the expected outcomes and predictions count:
     1. The number of correct predictions for each class
     2. The number of incorrect predictions for each class, organized by the class that was predicted
  - These numbers are then organized into a table, or a matrix as follows:
    - Expected down the side: Each row of the matrix corresponds to a predicted class
    - Predicted across the top: Each column of the matrix corresponds to an actual class

Confusion Matrix

EMAIL	Actual category (labelled by an expert)	Classifier prediction
1	Spam	Spam
2	Non spam	Non spam
3	Non spam	Non spam
4	Non spam	Non spam
5	Non spam	Non spam
6	Spam	Non spam
7	Spam	Non spam
8	Non spam	Non spam
9	Non spam	Spam
10	Spam	Spam

Confusion Matrix

- Example:

	Predicted Non Spam	Predicted Spam
Actual Non Spam	5	1
Actual Spam	2	2

Confusion Matrix

- Example:

	Predicted Spam	Predicted Non Spam
Actual Spam	2	2
Actual Non Spam	1	5

Confusion Matrix

- Example:

	Positive Prediction	Negative Prediction
Positive Class	True (Positive)	False (Negative)
Negative Class	False (Positive)	True (Negative)

1.3 F-score

Recall and Precision aggregated.

F_β Score

- **F_β score:**
  - Is a metric used in information retrieval to evaluate the performance of ranking algorithms, particularly in relevance feedback scenarios.
  - It's an extension of the F_1 score, which combines precision and recall into a single measure.
  - $F_\beta=(1+\beta^2)\text{x}\dfrac{Precision \times Recall}{(\beta^2 \times Precision)+Recall}$ where:
    - β is a parameter that controls the relative importance of precision and recall.
    - Precision is the ratio of relevant documents retrieved to the total number of documents retrieved.
    - Recall is the ratio of relevant documents retrieved to the total number of relevant documents.
    - The beta parameter represents the ratio of recall importance to precision importance. beta > 1 gives more weight to recall, while beta < 1 favors precision
    - For example, beta = 2 makes recall twice as important as precision, while beta = 0.5 does the opposite. Asymptotically, beta -> +inf considers only recall, and beta -> 0 only precision.

F_1 Score

- **F_1 score:** Is a metric used to evaluate the performance of a classification model, particularly when dealing with imbalanced classes. It is the harmonic mean of precision and recall, providing a single measure that balances both metrics.
  - It is very used in IR
  - $F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ onde

Precision$= \frac{True\ Positives}{True\ Positives + False\ Positives}$ **Recall**$= \frac{True\ Positives}{True\ Positives + False\ Negatives}$

Confusion Matrix

- Example:

	Predicted Non Spam	Predicted Spam
Actual Non Spam	5	1
Actual Spam	2	2

F_1 Score

- Example:
  - Suppose that we are trying to classify emails as either spam (positive) or non-spam (negative). After applying a classification algorithm, we obtain the following confusion matrix:

	Predicted Spam	Predicted Non Spam
Actual SPAM	1150	150
Actual Non Spam	200	8500

- True (Positives) (TP) = 1150 (number of correctly classified spam emails)
- False (Positives) (FP) = 200 (number of non-spam emails incorrectly classified as spam)
- False (Negatives) (FN) = 150 (number of spam emails incorrectly classified as non-spam)

F_1 Score

- Example:

	Predicted Spam	Predicted Non Spam
Actual SPAM	1150 (TP)	150 (FN)
Actual Non Spam	200 (FP)	8500 (TN)

- **Precision**= $\frac{1150}{1150+200} = \frac{1150}{1350} \approx 0.852$

- **Recall**= $\frac{1150}{1150+150} = \frac{1150}{1300} \approx 0.885$

- F1=2 x $\frac{0.852 \times 0.885}{0.852+0.885} = \frac{1.506}{1.737} \approx 0.866$

1.4 ROC curve

Binary classifier performance analysis.

Receiver Operating Characteristic(ROC) Curve

- Is a graphical plot that summarizes the performance of a binary classification model on the positive class across various threshold settings
- In IR are commonly used to evaluate the performance of binary classifiers
- Visualizes the trade-off between True Positive Rate (also known as recall) and False Positive Rate
- The x-axis indicates the False Positive Rate and the y-axis indicates the True Positive Rate.

