

# Cheat Sheet – Imbalanced Data in Classification



## Accuracy doesn't always give the correct insight about your trained model

**Accuracy:** %age correct prediction

**Precision:** Exactness of model

**Recall:** Completeness of model

**F1 Score:** Combines Precision/Recall

Correct prediction over total predictions  
From the detected cats, how many were actually cats

Correctly detected cats over total cats

Harmonic mean of Precision and Recall

One value for entire network  
Each class/label has a value

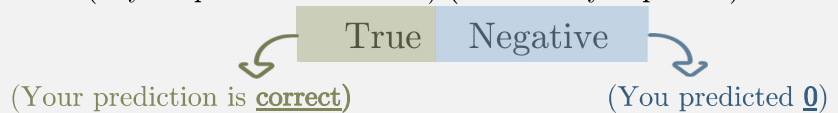
Each class/label has a value

Each class/label has a value

## Performance metrics associated with Class 1

|                  |   | Actual Labels  |                |
|------------------|---|----------------|----------------|
|                  |   | 1              | 0              |
| Predicted Labels | 1 | True Positive  | False Positive |
|                  | 0 | False Negative | True Negative  |

(Is your prediction correct?) (What did you predict)



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

$$\text{F1 score} = 2 \times \frac{(\text{Prec} \times \text{Rec})}{(\text{Prec} + \text{Rec})}$$

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

$$\text{False +ve rate} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

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

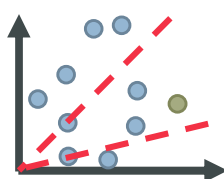
$$\text{Recall, Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

## Possible solutions

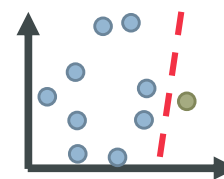
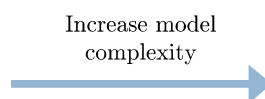
- Data Replication:** Replicate the available data until the number of samples are comparable
- Synthetic Data:** Images: Rotate, dilate, crop, add noise to existing input images and create new data
- Modified Loss:** Modify the loss to reflect greater error when misclassifying smaller sample set
- Change the algorithm:** Increase the model/algorithm complexity so that the two classes are perfectly separable (Con: Overfitting)



$$\text{loss} = a * \text{loss}_{\text{green}} + b * \text{loss}_{\text{blue}} \quad a > b$$



No straight line ( $y=ax$ ) passing through origin can perfectly separate data. **Best solution:** line  $y=0$ , predict all labels blue



Straight line ( $y=ax+b$ ) can perfectly separate data. Green class will no longer be predicted as blue