Classification Evaluation Metrics

TIET, PATIALA

Evaluation Metrics-Classification

- The performance of the classification model is evaluated with a number of evaluation metrics. Some of the most commonly used are:

1. Confusion Metrics
2. Accuracy
3. Misclassification Rate
4. Precision
5. Recall/ True Positive Rate/ Sensitivity/ Hit Rate
6. F-β Score
7. Specificity
8. ROC Curve

Confusion Matrix

- A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known.

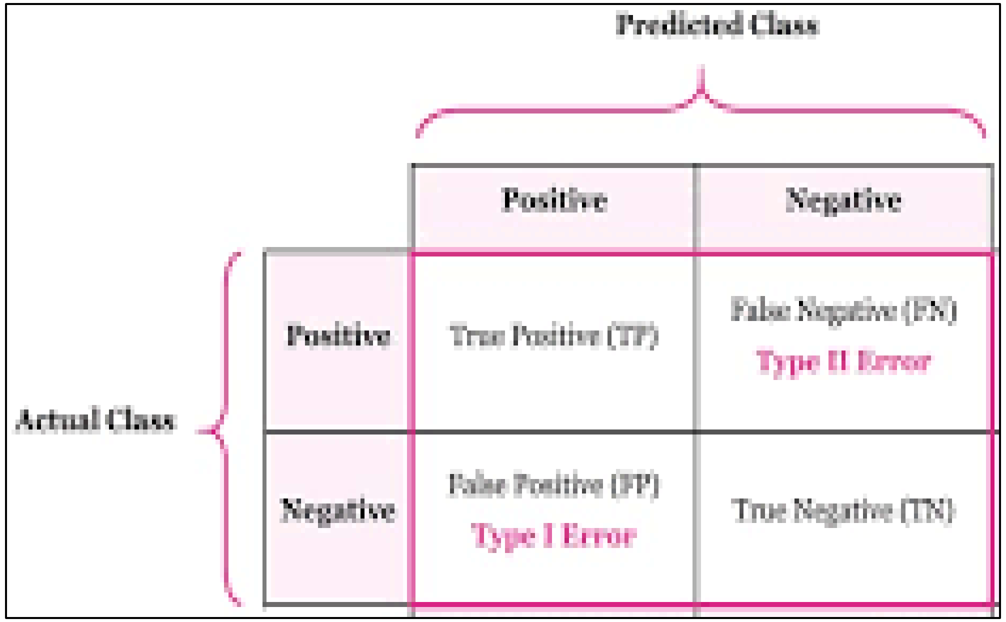
- It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions

- Each combination of dimension and class is a variable in the contingency table.

- For instance, for a binary classification models, both dimensions have two classes 0 and 1 and hence there are *four* variables namely *True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).*

Confusion Matrix Contd….

- **True positives (TP):** These are cases in which we predicted yes and the actually value is also yes.
- **True negatives (TN):** These are cases in which we predicted no, and the actual value is also no.
- **False positives (FP):** These are cases in which we predicted yes, but the actual value is no. (Also known as a "**Type I error**")
- **False negatives (FN):** These are cases in which we predicted no, but the actual value is yes. (Also known as a "Type II error.")



		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error
	Negative	False Positive (FP) Type I Error	True Negative (TN)

Which error (Type I/Type II) is important?

- Though, we want to minimize both False positive and False Negative errors. But, sometimes it is not possible to minimize both.
- So, it depends upon the application that which error must be minimized.
- For example, consider an application of classifying a patient as cancer patient (positive) or non-cancer patient (negative). In this application, diagnosing a cancer patient as healthy (False negative) is more important as compared to diagnosing a healthy patient as cancer (False Positive). So, False Negative must be minimized.
- Consider another application of classifying emails as spam (positive) or ham (negative). In this application, classifying a ham email as spam (False Positive) is more important than spam as ham (False Negative). So, False Positive must be minimized.

Accuracy

- Accuracy of a classification model is defined as number of correct predictions made by the model to the total number of predictions.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$$

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

- Accuracy is an important metric to be used when the data is approximately balanced (i.e., when the number of examples of each output class are approximately same). But, **it should not be used for imbalanced datasets.**
- For instance, in the cancer patient diagnosis system, if we have 100 patients (95 non-cancer, 5 cancer) and our model predicts every patient as non- cancer patient. So, for such a poor model the accuracy is 95%.

