Train In Data

One Sided Selection

One Sided Selection

The idea is to retain observations from the majority that are hard to classify, but remove the noise.

One Sided Selection

- First, selects samples at the boundary of the classes (hardest instances).
- Next, removes the Tomek Links (noise)

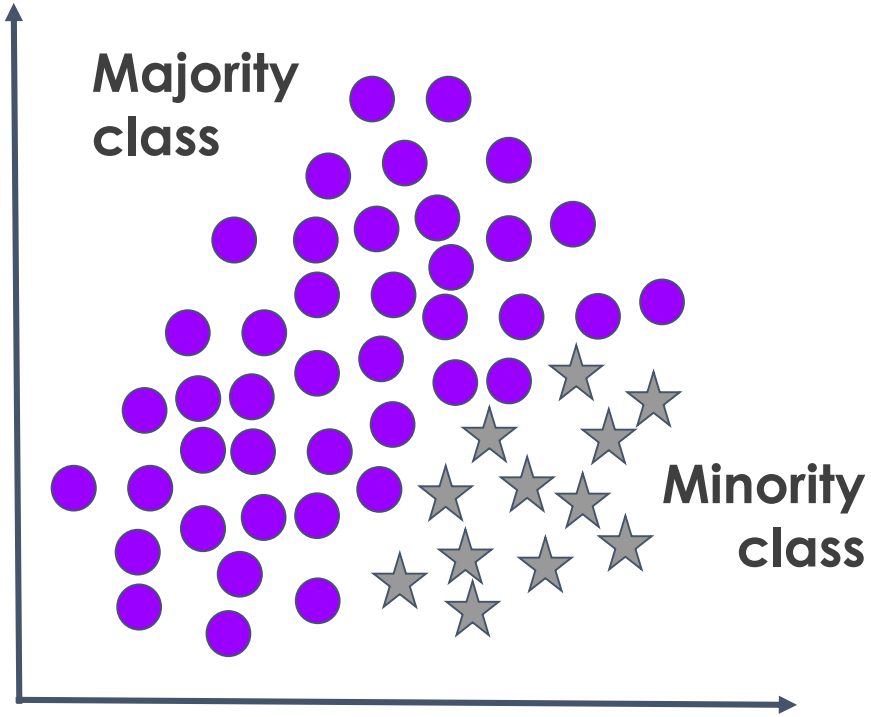
One Sided Selection

- Cleaning
- Final dataset shape varies
- Boundary matters

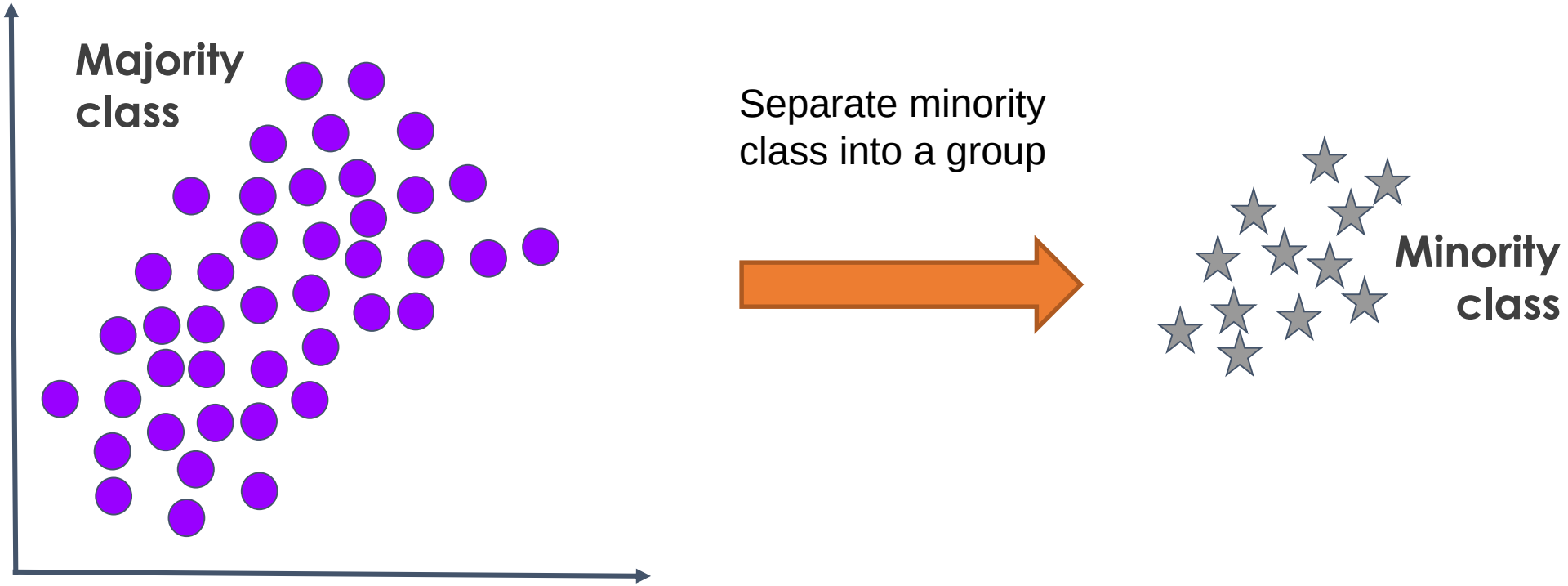
One Sided Selection: Procedure

1. Create group S with all samples from minority
2. Add 1 observation from the majority to S (at random)
3. Train a 1-KNN on S
4. Make predictions on the rest of the majority class observations
5. If predictions don't match the class, pass the samples to S
6. In S, find and and remove Tomek Links

