# Imbalanced datasets – Introduction

1. imbalanced datasets

* Imbalanced datasets have many more instances of certain classes than of others
* Most ML algorithms assume balanced distributions
* As the minority examples occur rarely, rules to predict the small classes are difficult to find
* Samples from the minority classes are most often misclassified
* The problem: Particularly interested in the minority classes
* E.g. Fraud detection, medical diagnosis

1. Imbalanced class distribution

* Class distribution: the proportion of instances belonging to each class
* Imbalanced datasets can have 1 or more minority classes
* Imbalance degree: ratio of the sample size of the minority class to that of the majority class
* Typical imbalanced ratios are 1:10 and smaller

1. Application domains

* In certain applications, the correct classification of samples in the minority classes often has a greater value than the contrary case
* Fraud detection
* Medical diagnosis
* Equipment manufacturing and testing
* Detection of oil spills from radar images of the ocean
* Network intrusion detection

# Nature of the imbalanced class

1. Imbalanced class distribution

* Class distribution
* Imbalance ratio
* Factors that influence the ability of a classifier to identify rare events
* Small sample size
* Class separability
* Within-class sub-clusters

1. Small sample size

* Sample size plays a crucial role in determining the ‘goodness’ of a model
* If a sample size is limited, finding patterns inherent to the small class is hard
* As the data size increases, the error in the prediction decreases
* Imbalanced classes may not be a problem if the data is big enough

1. Class separability

* If a pattern among classes overlap, it is harder to find rules
* The class imbalance per se may not be a problem, instead the separability makes it harder to find rules to classify correctly the minority class
* Linearly separable domains are not sensitive to any amount if imbalance

1. Within class sub-clusters

* In many classification problems, a single class is composed of various sub-clusters or concepts
* These sub-clusters do not always contain the same number of examples
* This phenomenon is referred to as within-class imbalance, corresponding to the imbalanced class distribution among subclasses
* Within class sub-clusters increases the complexity and makes it harder to find boundaries to separate the classes

# Solutions for imbalanced datasets

1. Solutions for imbalanced datasets

* Data-level:
* Undersampling
* Oversampling
* Cost-sensitive
* Higher misclassification costs
* Ensemble algorithms
* Boosting and bagging
* With sampling

1. Data-level approaches

* Changing the distribution of the data
* Random over- or under-sampling
* Creating new synthetic data
* Removing noise or alternatively, removing easy observations to classify

1. Cost-sensitive approaches

* Different cost to different errors
* The cost of misclassifying an instance of the minority class outweighs the cost of misclassifying an instance from the majority
* The cost-sensitive learning process seeks to minimize the cost error

1. Ensemble approaches

* Combine weak learners
* Construct multiple classifiers from the original data and then aggregate the predictions
* Combining classifiers generally improves their generalization ability