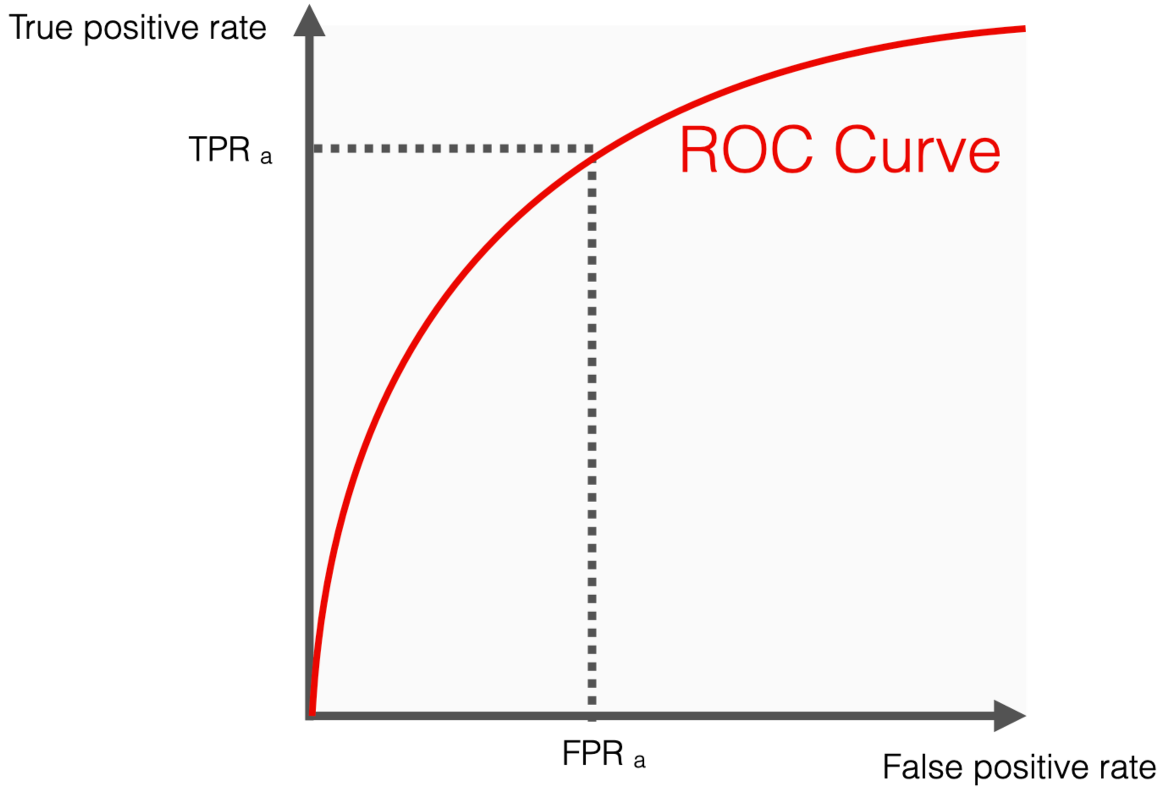
Receiver Operating Characteristic(ROC) Curve

- How to calculate True Positive and False Positive Rates
  - TPR=$\frac{TP}{TP+FN}$
  - FPR=$\frac{FP}{FP+TN}$

Receiver Operating Characteristic(ROC) Curve - Example

EMAIL	Actual category (labelled by an expert)	Classifier prediction
1	Spam	Spam
2	Non spam	Non spam
3	Non spam	Non spam
4	Non spam	Non spam
5	Non spam	Non spam
6	Spam	Non spam
7	Spam	Non spam
8	Non spam	Non spam
9	Non spam	Spam
10	Spam	Spam