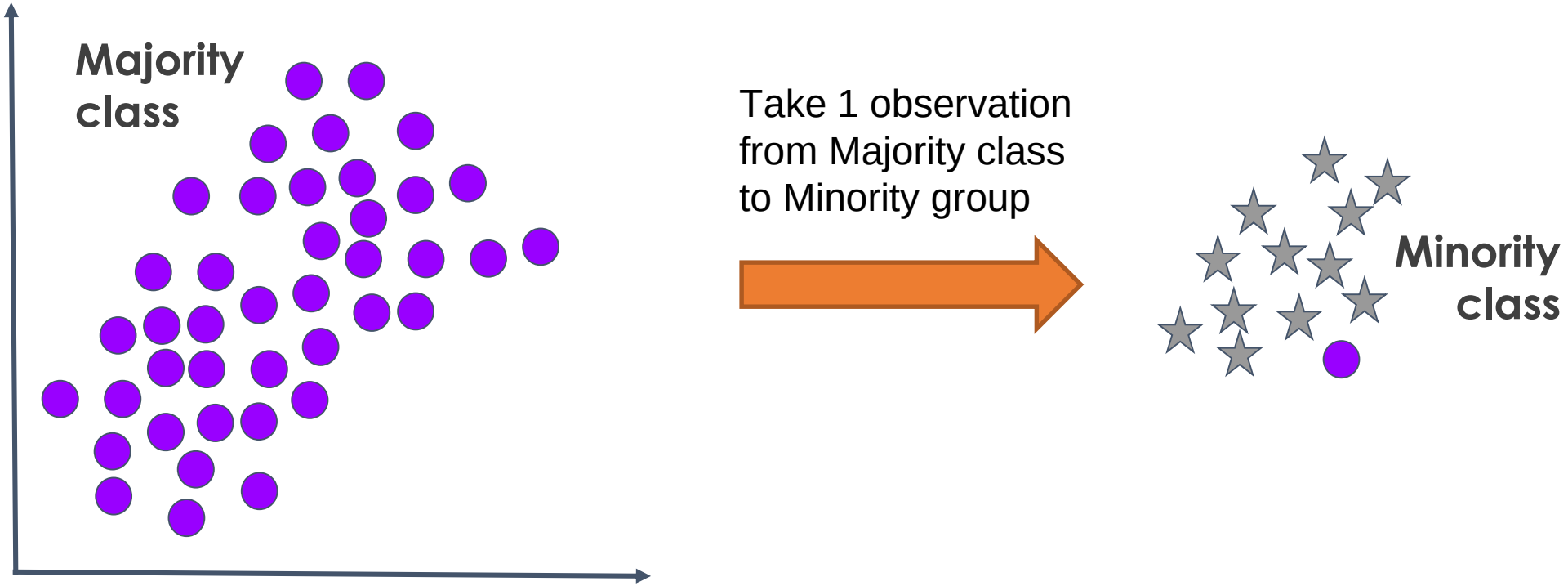
One Sided Selection



One Sided Selection

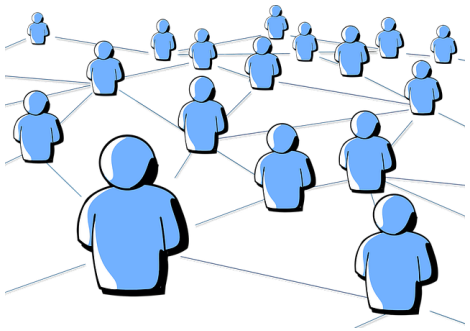


One Sided Selection