Misclassification Rate

- Misclassification Rate of a classification model is defined as number of incorrect predictions made by the model to the total number of predictions.

$$Accuracy = \frac{Incorrect\ Predictions}{Total\ Predictions}$$

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

- Lower the misclassification rate, better is the model.

Precision

- Precision of a classification model is defined as the number of correct positive predictions made by the model to the total positive predictions made by the model.

$$Precision = \frac{Correct\ Positive\ Predictions}{Total\ Positive\ Predictions}$$

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

- Precision is high when False Positive is low.
- Therefore, precision must be used when we want to minimize False Positive error (such as spam classification applications).

Recall

- Recall of a classification model is defined as the number of correct positive predictions made by the model to the total correct positive predictions for the model.

$$Recall = \frac{Correct\ Positive\ Predictions}{Total\ Correct\ Positive\ Predictions}$$

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

- Recall is high when False Negative is low.
- Therefore, recall must be used when we want to minimize False Negative error (such as medical diagnosis applications).
- **Recall is also called sensitivity/ hit rate/ True Positive Rate**

F-β Score

- It is a weighted metric that considers both Precision and Recall.
- It is a weighted harmonic mean of Precision and Recall.

$$F - \beta\ Score = (1 + \beta^2)\frac{Precision * Recall}{\beta^2 * Precision + Recall}$$

- If we want to give recall higher weightage over precision, then we take β > 1
- In case precision has to be given more weightage over recall, then we take β <1.
- Generally, we take value β =0.5 (for giving importance to precision over recall) and β = 2 (for giving importance to recall over precision)

F1-Score

- If we give equal weightage to Precision and Recall in F-β score, then we take β=1 and it is called as F1-score.

$$F1 - Socre = \frac{2 * Precision * Recall}{Precision + Recall}$$

- It is thus, the unweighted harmonic mean of Precision and Recall.

Macro & Weighted Precision

- The precision computed w.r.t positives (as discussed in slide 8) is called micro precision.
- But in many systems (especially for multi-class problems), precision is computed w.r.t each class and then macro (unweighted mean) , weighted (weighted mean) precision is computed.

$$Precision_{positive} = \frac{Correct\ positive\ prediciton}{Total\ positive\ prediction} = \frac{TP}{TP + FP}$$

$$Precision_{negative} = \frac{Correct\ negative\ prediciton}{Total\ negative\ prediction} = \frac{TN}{TN + FN}$$

$$Macro\ Precision = \frac{Precision_{positive} + Precision_{negative}}{2}$$

$$Weighted\ Precision = \frac{n_1 \times Precision_{positive} + n_2 \times Precision_{negative}}{n_1 + n_2}$$

Where n_1 is the number of support examples for positive class and n_2 is the number of support examples for negative class.

Macro & Weighted Recall

- The recall computed w.r.t positives (as discussed in slide 9) is called micro recall.
- But in many systems (especially for multi-class problems), recall is computed w.r.t each class and then macro (unweighted mean) , weighted (weighted mean) recall is computed.

$$Recall_{positive} = \frac{Correct\ positive\ prediciton}{Total\ correct\ positive\ prediction} = \frac{TP}{TP + FN}$$

$$Recall_{negative} = \frac{Correct\ negative\ prediciton}{Total\ correct\ negative\ prediction} = \frac{TN}{TN + FP}$$

$$Macro\ Recall = \frac{Recall_{positive} + Recall_{negative}}{2}$$

$$Weighted\ Recall = \frac{n_1 \times Recall_{positive} + n_2 \times Recall_{negative}}{n_1 + n_2}$$

Where n_1 is the number of support examples for positive class and n_2 is the number of support examples for negative class.

Macro & Weighted F1-Score

$$F1\ score\_positive = \frac{2 * Precsion_{positive} * Recall_{positve}}{Precsion_{positive} + Recall_{positve}}$$

$$F1\ score\_negative = \frac{2 * Precsion_{negative} * Recall_{negative}}{Precsion_{negative} + Recall_{negative}}$$

