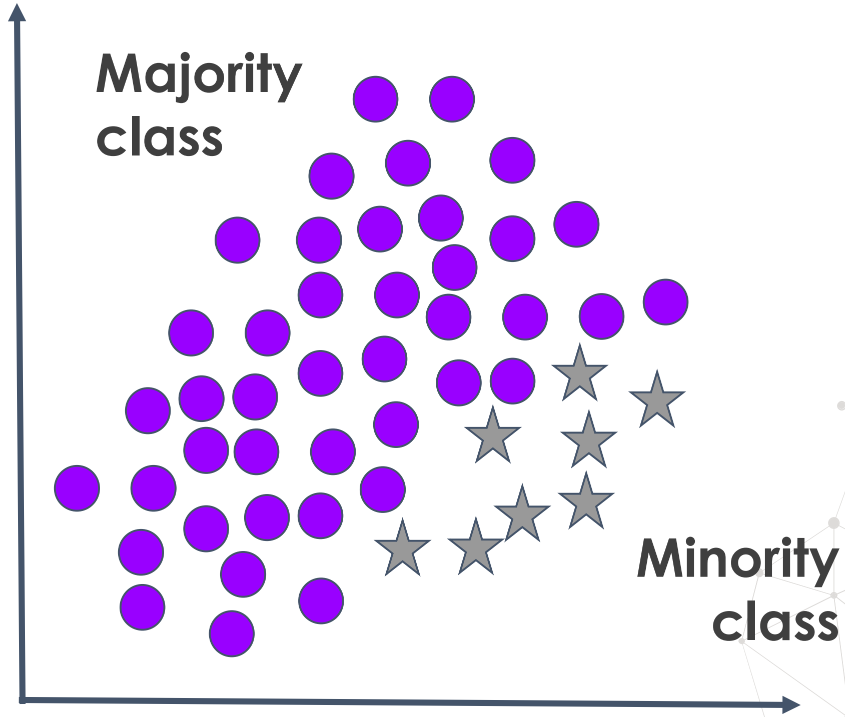
Train In Data

Imbalanced Datasets

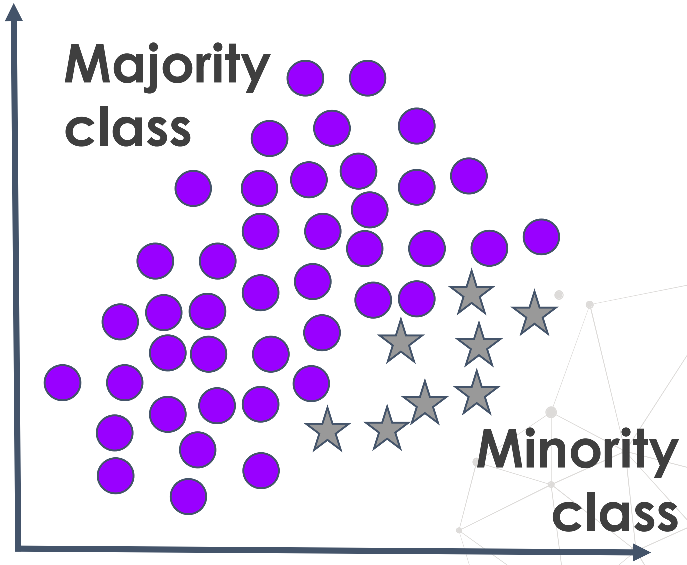
Imbalanced Datasets

Imbalanced datasets have many more instances of certain classes than of others.



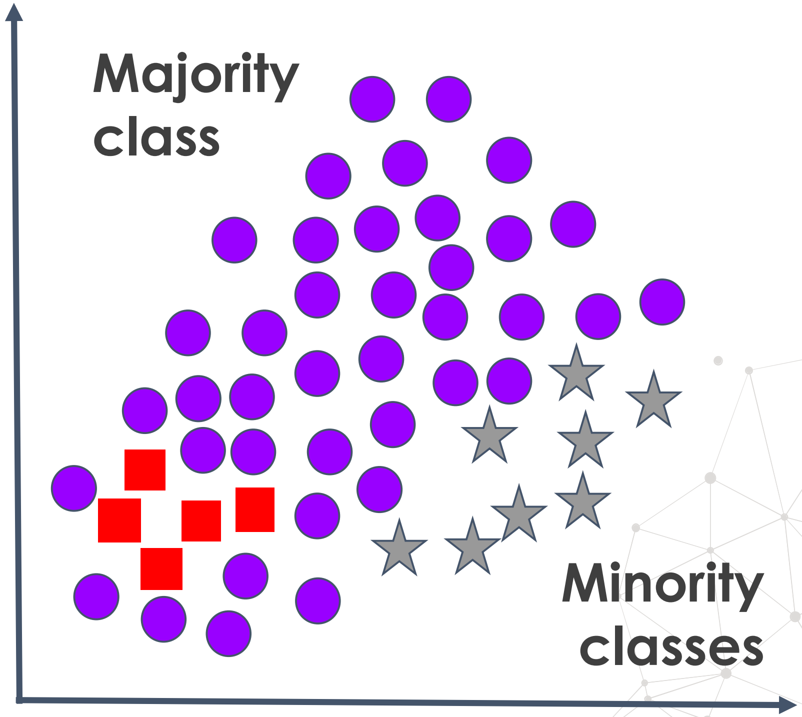
Imbalanced Datasets: The Problem

- Most machine learning algorithms assume balanced distributions
- As the minority examples occur rarely, rules to predict the small classes are difficult to find
- Samples from the minority class are most often misclassified
- **Particularly interested in the minority class!!!**



Imbalanced class distribution

- **Class distribution**: the proportion of instances belonging to each class.
- Imbalanced data-sets can have 1 or more minority classes
- **Imbalance degree**: ratio of the sample size of the minority class to that of the majority class i.e., 1:100
- Typical imbalanced ratios are 1:10 and smaller



The recurrence of imbalanced datasets in many real-world applications has sparked a huge amount of research.

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