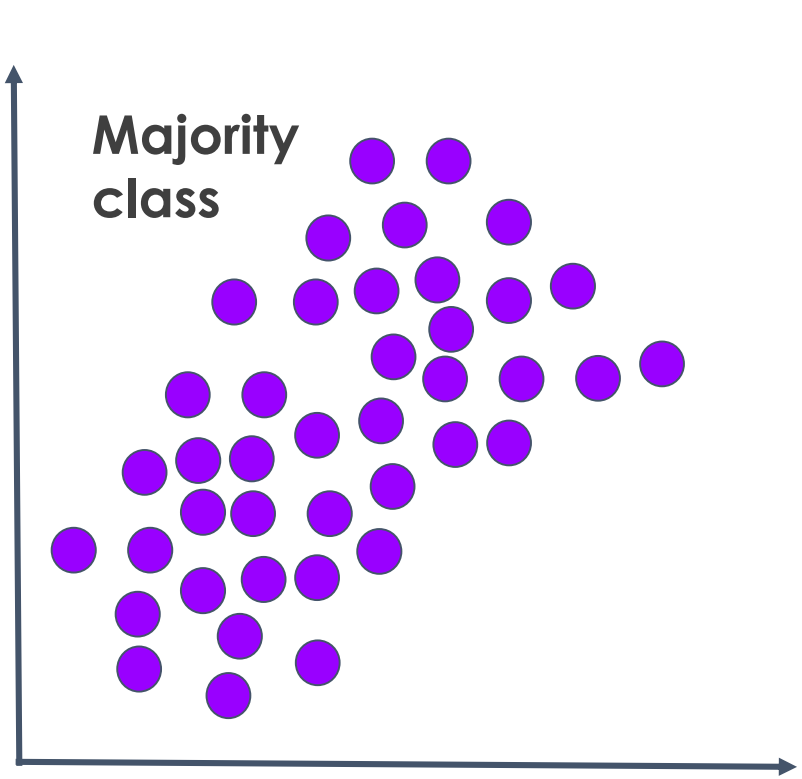


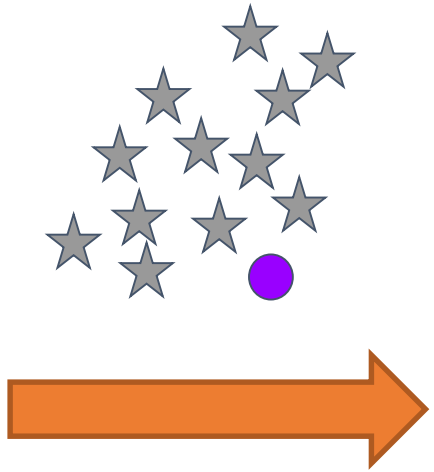
One Sided Selection

Train a 1 KNN algorithm

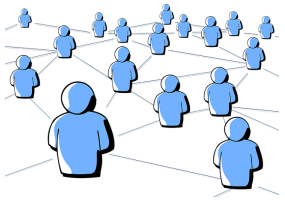