$$Macro\ F1\ score = \frac{F1\ score\_positive + F1\ score\_negative}{2}$$

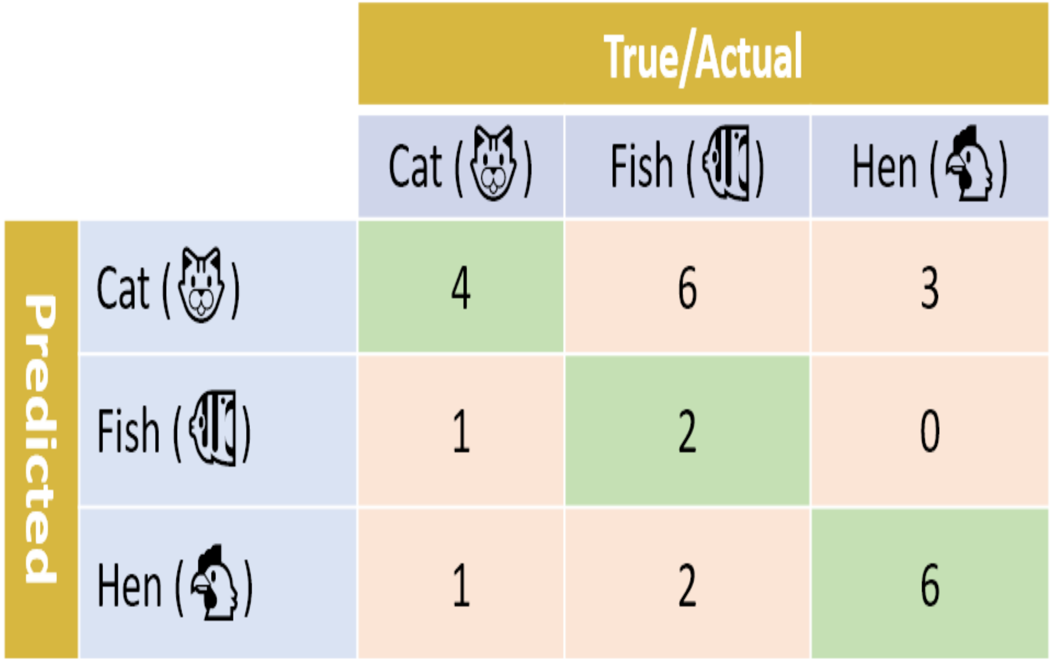
$$Weighted\ Precision = \frac{n_1 \times F1\ score\_positive + n_2 \times F1\ score\_negative}{n_1 + n_2}$$

Where n_1 is the number of support examples for positive class and n_2 is the number of support examples for negative class.

Numerical Example

For the confusion matrix, compute

1. Accuracy
2. Precision for each class, Macro Precision and Weighted Precision
3. Recall for each class, Macro Recall and Weighted Recall



		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

Numerical Example-Solution

$$Accuracy = \frac{Correct\ Prediciton}{Total\ Predicitons}$$

$$= \frac{4+2+6}{4+6+3+1+2+0+1+2+6} = \frac{12}{25} = 48\%$$

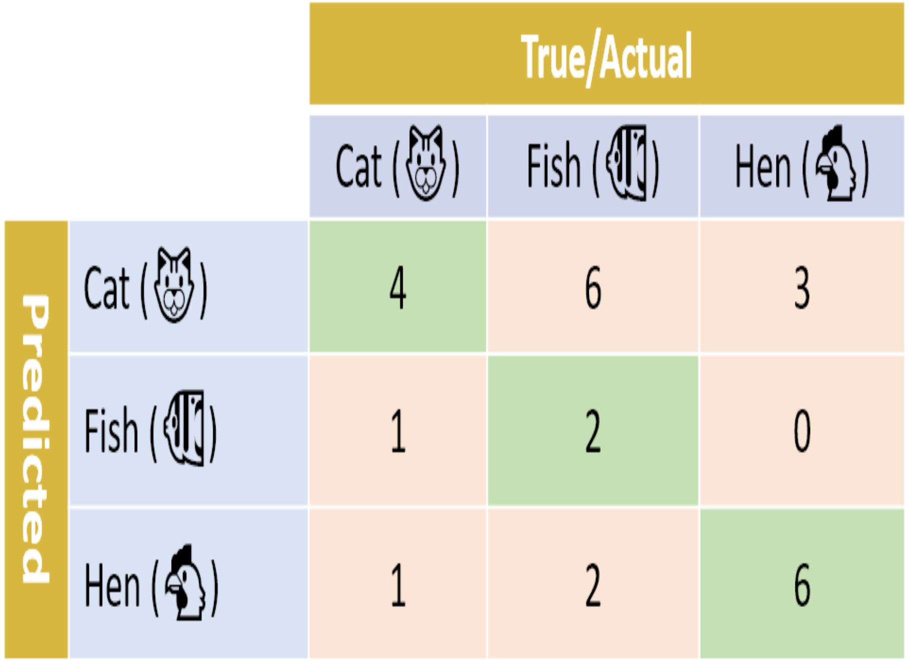
$$Precison\ cat = \frac{Correct\ cat\ pred}{Total\ cat\ pred} = \frac{4}{4+6+3} = \frac{4}{13} = 0.31$$

