Machine Learning Graz

Dealing with Imbalanced Data

10-Oct-2018

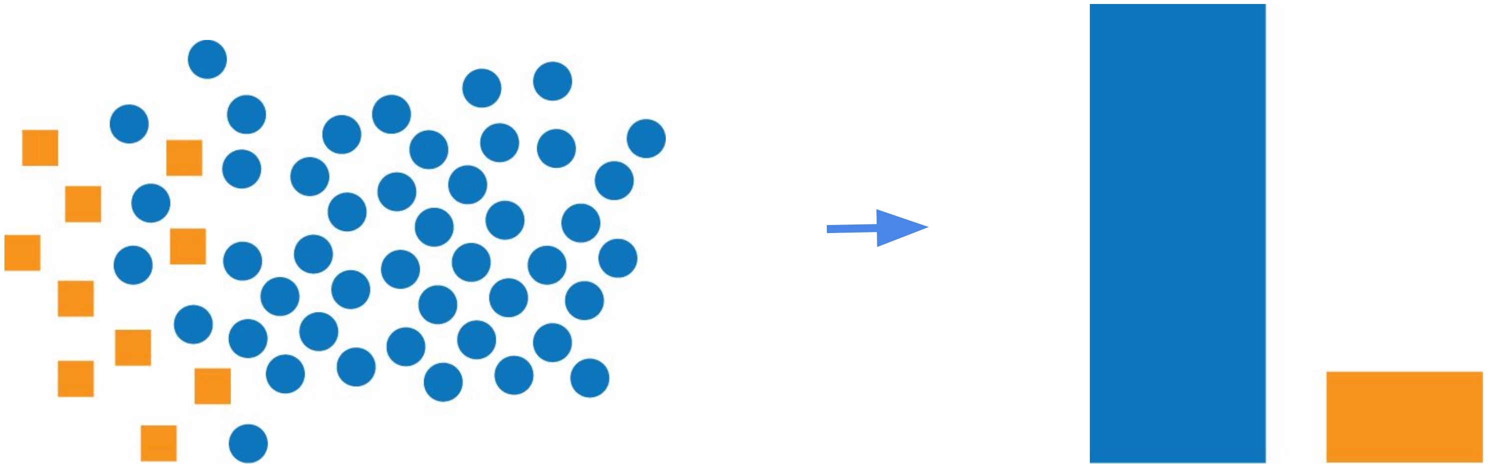
Adrian Spataru
Data Scientist at Know-Center
adrian@spataru.at
https://www.fb.me/adrian.spataru.5

Outline

- The Imbalanced Classes Problem
- Loss Function Weighing
- Undersampling Methods
- Oversampling Methods
- Feature Learning
- Feature Engineering
- Anomaly Detection
- Resources

What is unbalanced Data?

- When the minority class, is much rarer than the other classes.



Why is this a Problem?

- ML Algorithms perform poor in unbalanced data.
- Classifiers designed to optimize accuracy
- Assuming uniformity of misclassification costs

Fraud Detection

- Assume Fraud is only 1% of all transaction
- Create model and has 99% Accuracy

Fraud Detection

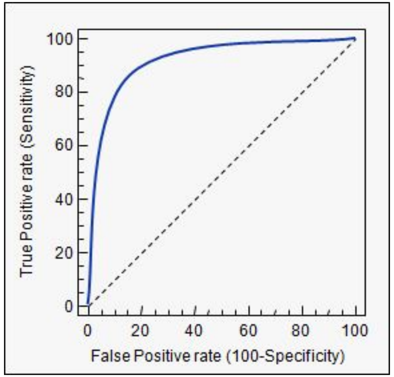
- Assume Fraud is only 1% of all transaction
- Create model and has 99% Accuracy
- Model classifies everything as not fraud.

Evaluation

- Don’t use Accuracy!

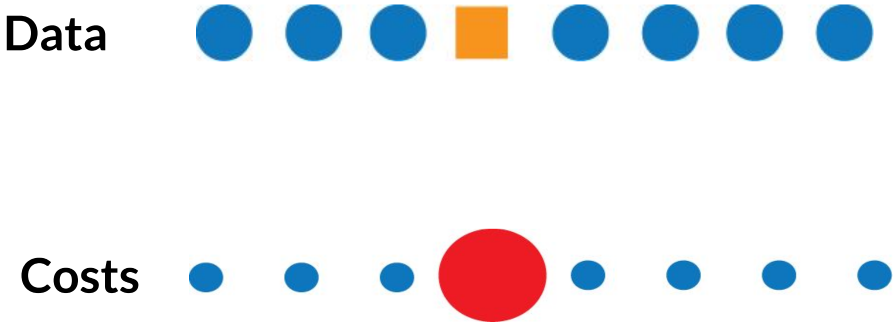
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- use Confusion Matrix
- use ROC Curves
- multiple metrics if possible



Fraud Detection

- Misclassify Fraud comes at a high cost.