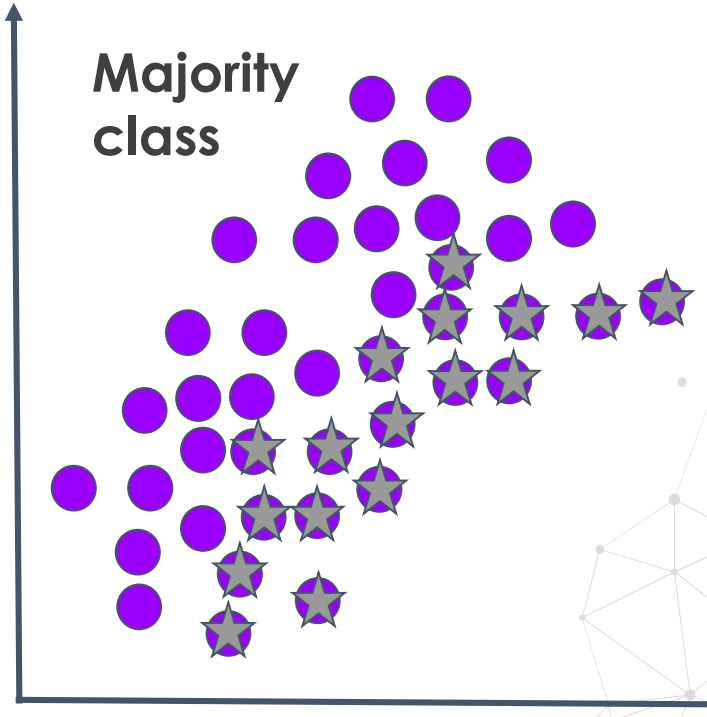
One Sided Selection



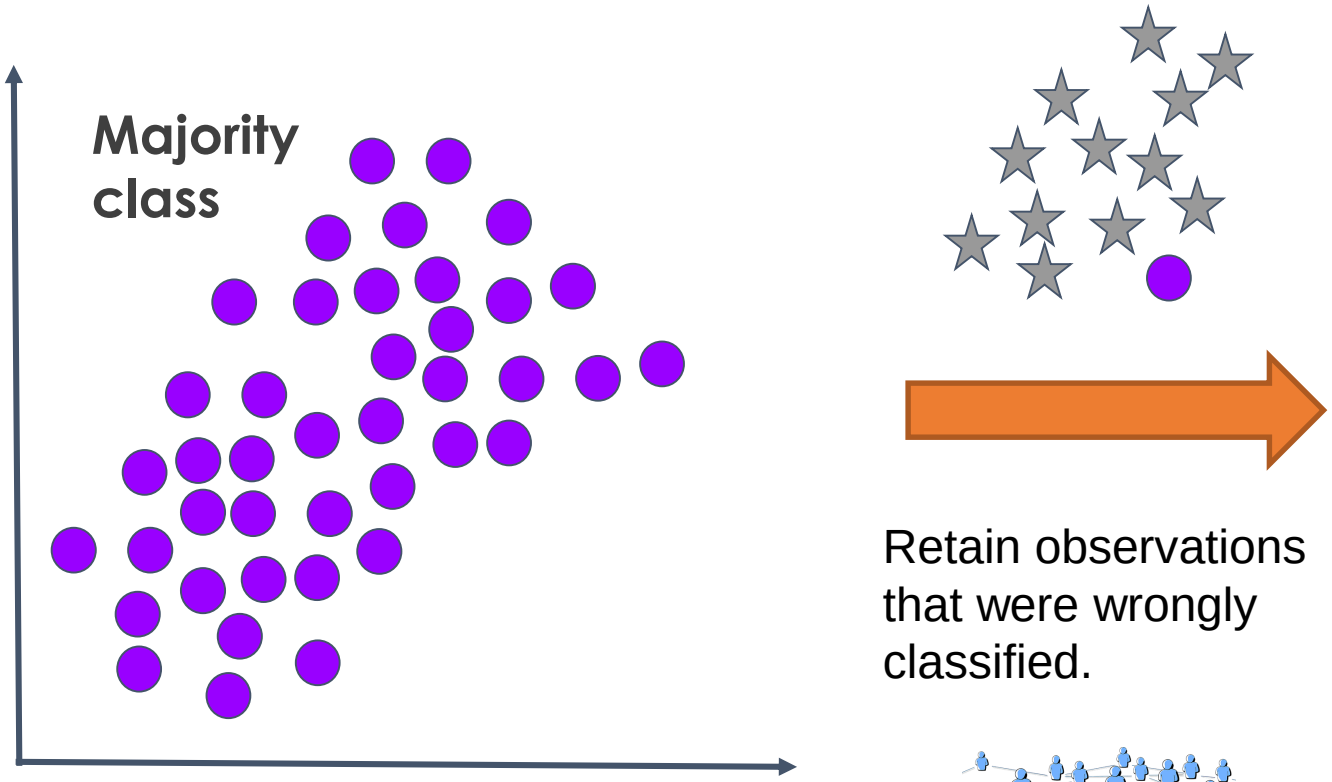


Use KNN to classify observations from majority class.

