Week 13

J.H. Lee, Dept. of Software, Sungkyunkwan Univ.

Imbalanced data

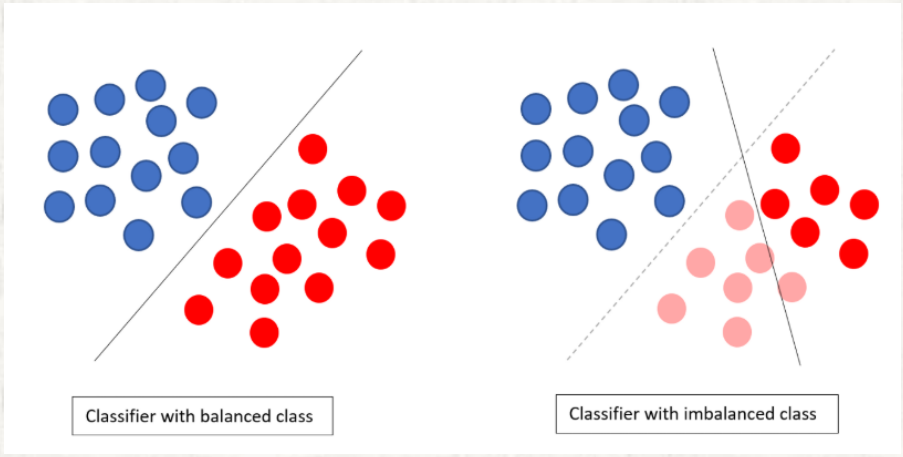
- Questions
  - Does bigger dataset always ensure higher performance?
  - Why do we need more data to learn better?

  - Is there any point that learning is saturated?
  - Then, How much data do we need?

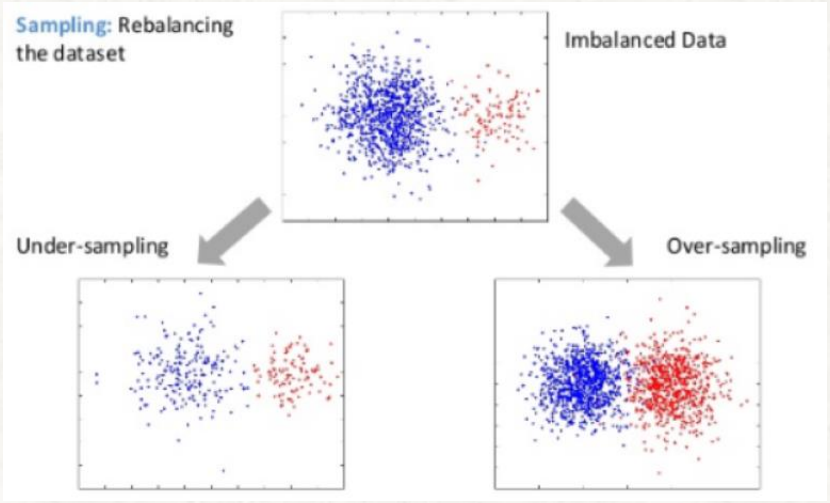
  - Should the number of data in each class be similar?
  - If so, why?
  - If not, why?

