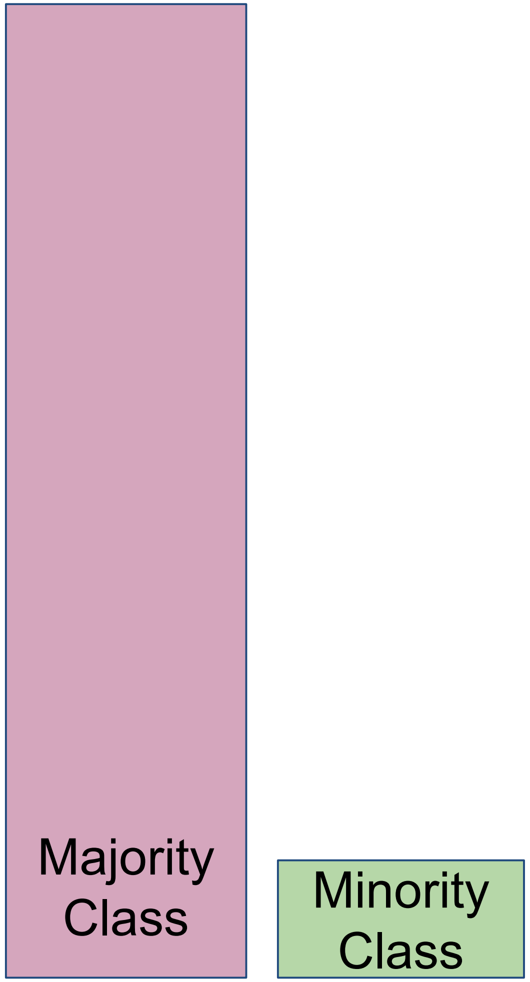
Imbalanced classification

Part 1: Introduction

By: Noureddin Sadawi, PhD
University of London

Class imbalance

- Class imbalance refers to the problem when a classification dataset contains more than one class and the number of instances in each class is **not approximately the same.**
- Imbalanced classification happens when the class distribution in the training dataset is unequal (skewed class distribution).
  - Imbalance in the class distribution might be negligible.
  - A severe imbalance is more challenging to model and may require specialised techniques.



Class imbalance is a challenge

- The main problem is the skewed class distribution.
- This is often exemplified by a binary (two-class) classification task where most of the examples belong to class 0 with only a few examples in class 1.
- Because the class distribution is not balanced, most machine learning algorithms will perform poorly and require modification to avoid simply predicting the majority class in all cases.
- Alternate methods for evaluating predictions on imbalanced examples are required.
- **Usually class 1 is the class of interest and it is the minority class.**

Dominance of majority class

- A model trained on heavily imbalanced data can easily predict everything as majority class.
- Misclassifying an example from the majority class as belonging to the minority class (called a false-positive) is often not desired, but less critical than classifying an example from the minority class as belonging to the majority class, a **false negative**.
- For example, in an AIDS prediction system, we may be far more concerned with having a low number of false negatives than a low number of false positives.
  - A false negative would mean not treating a patient when in fact they need treatment.
  - A false positive means the person would receive treatment (or another check) when they didn’t need to.

Model evaluation 1/2

- An evaluation metric must be carefully selected.
- It should best capture what is **important about the model or predictions**
- Accuracy: the number of correct predictions divided by the total number of predictions
  - Accuracy = Correct Predictions / Total Predictions.
  - Value is usually [0,1] or as a percentage [0-100].
  - A model that predicts everything as the majority class would give a high accuracy on an imbalanced test dataset (which is misleading).
- Error Rate = Incorrect Predictions / Total Predictions.
- Accuracy = 1 – Error Rate Error Rate = 1 – Accuracy.

