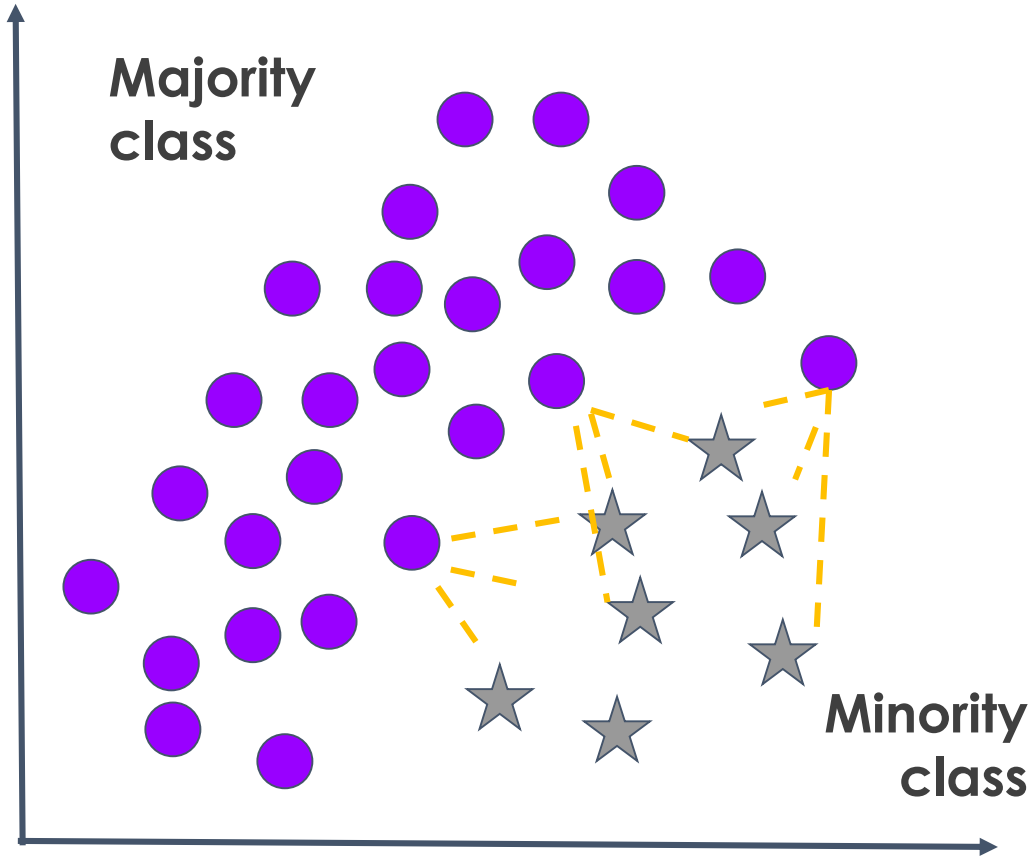
Train In Data

NearMiss

NearMiss

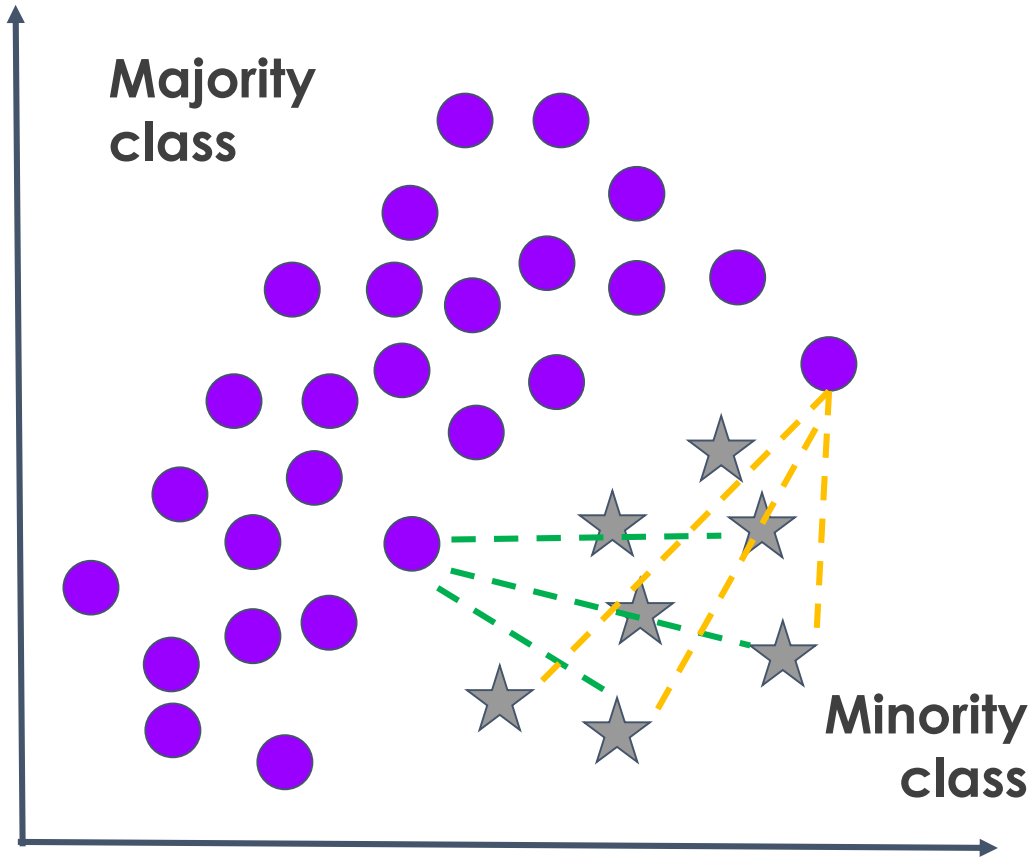
- 3 versions
  - Fixed method
  - Final dataset is 2 x minority for binary classification
  - Retains information closer to the minority class
