

4.1 - Model Valuation and Selection

Giovanni Della Lunga
giovanni.dellalunga@unibo.it

Bologna - February-April, 2025

Accuracy Metrics

Accuracy Metrics: how to choose a Model?

- Let's consider a simple problem of binary classification. Example: you have to classify bank customer into good or bad wrt the credit worthiness;
- Probably the first thing you think is to measure the goodness of your model using the classification accuracy...
- ... but this would turn out to be a very bad idea

Why?

Accuracy Metrics: how to choose a Model?

Credit Classification Example

- Train logistic regression model (say $y=1$ if the client is bad and 0 otherwise);
- Find that you got 10% error on test set (90% correct classification);
- Only 8% of clients are bad clients. You have a very **unbalanced** or skewed sample;

```
function predictBadLoad(x)  
    return 0
```

This very stupid model has a classification accuracy of 92%!!!

Threshold

- The threshold is an arbitrarily decided upon point between 0.0 and 1.0 that serves as your "cutoff" for which predicted probabilities you want to consider a True or a False, a Yes or a No, a 1 or a 0.
- Who decides it?
- You do, or whomever the decision-maker happens to be.
- You may have assumed this threshold is naturally located right in the middle, at 0.5, but you can move that threshold.

Threshold

- Why would you want to do that?
- Reasons include:
 - a) You are conservative about your guesses, so you set the threshold for a "Yes" to 0.7 (or 70%, if you will). Anything predicted to have less than a 70% probability is just too risky for you.
 - b) Alternatively, a risk-taker may want to call anything over 0.35 probability a "Yes", so that they don't miss any opportunities.
 - c) Lastly, perhaps you want to use the threshold that gives the highest performance, for whatever metric you choose.

Confusion Matrix

- A confusion matrix is $N * N$ dimension matrix wherein one axis represents **Actual** label while the other axis represents **Predicted** label.
- Confusion Matrix is the most intuitive and basic metric from which we can obtain various other metrics like precision, recall, accuracy, F1 score, AUC - ROC.

		Prediction	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Confusion Matrix

- For a better understanding of what TP, FP, TN, and FN are, we will consider an example of: **If received mail is spam or ham.**
- **Positive** - Mail received is ham
- **Negative** - Mail received is spam
- \Rightarrow **True Positive (TP)**: It represents the predicted label is positive and also actual label is positive - correctly predicted. We predicted mail received is 'ham' (positive) and actual mail received is also 'ham' (positive).
- \Rightarrow **True Negative (TN)**: It represents the predicted label is negative and also actual label is negative - correctly predicted. We predicted mail received is 'spam' (negative) and actual mail received is also 'spam' (negative).
- \Rightarrow **False Negative (FN)**: It represents the predicted label is negative but the actual label is positive - wrongly predicted. We predicted mail received is 'spam' (negative) but actual mail received is 'ham' (positive).
- \Rightarrow **False Positive (FP)**: It represents the predicted label is positive but the actual label is negative - wrongly predicted. We predicted mail received is 'ham' (positive) but actual mail received is 'spam' (negative).

Precision

- **General Definition:** Precision measures what proportion of predicted positive label is actually positive.
- **Precision** can be expressed in terms of True Positive and False Positive:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

- As 'False Positive' decreases, our precision increases and vice-versa
- When to use Precision?
- Precision is used when we want to mostly focus on false-positive i.e to decrease false-positive value thereby increase precision value.
- A question might arise why we want to mostly focus on false-positive and not false-negative. The answer to this question depends on the context.

Recall/Sensitivity

- **General Definition:** Recall measures what proportion of actual positive label is correctly predicted as positive.
- To explain recall and its use case, we shall consider 'Cancer Diagnosis' example i.e we have to predict if a patient is diagnosed with cancer or not.
- **Positive** - Patient diagnosed with cancer.
- **Negative** - Patient not diagnosed with cancer.
- Recall in terms of True Positive and False Negative:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

- From the above formula in the image, we can analyze that as 'False Negative' decreases, our recall increases and vice-versa.

Recall/Sensitivity

- When to use Recall?
- Recall is used when we want to mostly focus on false-negative i.e to decrease false negative value thereby increase recall value.
- A question might arise why we want to mostly focus on a false-negative and not false positive.
- To answer this question, let us consider 'Cancer Diagnosis' example...

