Machine Learning

Lecture 11: Evaluation Metrics for Classification

COURSE CODE: CSE451

2023

Course Teacher

Dr. Mrinal Kanti Baowaly

Associate Professor

Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh.

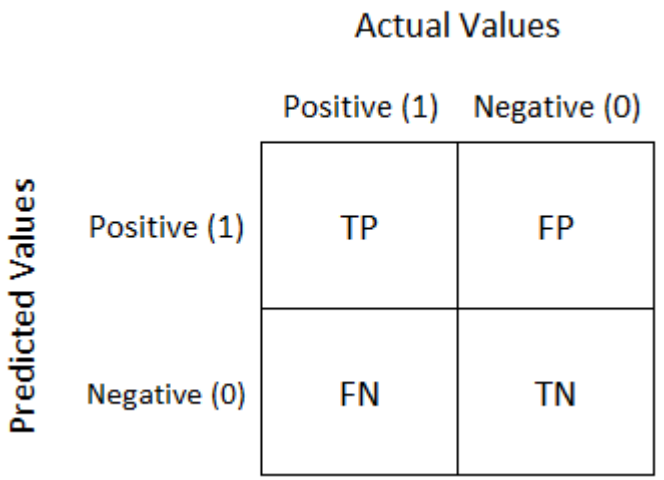
Email: mkbaowaly@gmail.com

Common Evaluation Metrics for Classification

1. Confusion Matrix
2. Accuracy
3. Precision
4. Recall/$Sensitivity$
5. Specificity
6. F1 Score
7. ROC (Receiver Operating Characteristics) Curve
8. AUC (Area Under the ROC curve) Score

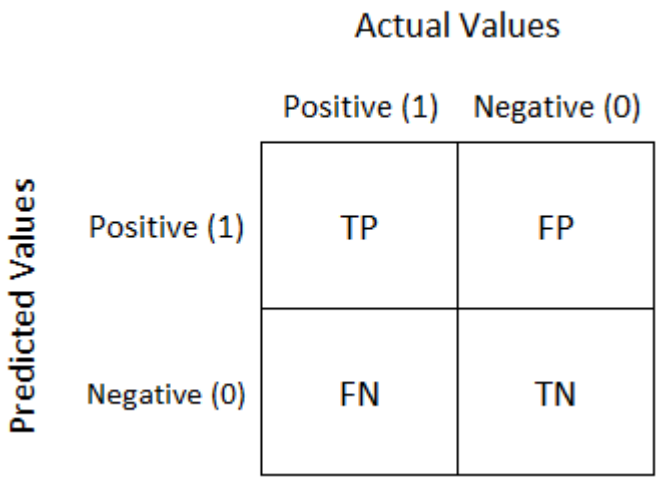
Confusion Matrix

- A confusion matrix is a table that describes the performance of a classification model on the test data
- It is an N X N matrix, where N is the number of classes being predicted
- Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (and vice versa).



Terms associated with Confusion matrix

- **True Positives :** The cases in which the model predicted 1(True) and the actual output was also 1(True).
- **True Negatives :** The cases in which the model predicted 0(False) and the actual output was also 0(False).
- **False Positives :** The cases in which the model predicted 1(True) and the actual output was 0(False).
- **False Negatives :** The cases in which the model predicted 0(False) and the actual output was 1(True).



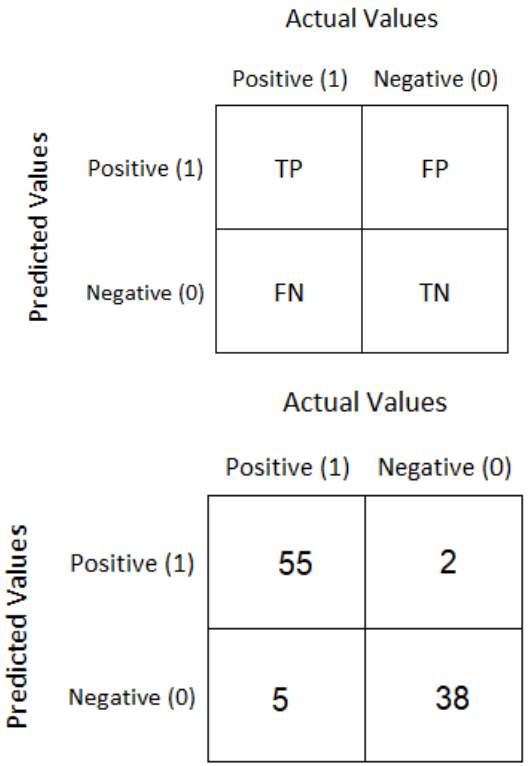
Accuracy

- It is the ratio of number of correct predictions to the total number of input samples (predictions).

