Imbalanced classification

Part 3: Undersampling and cost-sensitive learning

By: Noureddin Sadawi, PhD

University of London

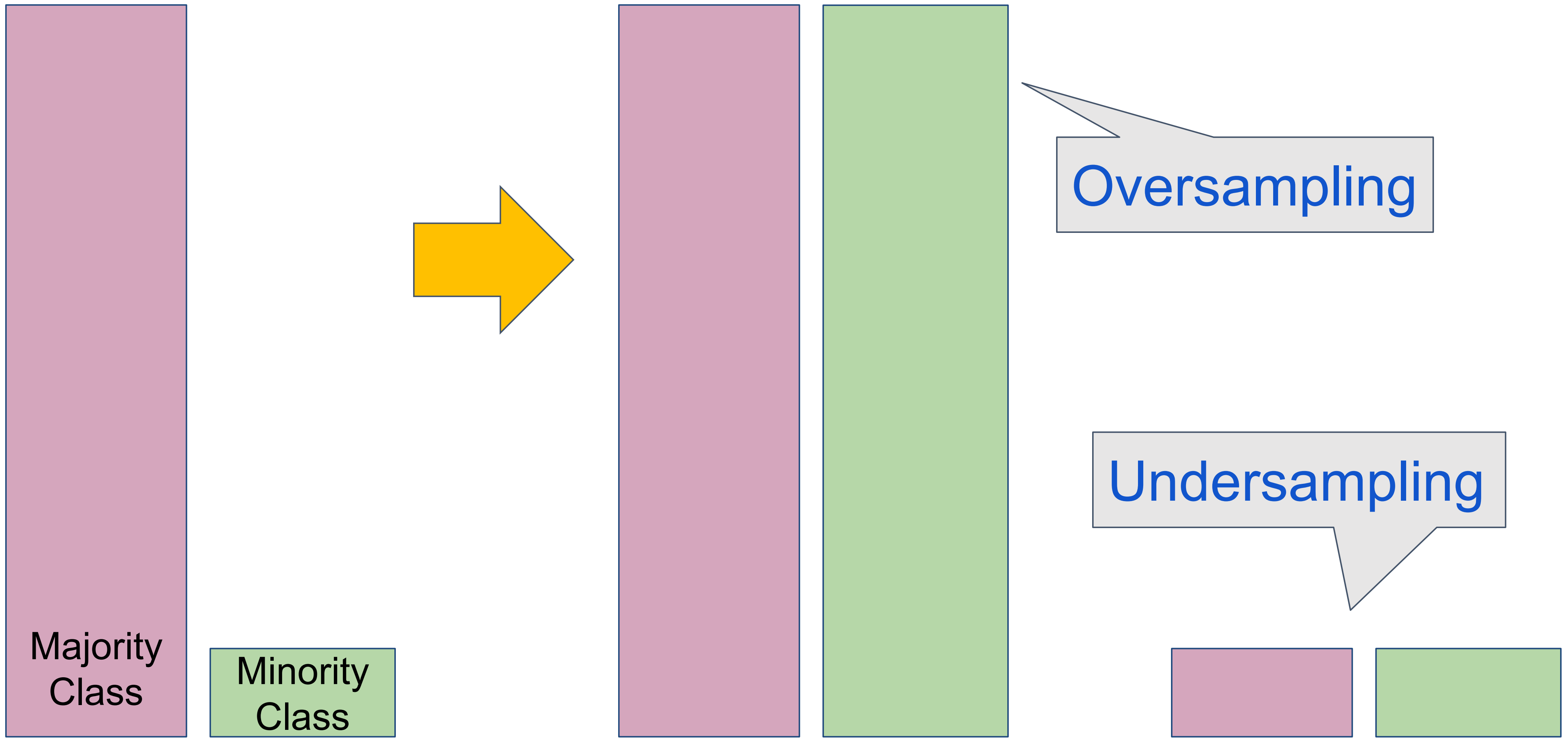
Undersampling

Undersampling methods delete or select a subset of examples from the majority class.

Example methods:

- random undersampling
- near miss undersampling
- condensed nearest neighbour rule (CNN)
- Tomek links undersampling
- edited nearest neighbours rule (ENN)
- one-sided selection (OSS)
- neighbourhood cleaning rule (NCR).

Data sampling



Some undersampling methods

- **Random undersampling:** as the name suggests, here we randomly delete examples from the majority class in the training dataset until the data is balanced.
- **Near miss undersampling:** uses kNN to select examples from the majority class that have the smallest average distance to the X closest/furthest examples from the minority class.
- **Condensed nearest neighbour rule (CNN):** uses a 1 nearest neighbour rule to find the subset of examples that can correctly classify the entire original dataset.

Combining over/undersampling methods

- Usually using one method or the other on the training dataset is effective.
- In some cases applying both types of techniques together can result in better overall performance of a model fit on the resulting transformed dataset.
- The purpose is to remove noisy points along the class boundary from both classes.
- Some common combinations:
    a. SMOTE and random undersampling
    b. SMOTE and Tomek links
    c. SMOTE and edited nearest neighbours rule.

Cost-sensitive learning

- Taking the costs of prediction errors (and potentially other costs) into account when training a machine learning model (a subfield of machine learning).
- Related to the field of imbalanced learning (which is concerned with classification on datasets with a skewed class distribution).
- Many techniques developed and used for cost-sensitive learning can be adopted for imbalanced classification problems.
- Based on the concept: **not all classification errors are equal.**

Classification errors

- **Majority class**: negative or no-event assigned the class label 0.
- **Minority class**: positive or event assigned the class label 1.
- In imbalanced classification, classifying a negative case as a positive case is typically far less of a problem than classifying a positive case as a negative case (**false positive vs false negative**).
- Remember: the goal of a classifier on imbalanced binary classification problems is to detect the positive cases correctly, and positive cases represent an exceptional event that we are most interested in.
- Predicting a positive case as a negative case is more harmful and more costly.

Cost of classification errors

- **Error minimisation**: the conventional goal when training a machine learning algorithm is to minimise the error of the model on a training dataset.
- **Cost**: the penalty associated with an incorrect prediction.
- **Cost minimisation**: the goal of cost-sensitive learning is to minimise the cost of a model on a training dataset.
- Assigning different costs to the types of misclassification errors that can be made, then using specialised methods to take those costs into account.

Cost matrix

- The varying misclassification costs are best understood using the idea of a cost matrix.
- **Cost matrix**: a matrix that assigns a cost to each cell in the confusion matrix.

		Predicted Class	
		Positive	Negative
Actual Class	Positive	0	C(FN)
	Negative	C(FP)	0