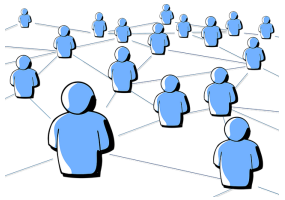


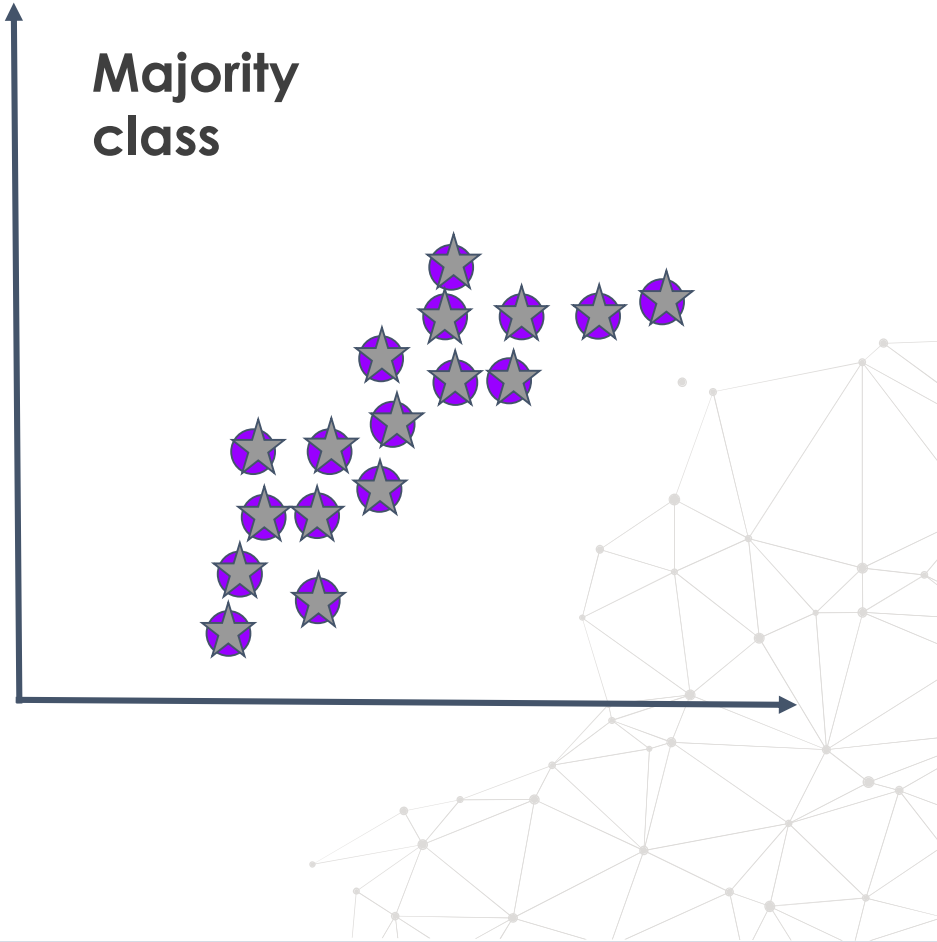


One Sided Selection



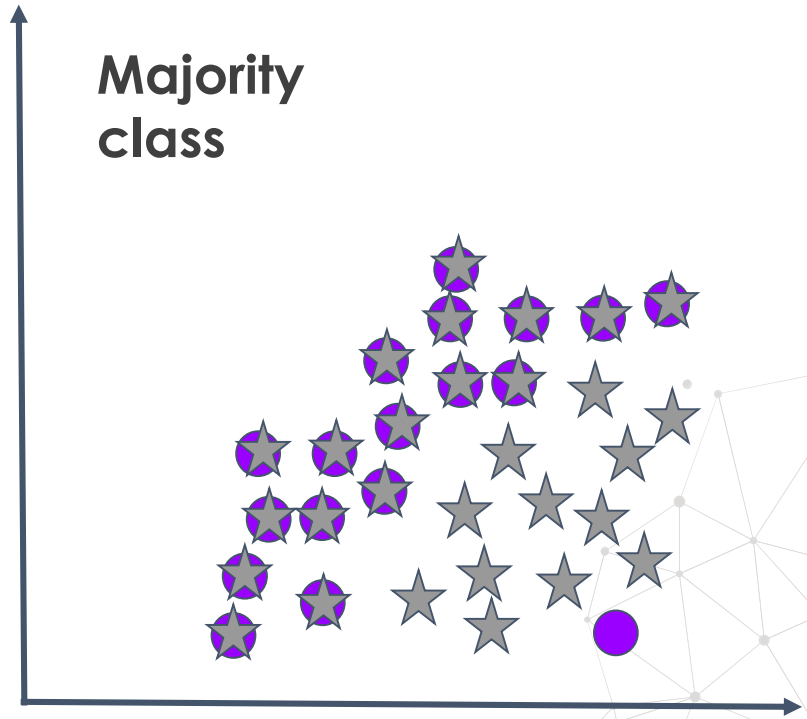
Retain observations that were wrongly classified.





