# Introduction to Performance Metrics

1. Metrics

A diagram of a company

Description automatically generated with medium confidence

1. Multiclass

A diagram of a company

Description automatically generated with medium confidence

# Accuracy

1. Overview

* Percentage of correct predictions
* Binary classification
* For imbalanced datasets, accuracy is not an appropriate metric since it does not distinguish between the numbers of correctly classified examples of different classes
* The minority class has very little impact on the accuracy as compared to that of the majority class

1. Example

A screenshot of a computer

Description automatically generated

# Precision, Recall and F-measure

1. Precision and recall

* **TPR (recall or sensitivity) – from the total positive observations, how many were correctly identified by the model?** -> Minority class
* Increase in recall decreases the probability of misclassifying a sample from minority class
* **TNR (specificity) -> Majority class**
* **Positive predictive value (precision) – how many observations among the observations that the model predicted to be positive are actually positive?** -> Minority class
* Increase in precision is reduction of negative positive
* Negative predictive value -> Majority class
* Bot precision and recall vary between 0 and 1
* To select and tune ML models, our goal is to maximize both precision and recall
* Both precision and recall **depend on a probability threshold**

1. F-measure

* F1 score is a **weighed harmonic mean** of the precision and recall
* Vary between 0 and 1
* Optimizing this metric produces the best balance between precision and recall

1. Support

* Support is the number of actual occurrences of the class in the specified dataset
* Indicates structural weaknesses in the reported scores. Highlights imbalanced datasets
* Doesn’t change between models but instead diagnoses the evaluation process

1. Example

* Recall

A diagram of a mathematical equation

Description automatically generated with medium confidence

* Precision

A screenshot of a computer

Description automatically generated

* F1 score

A math equations and formulas

Description automatically generated with medium confidence

* Support

A screenshot of a graph

Description automatically generated

* Breakdown

A screenshot of a computer

Description automatically generated

1. Optimizing the threshold

* The optimal threshold is that at which F-measure is highest (if we want to balance precision and recall

A graph of a graph showing the different colored lines

Description automatically generated with medium confidence

# Confusion tables, FPR, FNR

1. Confusion matrix

* In a 2-class problem, the results of the correctly and incorrectly classified samples are recorded in a confusion matrix

A grid of black and white squares with black text

Description automatically generated

1. FPR and FNR

* False positive rate:
* How many in the majority class were wrongly classified
* False negative rate:
* How many in the minority class were wrongly classified
* Both vary between 0 and 1
* Goal: minimize both FPR and FNR
* Depend on a probability threshold
* Minimize FNR:
* Disease diagnosis – we want to minimize the number of sick people that we do not diagnose correctly
* Minimize FPR:
* Drug discovery – we want to minimize the number of drugs that we think could be beneficial, but they are not

1. Example

* FPR

A screen shot of a computer

Description automatically generated

* FNR:

A screenshot of a computer

Description automatically generated

1. FPR, FNR, vs. threshold

A graph with a line and a line

Description automatically generated with medium confidence

# Balanced Accuracy

1. Accuracy
2. Recall

* The percentage of observations of each class that were correctly classified
* Accuracy within each class

1. Balanced accuracy

* Average recall obtained in each class

A math equation with arrows

Description automatically generated with medium confidence

# Geometric mean, Dominance, Index of Imbalanced Accuracy

1. Geometric mean

* The Geometric mean tries to maximize the accuracy on each of the classes while keeping these accuracies balanced
* Range from 0 to 1

1. Dominance

* Range from -1 to 1
* A value of 1 indicates perfect accuracy on the minority (positive) class, but all cases of the majority class are misclassified
* A value of -1 corresponds to the opposite situation

1. Index of imbalanced accuracy – IBA

* Quantifies a trade-off between an index of how balanced both class accuracies are and a chosen unbiased measure

Where (1 + dominance) is the weighting factor and M represents any performance metric

# ROC-AUC

1. Receiving Operating Characteristic (ROC)

* The ROC curve plots the benefits (TPR) and costs (FPR) at different classification thresholds
* Every point on the ROC curve represents a probability threshold and the model performance trade-off
* Evaluates how well a classifier can separate positive and negative examples
* Helps identify the best threshold to separate them

A diagram of a path

Description automatically generated

1. Area under the ROC: AUC

* AUC is the area under the ROC curve
* AUC provides an aggregate measure of performance across all possible classification thresholds
* Higher AUC indicates the model is better at predicting both classes

A graph with lines and numbers

Description automatically generated

A graph with arrows pointing to the top

Description automatically generated with medium confidence

1. Known issues of the ROC

* ROC curves of different classifiers cross each other
* The performance of the classifiers is not comparable for all discriminant thresholds

A graph with a point and a point

Description automatically generated with medium confidence

# Precision-Recall Curve

1. Precision and recall
2. Precision-recall curve

* Shows the relationship between precision and recall for every cut-off / discriminant probability threshold
* A graph with
* Recall in x-axis
* Precision in y-axis
* Every point on the PRC represents a chosen cutoff
* Every point provides the precision and recall for a certain cutoff/threshold

A graph of a normal distribution

Description automatically generated with medium confidence

A graph with text overlay

Description automatically generated

1. PRC: perfect model

A screenshot of a graph

Description automatically generated

1. PRC: Random model

A graph of a graph with a line and a line

Description automatically generated with medium confidence

* The area under the PRC provides an aggregate measure of performance across all possible classification thresholds
* Larger area indicates better model performance

# Probability

1. Balanced dataset

* The likelihood of receiving on observation of class 1 is 0.5
* If the model outputs a prob > 0.5 -> class 1
* If the model outputs a prob < 0.5 -> class 0

1. Imbalanced dataset

* The likelihood of receiving an observation of class 1 is 0.1
* If the model outputs a prob > 0.1 -> class 1
* If the model outputs a prob < 0.1 -> class 0
* When we train models on imbalanced datasets, we tend to obtain lower probability values for the rare class

1. Fine tune the probability threshold

* Using 0.5 as a default threshold does not make sense when we have imbalanced datasets
* We need to fine tune the threshold depending on what we want to optimize
* Precision, recall, or F-score
* False positive or false negative discovery rate

A graph of a patient

Description automatically generated with medium confidence