Imbalanced data

- Imbalanced data problem



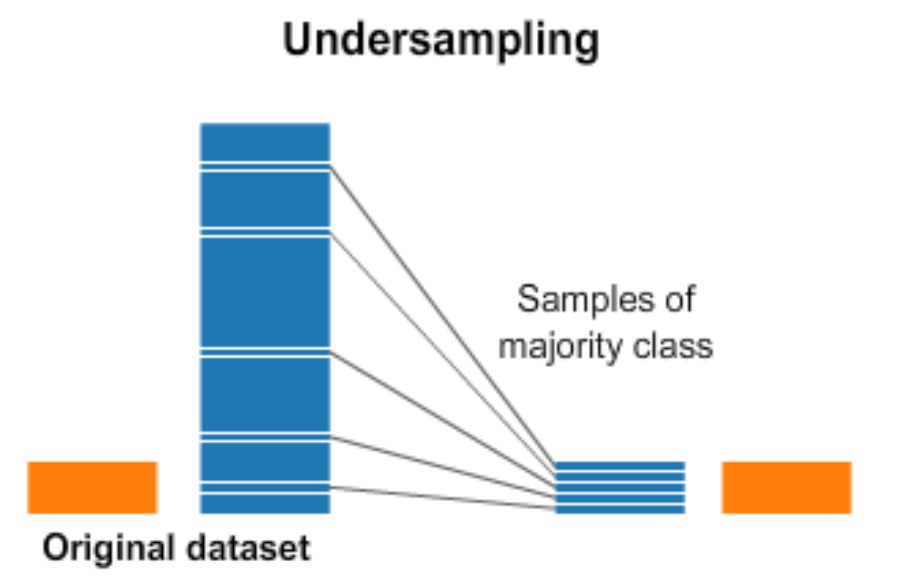
Problem



Solving method

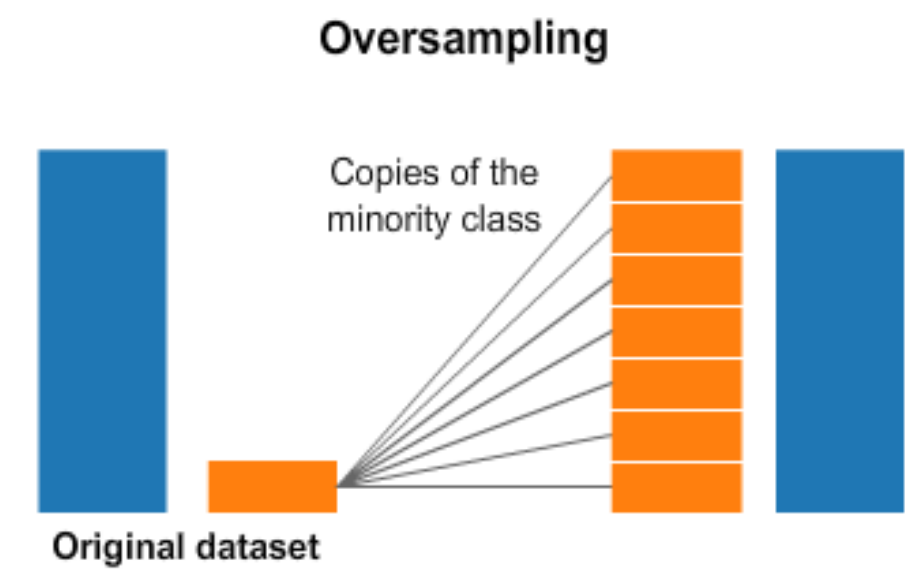
Imbalanced data

- Undersampling technique
  - reducing the data by eliminating examples belonging to the majority class with the objective of equalizing the number of examples of each class



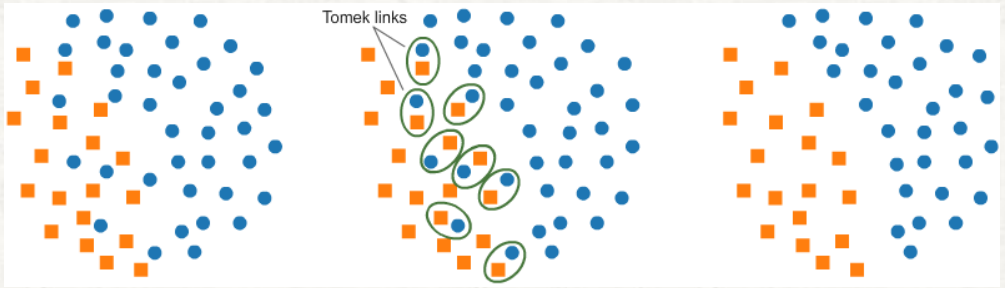
Imbalanced data

- Oversampling technique
  - Increasing the data by duplicating examples belonging to the minority class with the objective of equalizing the number of examples of each class