Loss Weighting

- Most ML Algorithms have loss functions
- increase the loss when misclassify the minority class

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}(y_i - 1)h(x_i) + y_i(h(x_i) - 1)$$

Loss Weighting

- Most ML Algorithms have loss functions
- increase the loss when misclassify the minority class

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N}(y_i - 1)h(x_i) + y_i(h(x_i) - 1)$$

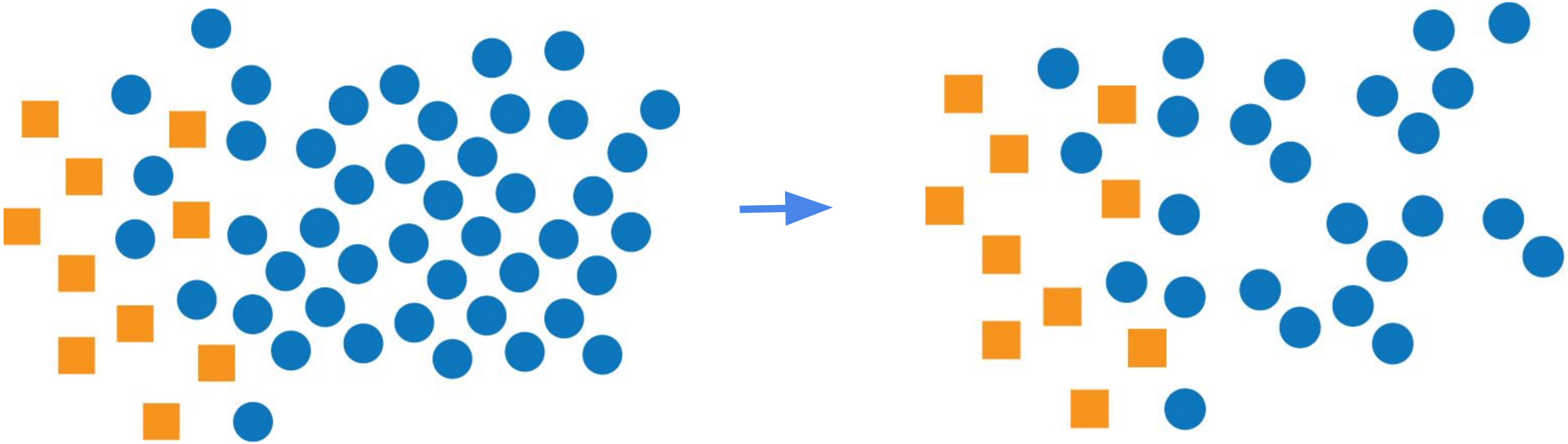
Multiply with a value

Sampling Methods

- If our data is imbalanced, we will can make it balanced!
- Generate/Remove data to get it balanced