$$Accuracy = \frac{No.\,of\ correct\ predictions}{Total\ no.\,of\ predictions}$$

$$= \frac{TP + TN}{TP + FP + FN + TN}$$

- It is the most commonly used metric to judge a model and is a good measure when the target variable classes in the data are nearly balanced.
- It should NEVER be used as a measure when the target classes are imbalanced.



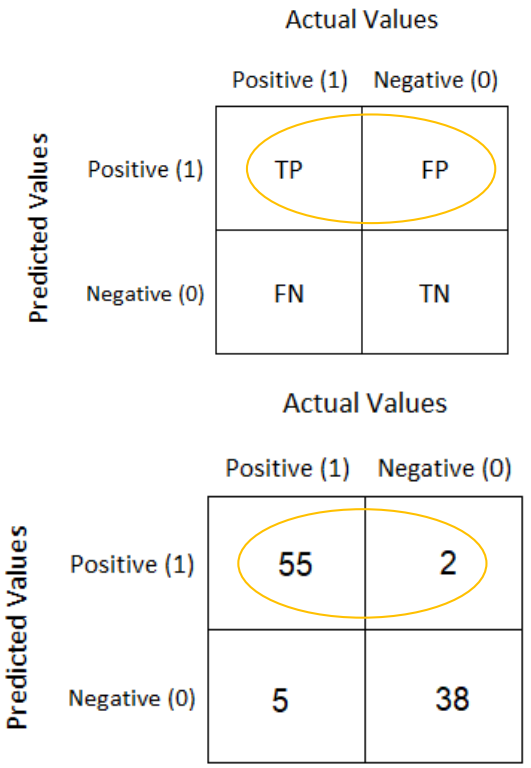
Accuracy = 93%
Error = 7%

Precision

- Out of all the positive classes we have predicted, how many are actually positive