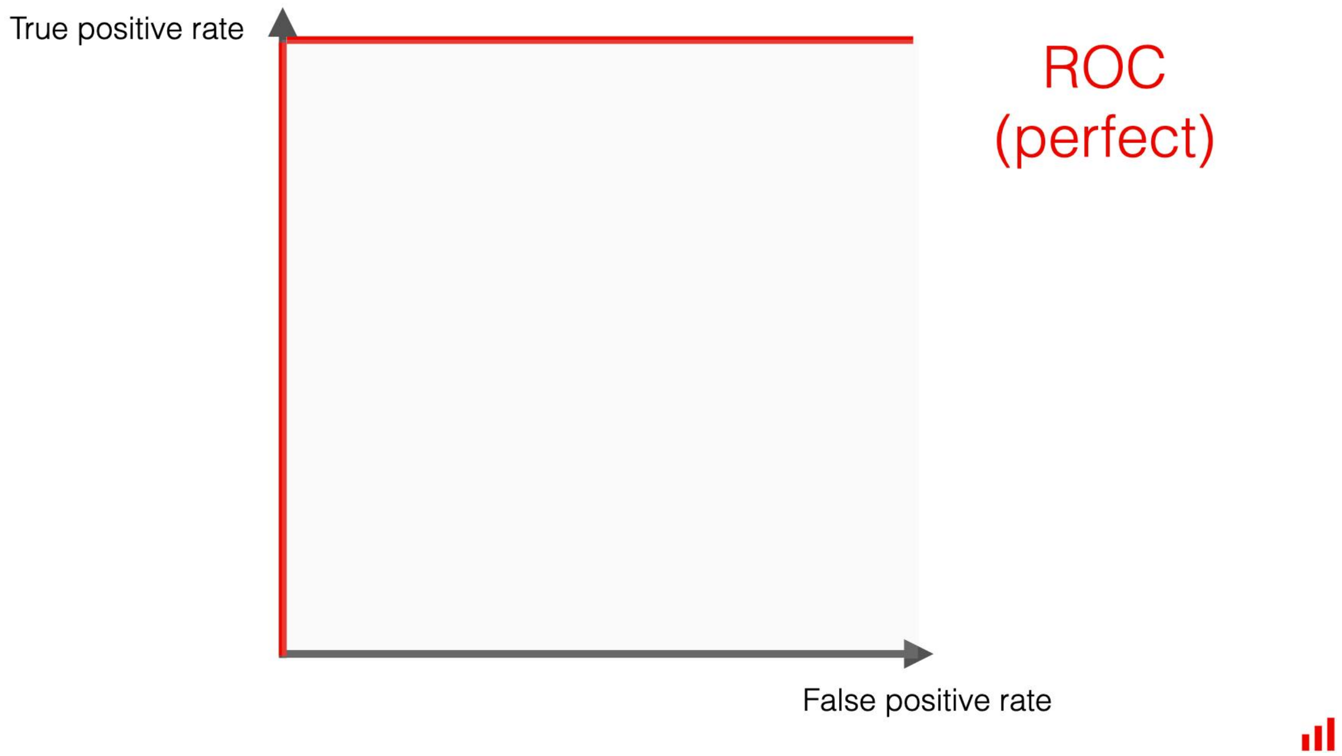
Receiver Operating Characteristic(ROC) Curve