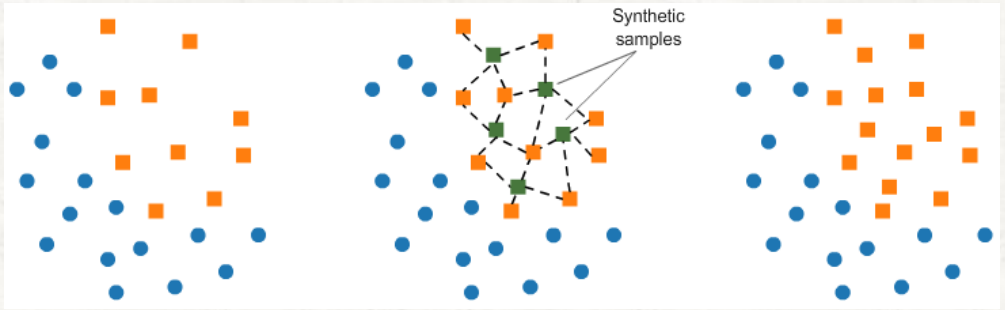


Imbalanced data

- Undersampling technique
  - Tomek links

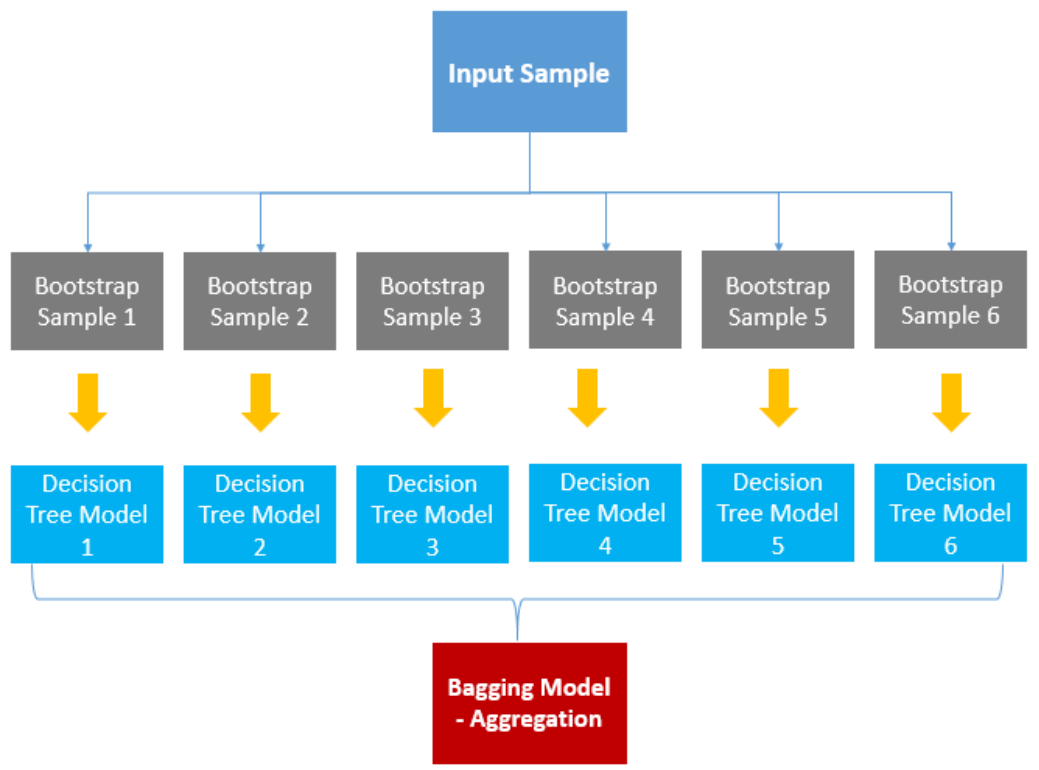


- Oversampling technique
  - SMOTE



Imbalanced data

- Oversampling Algorithm
  - Bootstrapping
  - Bagging



Imbalanced data

- Oversampling Algorithm
  - Boosting