$$Precison\ fish = \frac{Correct\ fish\ pred}{Total\ fish\ pred} = \frac{2}{1+2+0} = \frac{2}{3} = 0.66$$

$$Precison\ hen = \frac{Correct\ hen\ pred}{Total\ hen\ pred} = \frac{6}{1+2+6} = \frac{6}{9} = 0.66$$

$$Macro\ Precision = \frac{0.31+0.66+0.66}{3} = 0.54$$

$$Weighted\ Precision = \frac{6*0.31+10*0.66+9*0.66}{25} = \frac{14.4}{25} = 0.58$$



		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

Numerical Example-Solution

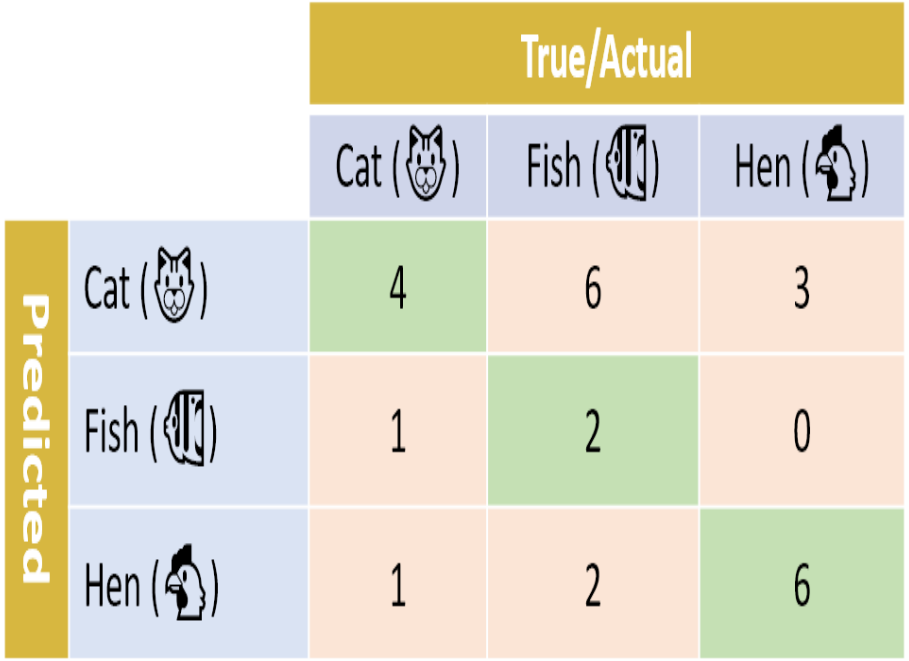
$$Recall\ cat = \frac{Correct\ cat\ pred}{Total\ correct\ cat\ pred} = \frac{4}{4+1+1} = \frac{4}{6} = 0.66$$

$$Recall\ fish = \frac{Correct\ fish\ pred}{Total\ correct\ fish\ pred} = \frac{2}{6+2+2} = \frac{2}{10} = 0.2$$

$$Recall\ hen = \frac{Correct\ hen\ pred}{Total\ correct\ hen\ pred} = \frac{6}{3+0+6} = \frac{6}{9} = 0.66$$

$$Macro\ Recall = \frac{0.66+0.2+0.66}{3} = 0.51$$

$$Weighted\ Recall = \frac{6*0.66+10*0.2+9*0.66}{25} = \frac{11.9}{25} = 0.48$$



		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

Sensitivity and Specificity

- The terms "sensitivity" and "specificity" were introduced by American biostatistician Jacob. It is widely used in medical diagnosis system.
- Sensitivity (as discussed earlier) is the true positive rate or recall.
- **Specificity is True Negative rate** that measures the proportion of negatives that are correctly identified.
- In medical diagnosis systems, both are important. As sensitivity tells that how many sick patients are correctly identified as sick and specificity tells that how many healthy patients are identified as healthy.
- Thus, if a test's sensitivity is 97% and its specificity is 92%, its rate of false negatives is 3% and its rate of false positives is 8%.