https://www.evidentlyai.com/classification-metrics/explain-roc-curve

Receiver Operating Characteristic(ROC) Curve

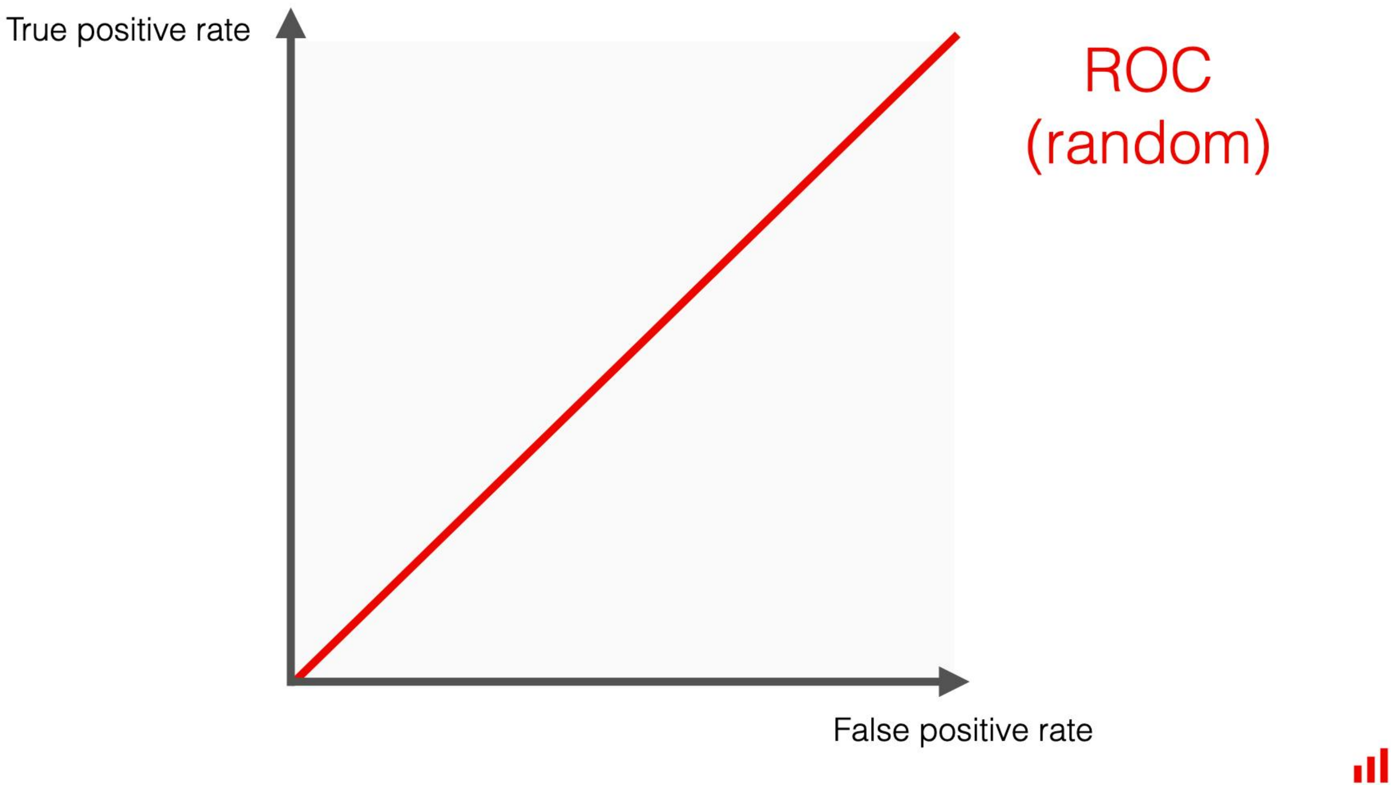
- Perfect Scenario



https://www.evidentlyai.com/classification-metrics/explain-roc-curve

Receiver Operating Characteristic(ROC) Curve

- Worst Scenario

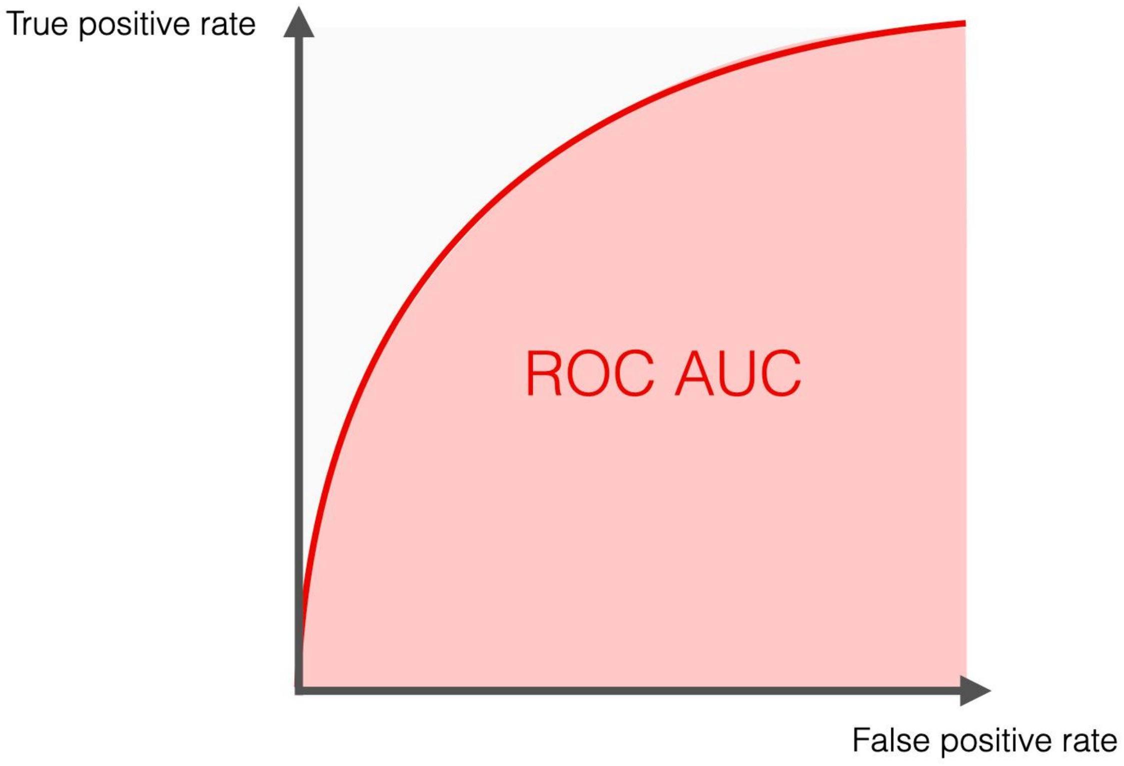


https://www.evidentlyai.com/classification-metrics/explain-roc-curve

ROC Area Under Curve (AUC) Score

- ROC AUC is a score that shows the quality (performance) of the classifier across all possible classification thresholds.
- To get the score, you must measure the area under the ROC curve.
- The ROC AUC score range from 0 to 1, with 0,5 indicating random guessing
- Shows how well the classifier distinguishes positive and negative classes

Receiver Operating Characteristic(ROC) Curve



https://www.evidentlyai.com/classification-metrics/explain-roc-curve

References

- https://users.dcc.uchile.cl/~rbaeza/mir2ed/
- https://machinelearningmastery.com/
- https://www.evidentlyai.com/classification-metrics/explain-roc-curve