2695 Introduction to Machine Learning
Masters Program in Economics, Finance and Management

NOVA SCHOOL OF BUSINESS & ECONOMICS

5c Video Lecture

CLASS IMBALANCE

Imbalanced classification

- **Class distribution:** the number of examples that belong to each class.
- **Imbalanced classification:** classification tasks where the distribution of examples across the classes is not equal.
- Imbalanced classification: class distribution is severely skewed such that for each example in the minority class there are many more examples in the majority class.
- Many real-world problems that we are interested in solving are imbalanced:
    - Fraud Detection
    - Churn Prediction
    - Spam Detection
    - Anomaly Detection
    - Intrusion Detection
    - Conversion Prediction

Causes of Class Imbalance

- Data sampling:
  - Biased Sampling: data was collected from a narrow geographical region, or slice of time, or from a particular segment of customers
  - Measurement Errors (data was mislabeled)

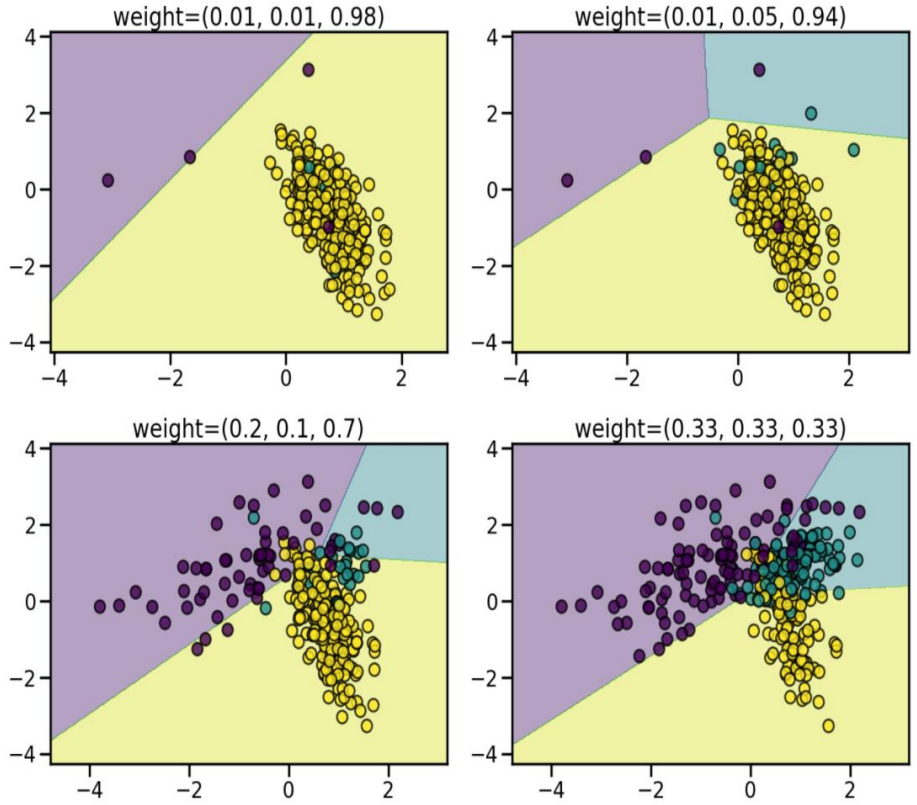
- Property of the phenomenon:
  - Asymmetric data: fraud
  - Asymmetric cost: disease detection

Class Imbalance terminology

- Terminology:
  - **Majority class**: the class or classes with many examples
  - **Minority class:** the class with fewer examples (and there is typically just one)
- **Minority class is typically of the most interest:**
  Model's skill in correctly predicting the class label or probability for the minority class is more important than the predicting the majority class or classes.
- A slight imbalance is often not a concern, and the problem can often be treated like a normal classification problem. A severe imbalance of the classes can be challenging to a model.

Challenges of Class Imbalance

- The minority class is harder to predict because there are fewer examples of this class: more challenging to learn the characteristics of examples from this class, and to differentiate examples from this class from the majority class:
  - If 99% of examples in a dataset belong to one class, then standard models fit on this dataset would focus attention on the majority class at the expense of the minority class.
  - Accuracy is dangerously misleading: if 99% of examples in a dataset belong to one class, a model that always predicts that class will achieve a classification accuracy of 99%.



Decision function of logistic regression:

With a greater imbalance ratio, the decision function favors the class with the larger number of samples.