The confusion matrix

Predicted class

Actual Class

	Positive	**Negative**	
Positive	**True Positive (TP)**	**False Negative (FN)** **Type II Error**	Sensitivity (i.e. Recall) $\frac{TP}{TP+FN}$
Negative	**False Positive (FP)** **Type I Error**	**True Negative (TN)**	Specificity $\frac{TN}{TN+FP}$
	Precision $\frac{TP}{TP+FP}$	Negative Predicted Value $\frac{TN}{TN+FN}$	Accuracy $\frac{TP+TN}{TP+FP+TN+FN}$

Model evaluation 2/2

- F-Measure: the harmonic mean of Precision and Recall
  - Value is between [0,1] the closer to 1 the better the model:

$$F = 2 * \frac{Precision \; * Recall}{Precision + Recall}$$

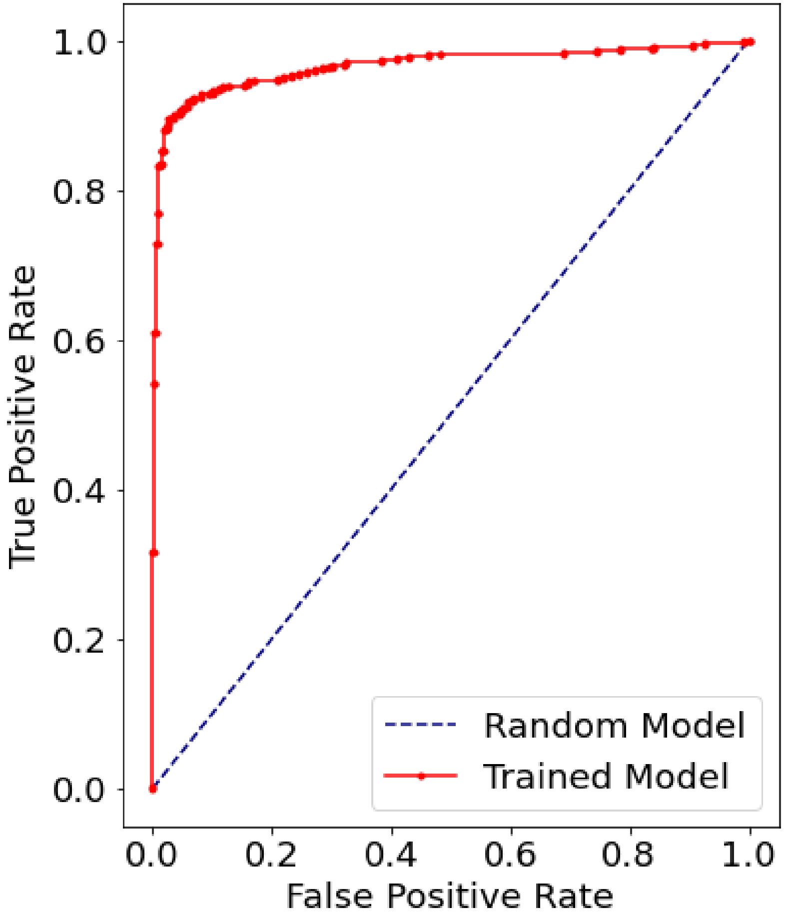
- Area under the curve:
  - The receiver operating characteristic curve, or ROC curve.
  - The precision-recall curve, or PRC curve.
  - Both areas are usually between [0,1] the closer to 1 the better the model.

ROC and PRC curves

- Instead of predicting a class value directly, it is useful to predict a class probability.
- This is to provide the capability to choose and even calibrate the threshold for how to interpret the predicted probabilities.
- For example, a default might be to use a threshold of 0.5, meaning that a probability in [0.0, 0.49] is a negative outcome (0) and a probability in [0.5, 1.0] is a positive outcome (1).
- This threshold can be adjusted to tune the behaviour of the model for a specific problem.
  - An example would be to reduce one or another type of error.

ROC curve

- It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0.
    - It plots the false alarm rate versus the hit rate.
- True Positive Rate (Sensitivity) = True Positives / (True Positives + False Negatives).
- Specificity = True Negatives / (True Negatives + False Positives).
- False Positive Rate (1 - Specificity) = False Positives / (False Positives + True Negatives).



PRC curve

- Positive Predictive Power (Precision) = True Positives / (True Positives + False Positives)
- Recall (Sensitivity) = True Positives / (True Positives + False Negatives)
- Because of the high class imbalance, we are less interested in the skill of the model at predicting class 0 correctly, e.g. high true negatives
- Key to the calculation of precision and recall is that the calculations **do not make use of the true negatives**
- It is only concerned with the correct prediction of the minority class, class 1