Recall/Sensitivity

- **False Negative (FN)**

- It represents our predicted label is negative but the actual label is positive - wrongly predicted.
- Applying false negative on our example- it means we have predicted that the patient is not diagnosed with cancer but the actual patient is diagnosed with cancer.
- If this is the case, patient, as per prediction might not get treatment to cure cancer.
- But the truth is patient is diagnosed with cancer.
- Our wrong negative prediction will lead to death of a patient.
- So, **we mostly focus on false-negative** value and try to decrease it to the least possible value.

Recall/Sensitivity

- **False Positive (FP)**

- It represents our predicted label is positive but the actual label is negative - wrongly predicted.
- Applying false positive on our example- it means we have predicted that the patient is diagnosed with cancer but the actual patient is not diagnosed with cancer.
- If this is the case, patient, as per prediction will get check-up for cancer diagnosis.
- To his happiness, he will come to know that he is not diagnosed with cancer. Hurrah! He is free from cancer now.
- So, we don't much focus on false-positive value.

F1-Score

- F1-score is another one of the good performance metrics which leverages both precision and recall metrics.
- F1-score can be obtained by simply taking 'Harmonic Mean' of precision and recall.
- Unlike precision which mostly focuses on false-positive and recall which mostly focuses on false-negative, **F1-score focuses on both false positive and false negative.**

F1-Score

- F1-score in terms of Precision and Recall;
- The F-Score is the Harmonic Mean of Precision and Recall:

$$F = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

- Alternatively:

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

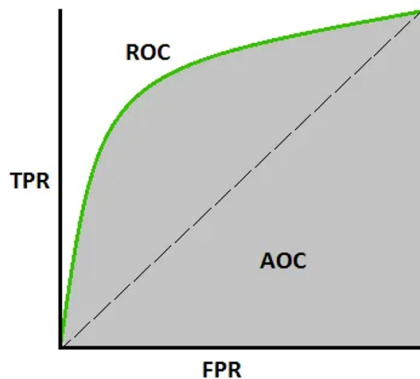
- When to use F1-score:
- As mentioned above, F1-score focuses on both false positive and false negative and try to decrease both false positive and false negative thereby increase F1-score.

AUC - ROC curve

- AUC - ROC is one of the most important performance metric used to check model performance.
- AUC - ROC is used for binary and also multi-class classification but mostly used in binary classification problems.
- In this lesson, we will consider a binary class classification.
- AUC-ROC is a graphical representation of model performance. ROC is a probability curve and AUC is the measure of separability.
- Depending on the threshold set, we can analyze how well our model has performed in separating two classes.
- Higher the AUC better is our model in separating two classes.

AUC - ROC curve

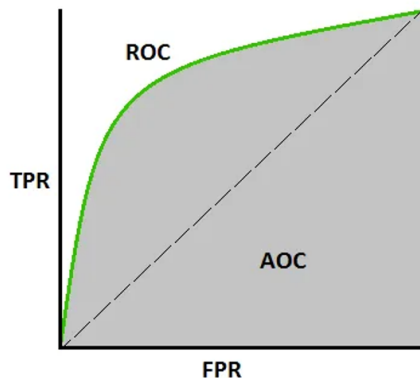
- Graphical representation of AUC - ROC
- Referring the image, we can see that AUC - ROC curve is plotted with FPR against TPR where FPR (False Positive Rate) is on X-axis while TPR (True Positive Rate) is on Y-axis.
- The green curve represents ROC curve while the area/region under ROC curve (green curve) represents AUC.