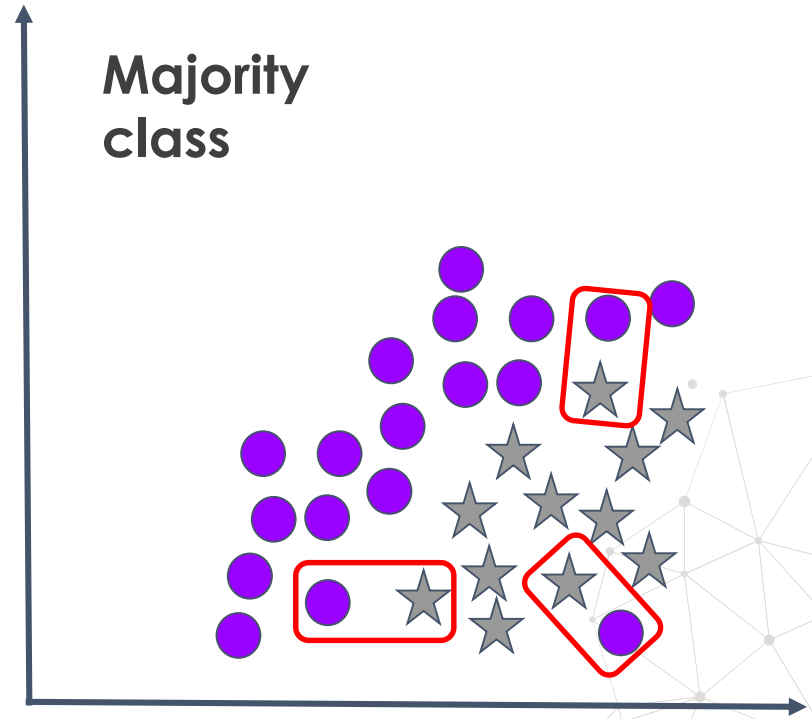
One Sided Selection

Resampled data consists of the minority group + those observations wrongly classified.



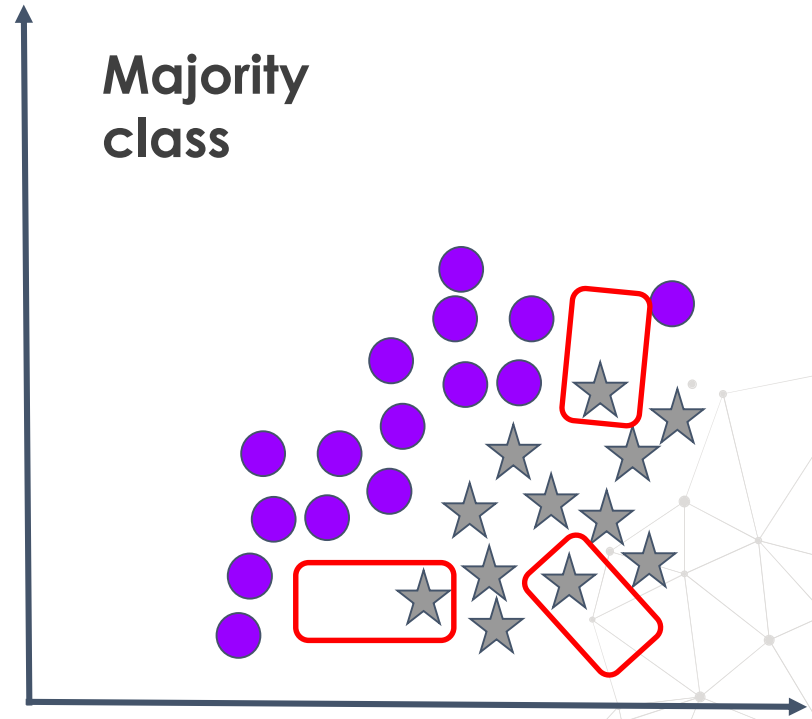
One Sided Selection

With the resampled data, find and remove the Tomek Links.