Techniques to deal with Imbalanced Classification

- Select **performance metrics** that focus on the minority class, and tuning the threshold for mapping the score to the class label:
  - Precision
  - Recall
  - F1-score
  - ROC

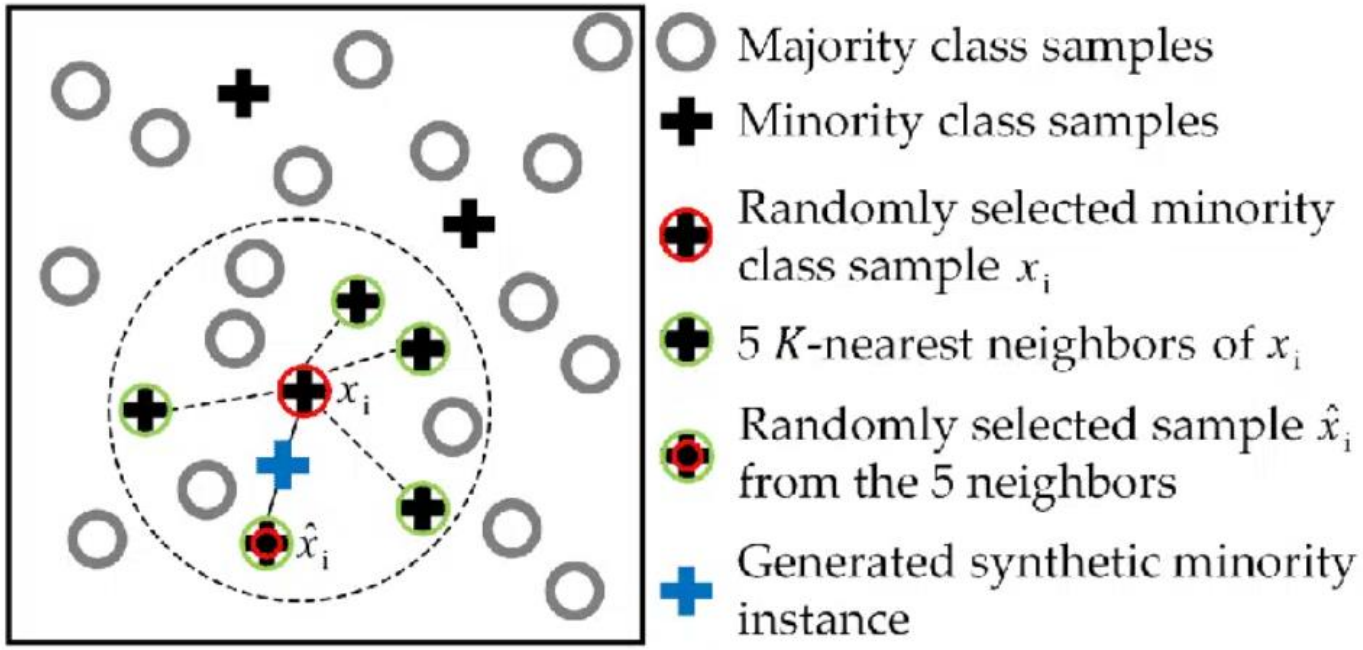
- Select **data preparation methods** that attempt to re-balance the classes in the training set.
- Select **classification algorithms, such as those that penalize misclassification** errors differently.
- Select a combination or all of the above techniques.

Balancing the dataset

- Collecting more data from minority class.
- **Under-sampling**: deleting instances from the majority class. This method is used when we have sufficient data. All instances from the minority class are kept, and equal number of instances are sampled from the majority class.
- **Over-sampling**: adding copies of instances from the minority class. This method is used when the quantity of data is insufficient. The simplest approach is to add copies of random instances.
- **Generating synthetic samples** (example: SMOTE algorithm and its extensions).
- Combination of the above approaches.

Synthetic Minority Oversampling Technique (SMOTE)

- Adds synthetic interpolated data to minority class.
- For each sample in minority class:
  - Pick a data instance and its random neighbor from k neighbors
  - Pick point on the line connecting the two samples
  - Repeat (until a desired percentage is reached)



https://rikunert.com/smote_explained

Data balancing and data leakage

- Data leakage occurs when information that would not be available at prediction time is used when building the model.
- Data balancing **is only performed on the training dataset** in order that the algorithm learns a model.
- Balancing is **not performed on the holdout test or test dataset**.
- Doing the balancing on the entire dataset before splitting it into a train and a test partitions leads to two problematic issues:
    - The resampling procedure might use information about test samples to either generate or select some of the samples for training.
    - The model will not be tested on a dataset with class distribution similar and representative to the real use-case: misleading and perhaps overly optimistic estimation of performance.

Class weights

- Modify the ML algorithms to take into account the skewed distribution of the classes.
- Give different weights to the majority and minority classes.
- The difference in weights will influence the classification of the classes during the training phase.

  Set a higher class weight for minority class and at the same time reduce weight for the majority class.
- Some algorithms have their own penalized versions, or there are implementations of generic frameworks that apply a custom penalty matrix.