  - Design to work with text, where each word is a complex representation of words and tags

NearMiss, version 1



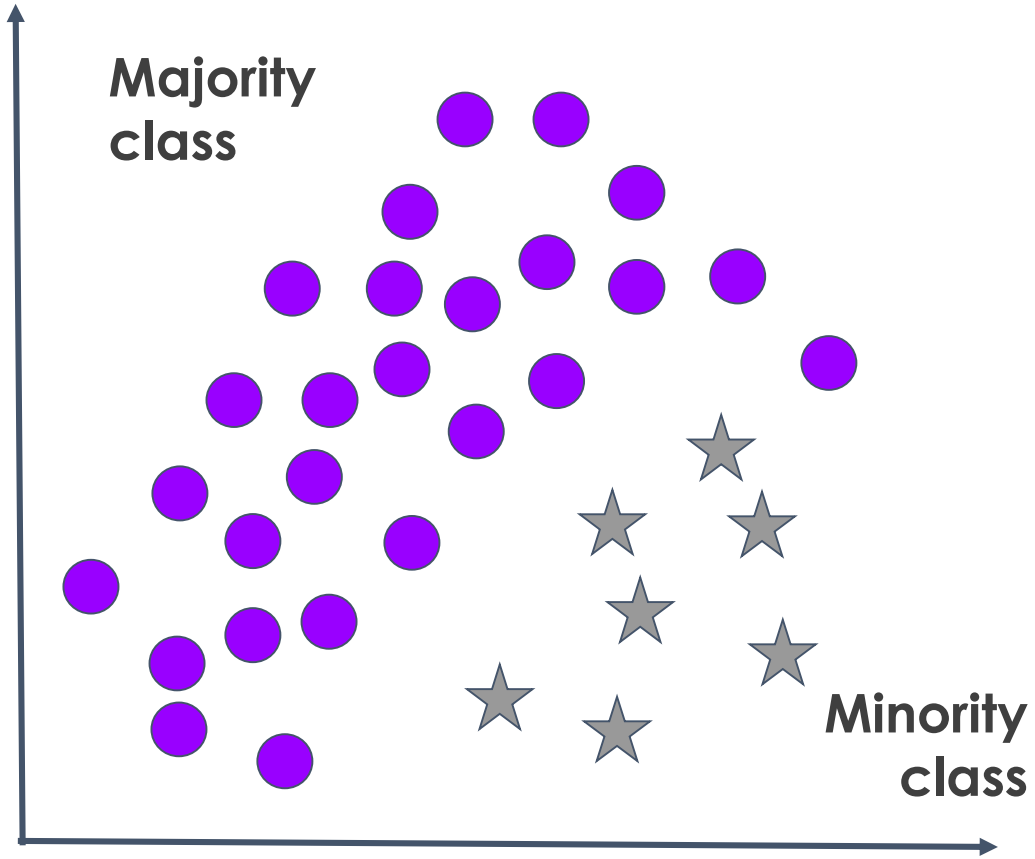
- Determine the mean distance to each k **closest** neighbour from X(min)
- Retain observations from X(maj) with the **smallest** average distance