ROC Curve

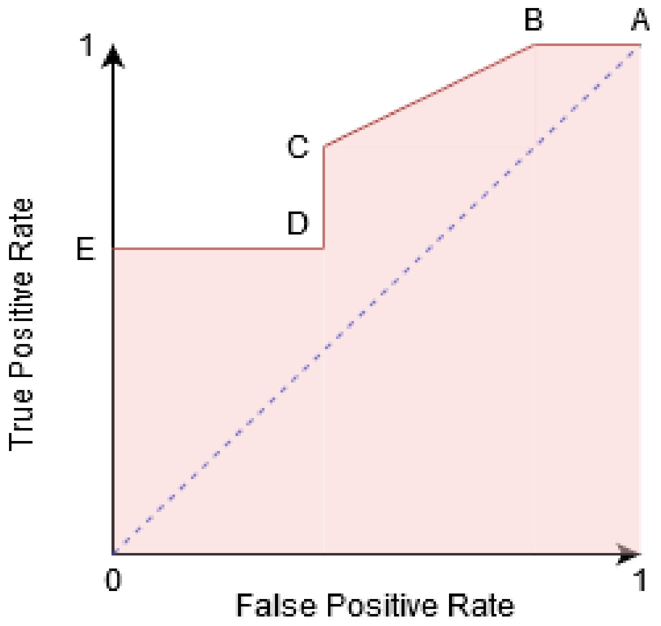
- The **Receiver Operator Characteristic (ROC)** curve is an evaluation metric for binary classification problems.
- . It is a probability curve that plots the **TPR** against **FPR (FPR=1-Specificity)** at various threshold values.
- The **Area Under the Curve (AUC)** is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.
- When AUC = 1, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly.
- If, however, the AUC had been 0, then the classifier would be predicting all Negatives as Positives, and all Positives as Negatives.

ROC Curve (Contd...)

- When 0.5<AUC<1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than False negatives and False positives.

ROC Curve (Contd...)

- In a ROC curve, a higher X-axis value indicates a higher number of False positives than True negatives. While a higher Y-axis value indicates a higher number of True positives than False negatives.
- So, the choice of the threshold depends on the ability to balance between False positives and False negatives.
- Point A is where the Sensitivity is the highest and Specificity the lowest. This means all the Positive class points are classified correctly and all the Negative class points are classified incorrectly.

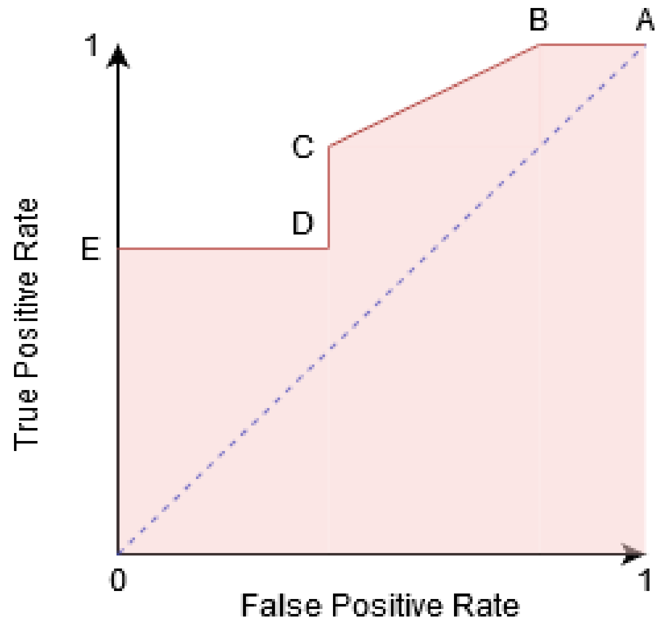


ROC Curve (Contd…)

- Although Point B has the same Sensitivity as Point A, it has a higher Specificity. Meaning the number of incorrectly Negative class points is lower compared to the previous threshold. This indicates that this threshold is better than the previous one.

- Between points C and D, the Sensitivity at point C is higher than point D for the same Specificity. This means, for the same number of incorrectly classified Negative class points, the classifier predicted a higher number of Positive class points.

- Point E is where the Specificity becomes highest. Meaning there are no False Positives classified by the model. The model can correctly classify all the Negative class points!


B
A
C
D
E
True Positive Rate
False Positive Rate