$$Precison = \frac{TP}{TP + FP}$$

$$= \frac{55}{57} = 0.9649$$



Accuracy = 93%

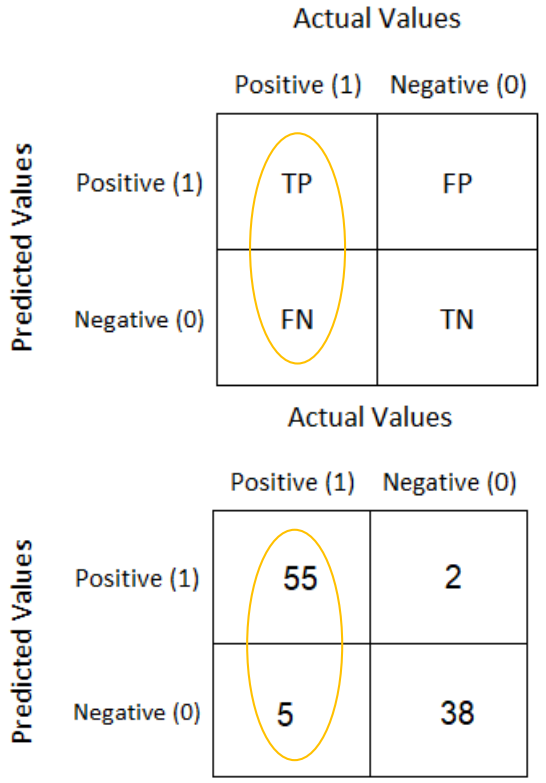
Error = 7%

Recall/Sensitivity

- Out of all the positive classes, how many are predicted correctly

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

$$= \frac{55}{60} = 0.9166$$



Accuracy = 93%

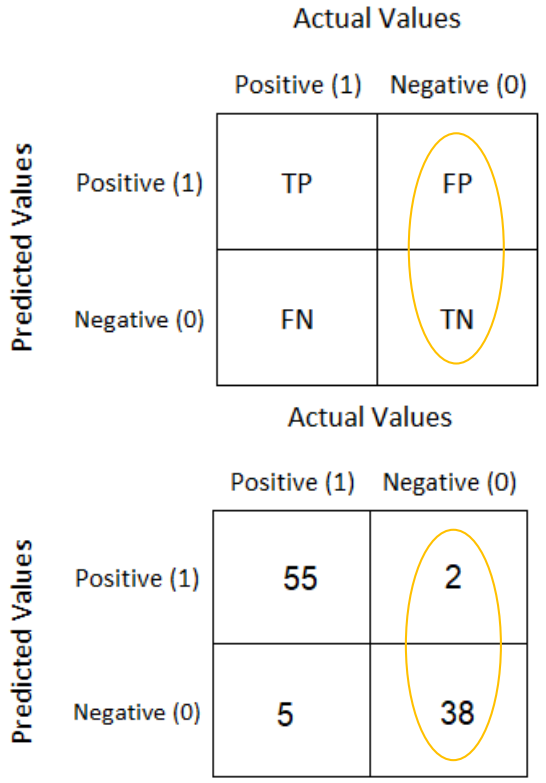
Error = 7%

Specificity

- Out of all the negative classes, how many are predicted correctly

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

$$= \frac{38}{40} = 0.95$$



Accuracy = 93%

Error = 7%

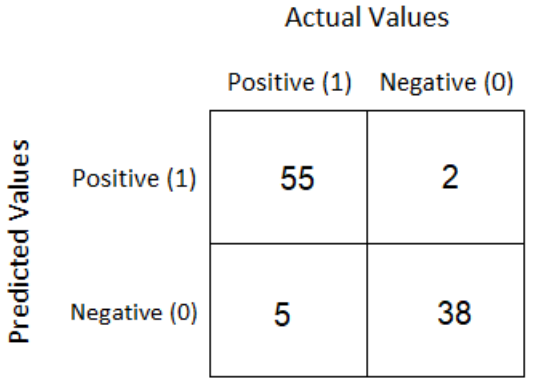
F1 Score

- Harmonic mean of the Precision and Recall

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$$

F1 = 0.94

- It makes a balance between Precision and Recall
- Rather than measure recall and precision every time, it would be easier to use a single F1 score
- It is a better choice when the target classes are imbalanced



Accuracy = 93%
Error = 7%

HW: Why classification accuracy is not enough?

Hints:

- Suppose you have the problem of detecting cancer. You have two classes for that:
  1. Having cancer, the positive class, denoted by 1
  2. No cancer, the negative class, denoted by 0

Lets assume that you have 1000 patient records. The confusion matrix of a predictive model is as in the right side.

It yields very high accuracy (99.4%) but fails to detect the patients with cancer (4 out of 5). F1 score can be a proper metric in this case of imbalanced target classes.

		Actual Values: Positive (1)	Actual Values: Negative (0)
Predicted Values	Positive (1)	1	2
	Negative (0)	4	993

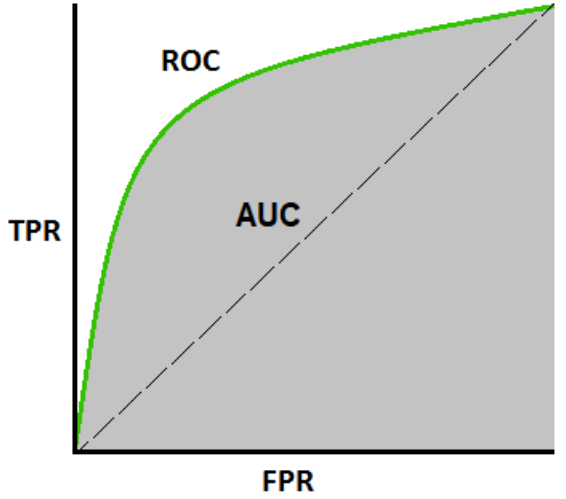
Accuracy = 0.994
Error = 0.006
F1 Score = 0.249

ROC (Receiver Operating Characteristics) Curve

- A ROC is a graphical plot that is used as a performance measurement for classification problem
- The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings

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

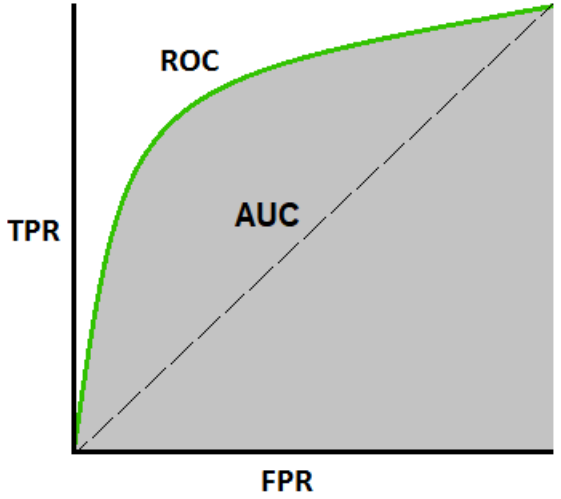
$$FPR = 1 - Specificity = 1 - \frac{TN}{FP+TN} = \frac{FP}{FP+TN}$$



- It tells how much model is capable of distinguishing between classes (i.e. Separability/Discrimination capacity)

AUC (Area Under the ROC curve) Score

- The AUC is the area under the ROC curve.
- This score gives us a good idea of how well the model performances.
- AUC Score ranges 0 to 1
- An ideal model has AUC near to the 1 which means it has excellent discrimination capacity.
- An poor model has AUC near to the 0.5 which means it has no discrimination capacity.
- When AUC is approximately 0, model is actually reciprocating the classes. It means the model is predicting negative class as a positive class and vice versa (Worst model).



Example: Confusion Matrix

```
# import confusion matrix
from sklearn.metrics import confusion_matrix
# actual values
actual = [1,0,0,1,0,0,1,0,1,1]
# predicted values
predicted = [1,0,0,1,0,0,1,1,0,0]
# confusion matrix
matrix = confusion_matrix(actual, predicted, labels=[1,0])
print('Confusion matrix : \n',matrix)
# outcome values order in sklearn
TP,FN,FP,TN = matrix.reshape(-1)
print('Outcome values : \n', TP,FN,FP,TN)
```

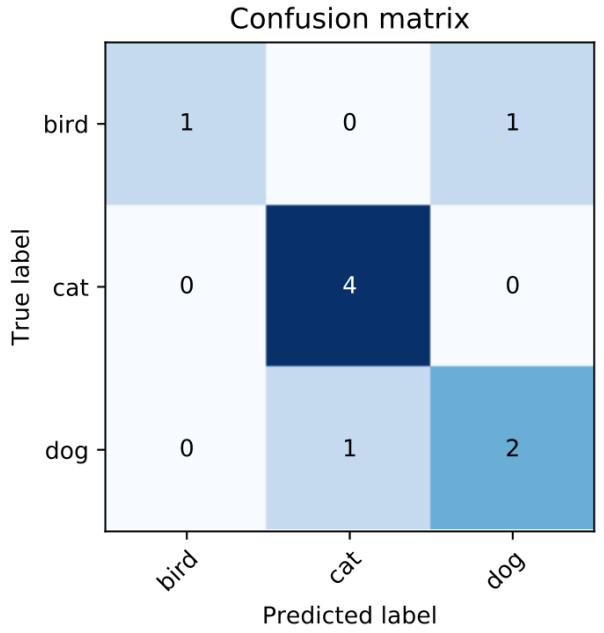
Assignment: How to Use Various Metrics in Classification Problems?

1. Let us investigate the Lung Cancer Dataset from here: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link
2. There are 1000 items (patients) and 24 predictor variables (age, gender, air pollution exposure, alcohol use, dust allergy, etc.) without index and ID. The variable (level) to predict the risk of lung cancer is encoded as 0 and 1 where 0 means low risk of lung cancer and 1 means medium or high risk of lung cancer.
3. Build a binary classification model to predict the risk of lung cancer (0, 1) of the patients. Estimate and compare Accuracy, Precision, Recall, Specificity, F1 Score and AUC Score to evaluate the performance of the model. And plot the ROC curve also.

Evaluation Metrics for Multi-class Classification

- **Micro-averaged Precision** is calculated as precision of total values.
- **Macro-averaged Precision** is calculated as an average of Precisions of all classes.
- **Weighted-averaged Precision** is also calculated based on Precision per class but takes into account the number of samples of each class in the data
- **HW:** Find which type of averaging is preferable?

Source link: Maria Khalusova



Some Learning Materials

AnalyticsVidhya: How to Choose Evaluation Metrics for Classification Models

RitchieNg: Evaluating a Classification Model

TowardsDatascience: Various ways to evaluate a machine learning model's performance

Understanding Micro, Macro, and Weighted Averages for Scikit-Learn metrics in multi-class classification with example