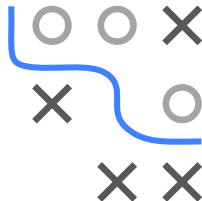
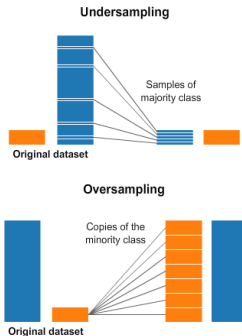


SAMPLING METHODS: OVERVIEW

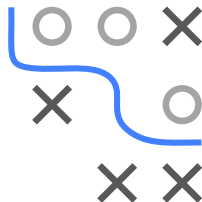
- Balance training data distribution to perform better on minority classes.
- Independent of classifier \rightsquigarrow very flexible and general.
- Three groups:

- Undersampling — Removing instances of majority class(es).
- Oversampling — Adding/Creating new instances of minority class(es). (Slower, but usually works better.)
- Hybrid — Combining both methods.



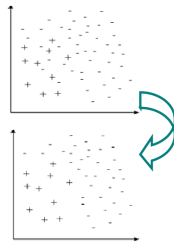
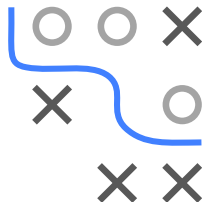
RANDOM UNDERSAMPLING/OVERSAMPLING

- Random oversampling (ROS):
 - Randomly **replicate minority** instances.
 - Prone to overfitting due to multiple tied instances.
- Random undersampling (RUS):
 - Randomly **eliminate majority** instances.
 - Might remove informative instances and destroy important concepts in data.
- Better: Introduce heuristics in removal process (RUS) and do not create exact copies (ROS).



UNDERSAMPLING: TOMEK LINKS

- Remove “noisy borderline” examples (very close observations of different classes) of majority class(es).
- Let $E^{(i)} = (\mathbf{x}^{(i)}, y^{(i)})$ and $E^{(j)} = (\mathbf{x}^{(j)}, y^{(j)})$ be two data points in \mathcal{D} .
- A pair $(E^{(i)}, E^{(j)})$ is called *Tomek link* iff there is no other data point $E^{(k)} = (\mathbf{x}^{(k)}, y^{(k)})$ such that
$$d(\mathbf{x}^{(i)}, \mathbf{x}^{(k)}) < d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \text{ or } d(\mathbf{x}^{(j)}, \mathbf{x}^{(k)}) < d(\mathbf{x}^{(j)}, \mathbf{x}^{(i)})$$
holds, where d is some distance on \mathcal{X} .
- $y^{(i)} \neq y^{(j)} \rightsquigarrow$ noisy borderline examples.
- Remove majority instance in each data pair in a Tomek link where $y^{(i)} \neq y^{(j)}$.



Franciso Herrera (2013), Imbalanced Classification: Common Approaches and Open Problems ([URL](#)).

UNDERSAMPLING: OTHER APPROACHES

- Neighborhood cleaning rule (NCL):
 - 1 Find 3 nearest neighbors for each $(\mathbf{x}^{(i)}, y^{(i)})$ in \mathcal{D} .
 - 2 If $y^{(i)}$ is majority class *and* 3-NN classifies it as minority \rightsquigarrow Remove $(\mathbf{x}^{(i)}, y^{(i)})$ from \mathcal{D} .
 - 3 If $y^{(i)}$ is minority class *and* 3-NN classifies it as majority \rightsquigarrow Remove 3 nearest neighbors from \mathcal{D} .
- Condensed Nearest Neighbor (CNN): Construct a **minimally consistent** subset $\tilde{\mathcal{D}}$ of \mathcal{D} .
- One-sided selection (OSS): Tomek link + CNN
- CNN + Tomek link: to reduce computation of finding Tomek links \rightsquigarrow first use CNN and then remove the Tomek links.
- Clustering approaches: Class Purity Maximization (CPM) and Undersampling based on Clustering (SBC).