One Sided Selection

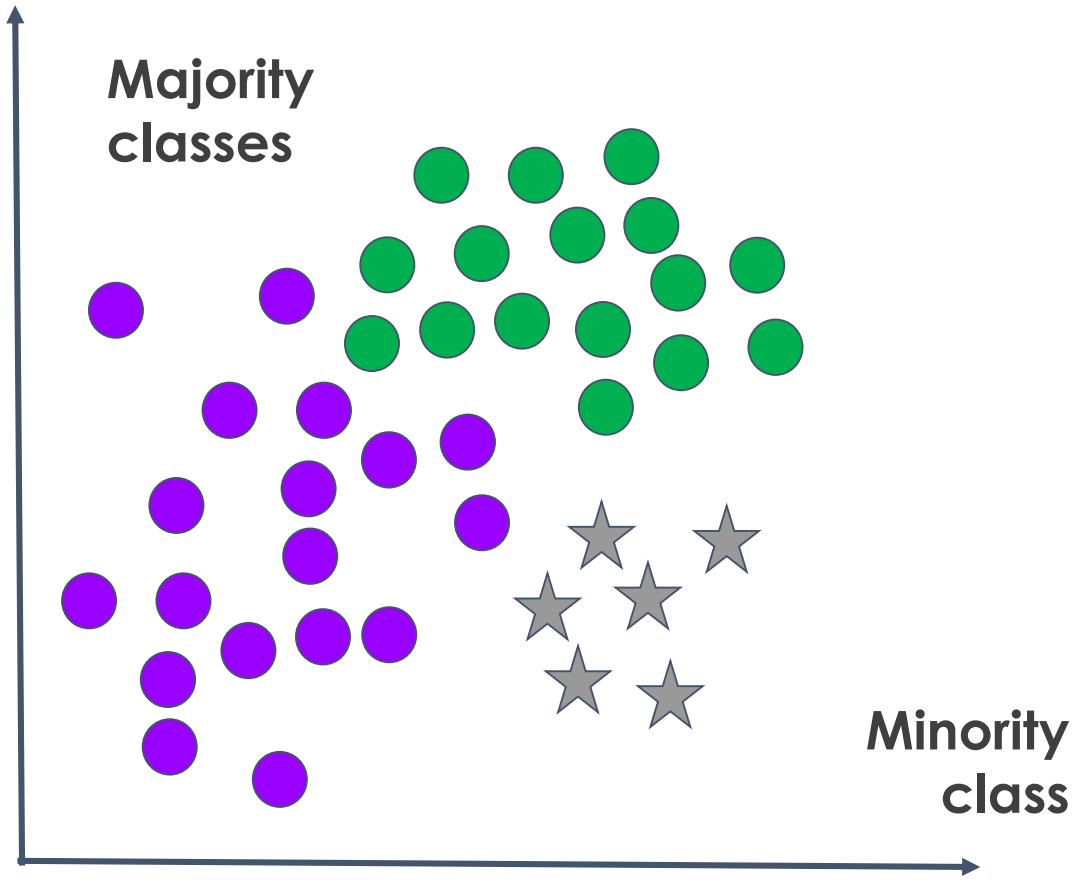
With the resampled data, find and remove the Tomek Links.



Imbalanced-learn: OSS

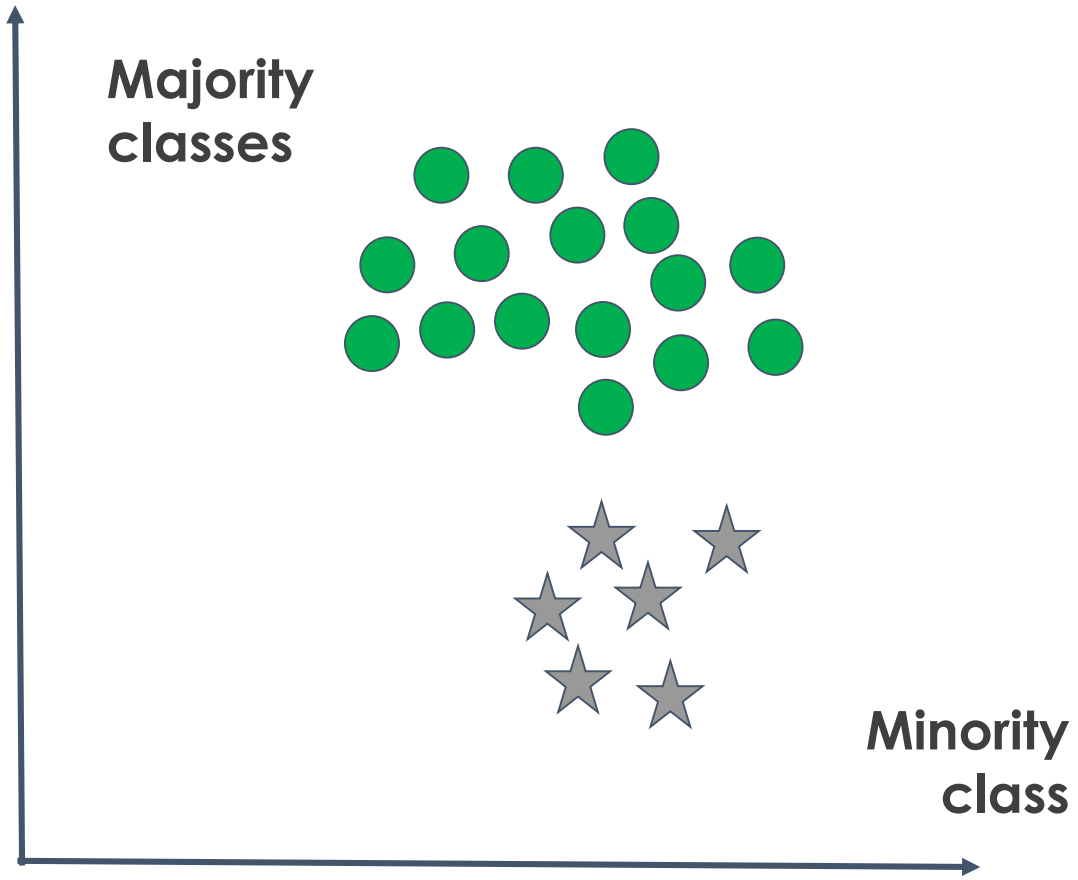
```
# create data

X, y = make_data(sep=2)

# set up OSS

oss = OneSidedSelection(
    sampling_strategy='auto',  # undersamples only the majority class
    random_state=0,  # for reproducibility
    n_neighbors=1,# default
    n_jobs=4)  # I have 4 cores in my laptop

X_resampled, y_resampled = oss.fit_resample(X, y)
```