NearMiss, version 2



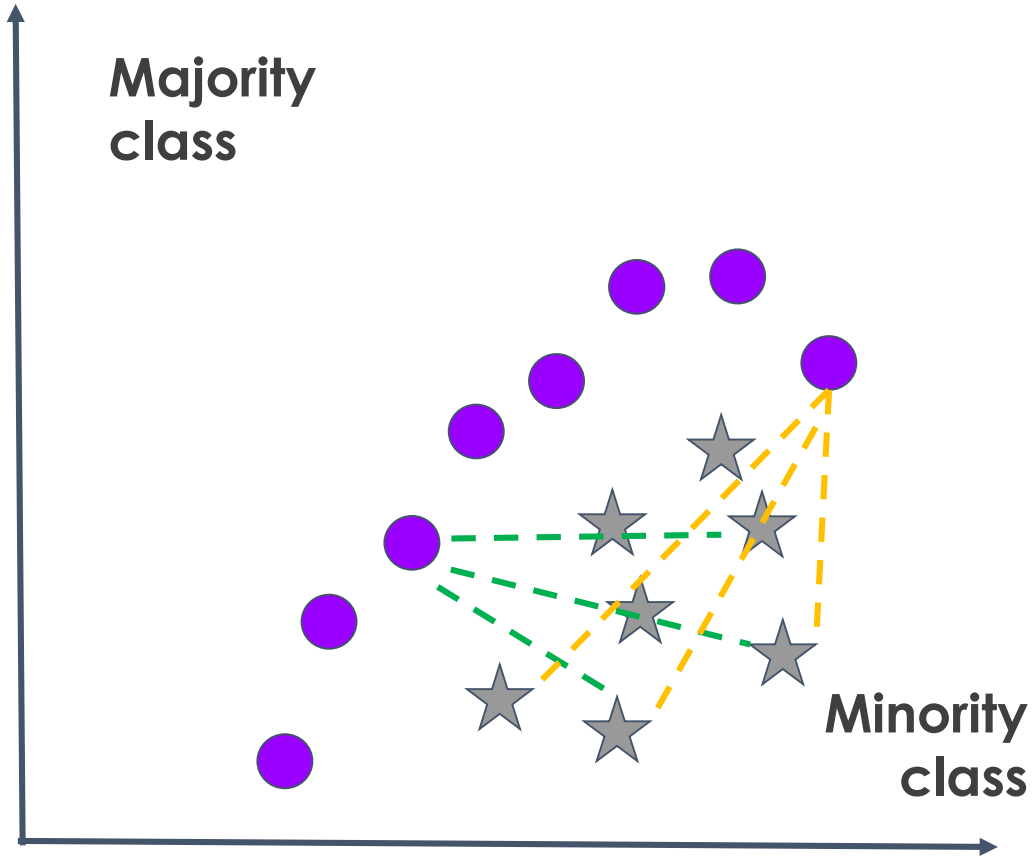
- Determine the mean distance to each k **furthest** neighbour from X(min)
- Retain observations from X(maj) with the **smallest** average distance

NearMiss, version 3



- Retain the 3 closest K to the minority sample
- Intermediate dataset

NearMiss, version 3



- Retain the 3 closest K to the minority sample
- Select those which average distance to X(min) is the largest

Imbalanced-learn: NearMiss

```
# create data

X, y = make_data(sep=2)

# set up Near Miss, first method
# that is, version = 1

nm1 = NearMiss(
    sampling_strategy='auto',  # undersamples only the majority class
    version=1,
    n_neighbors=3,
    n_jobs=4)  # I have 4 cores in my laptop

X_resampled, y_resampled = nm1.fit_resample(X, y)
```

Multi-class

One vs Rest

Train In Data

THANK YOU

www.trainindata.com