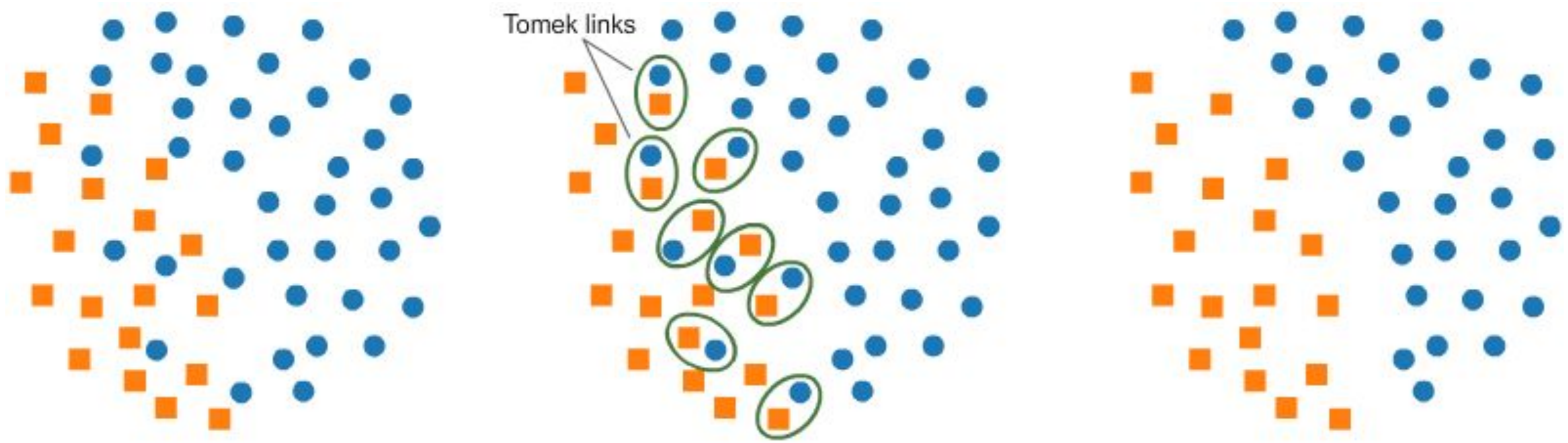
Random Undersampling

- Randomly Remove Data from Majority Class



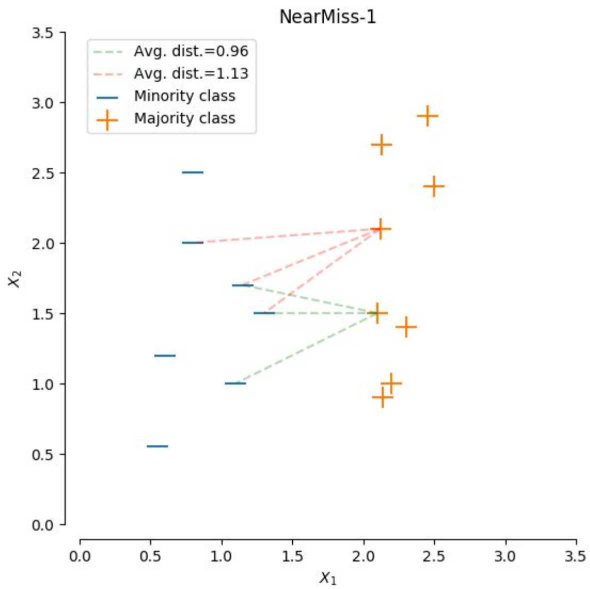
Tomek Links

- Tomek links are pairs of opposite classes which are close
- Increases the separation between classes



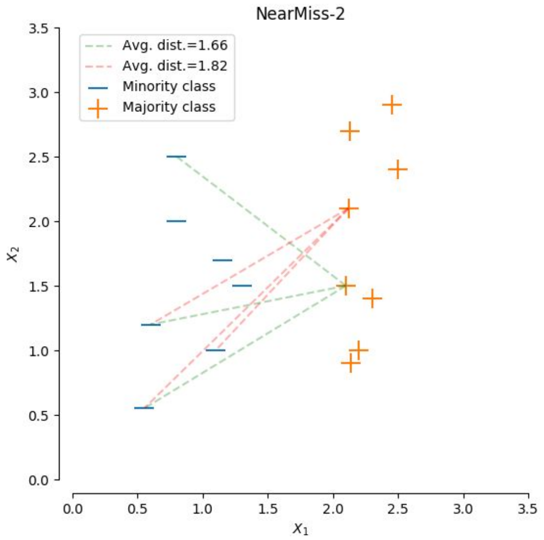
NEARMISS- 1

- NearMiss-1 selects samples from the majority class for which the average distance to K nearest neighbours is the smallest.