Multi-class



One vs One.

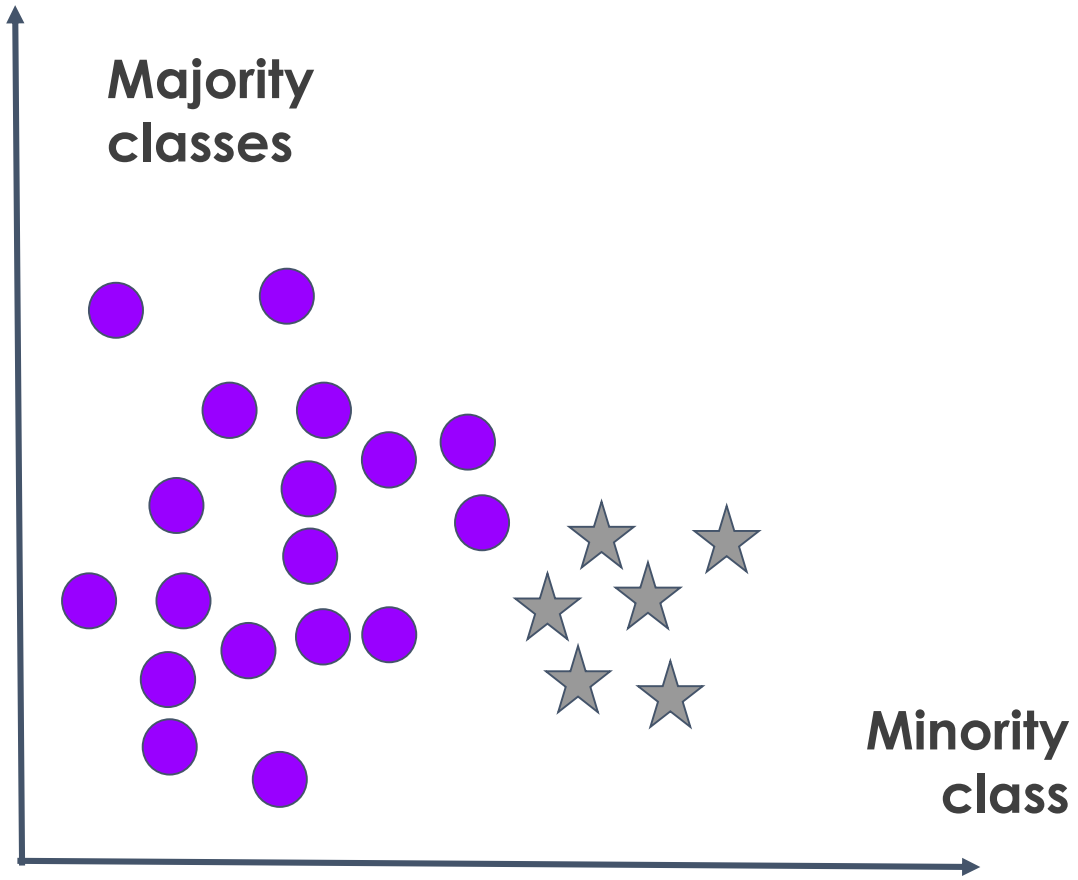
Multi-class



One vs One.

Run entire procedure over 1 majority class first

Multi-class



One vs One.

Repeat the procedure for the other majority class.

Train In Data

THANK YOU

www.trainindata.com