AUC - ROC curve

- **True Positive Rate (TPR):**
TPR is nothing but Recall / Sensitivity.
- The formula for TPR as follows:

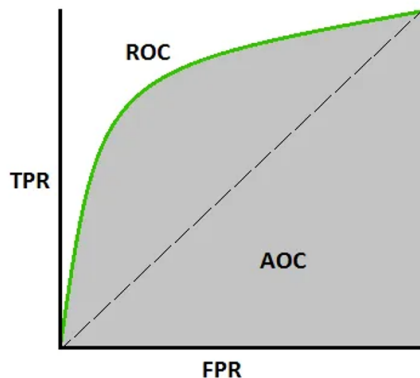
$$TPR = \frac{TP}{TP + FN}$$



AUC - ROC curve

- **False Positive Rate (FPR)**
- The formula for TPR as follows:

$$FPR = \frac{FP}{TN + FP}$$

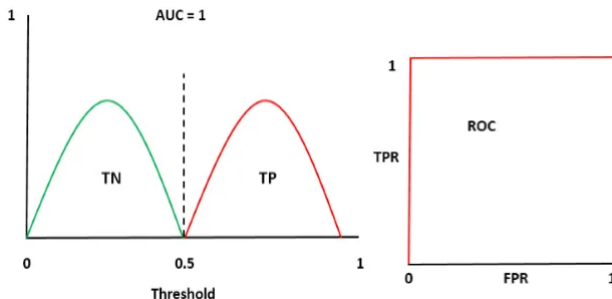


AUC - ROC curve

- Interpretation of AUC-ROC curve
- Let's now look into the analysis of binary class classification based on the AUC score and ROC curve...

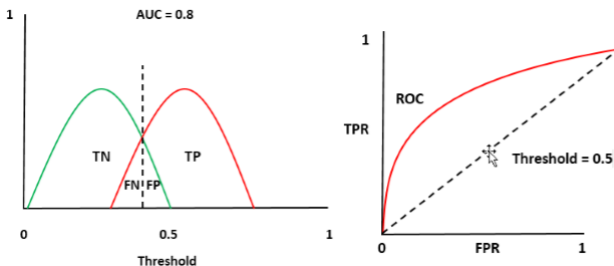
AUC - ROC curve

Threshold set to 0.5. There is no overlap between the two curves (green and red). This is the best model with AUC score of 1.0. This indicates that the probability of a model to separate positive and negative class is 1.0. In other words, we can say that there is 100% chance model can separate positive and negative class.



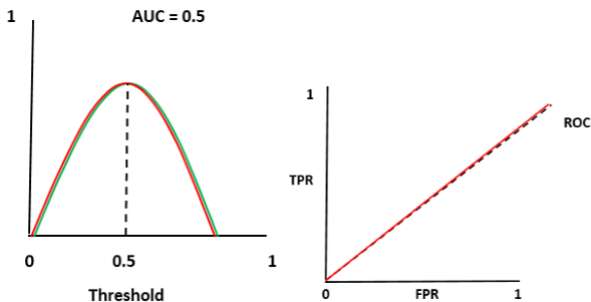
AUC - ROC curve

Threshold set to 0.5. There is a little bit of overlap between the two curves (green and red). This is a good model with AUC score of 0.8. This indicates that the probability of a model to separate positive and negative class is 0.8. In other words, we can say that there is 80% chance model can separate positive and negative class.



AUC - ROC curve

Threshold set to 0.5. We can see the full overlap between the two curves (green and red). This is a bad model with AUC score of 0.5. This indicates that the probability of a model to separate positive and negative class is 0.5. In other words, we can say that there is 50% chance model can separate positive and negative class.





Summary

- **Precision:** Precision measures what proportion of predicted positive label is actually positive. We mostly focus on false-positive value and try to decrease it to the least possible value thereby increase in precision value.
- **Recall:** Recall measures what proportion of actual positive label is correctly predicted as positive. We mostly focus on false-negative value and try to decrease it to the least possible value thereby increase in recall value.
- **F1-Score:** F1 Score is the 'Harmonic Mean' of precision and recall. We focus on both false-positive and false-negative and try to decrease both false-positive and false-negative thereby increase F1 Score.
- **AUC- ROC curve:** It is a graphical representation of ROC curve and region/area under curve i.e AUC. It is mostly used in binary class classification. It interprets the probability or percentage of separability of positive and negative classes. Higher the AUC - ROC, better is our model in separating positive and negative classes.

References and Credits

Bibliography

-  John C. Hull, *Machine Learning in Business: An Introduction to the World of Data Science*, Amazon, 2019.
-  Paul Wilmott, *Machine Learning: An Applied Mathematics Introduction*, Panda Ohana Publishing, 2019.