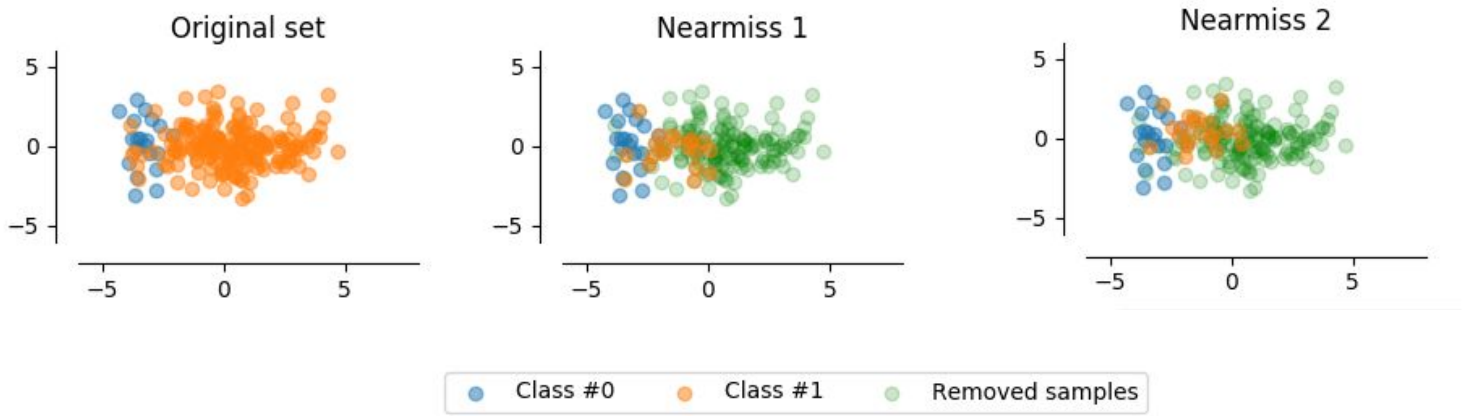


NEARMISS- 2

- NearMiss-2 selects samples from the majority class for which the average distance to the K farthest neighbors is the smallest.



Nearmiss combined

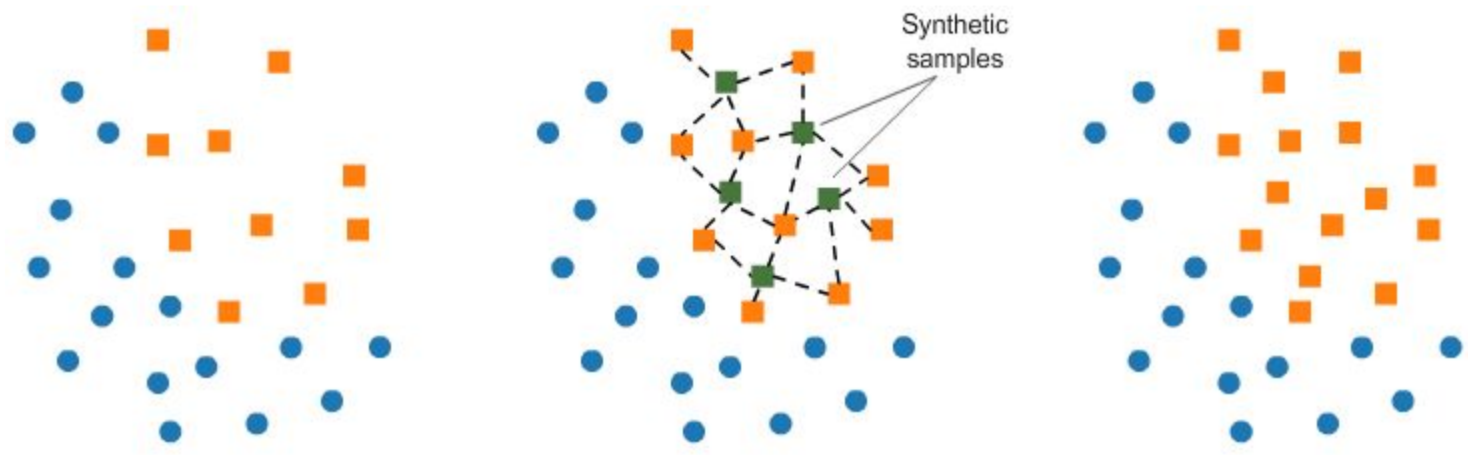


Random Oversampling

- Oversample minority class by randomly "copying" points from the class

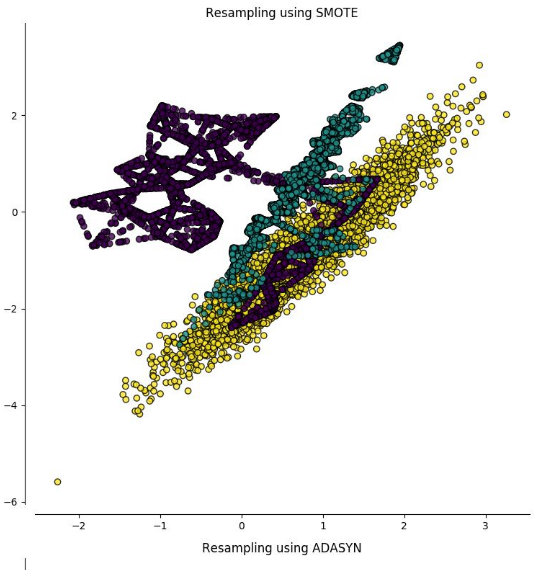
Synthetic Minority Oversampling Technique(SMOTE)

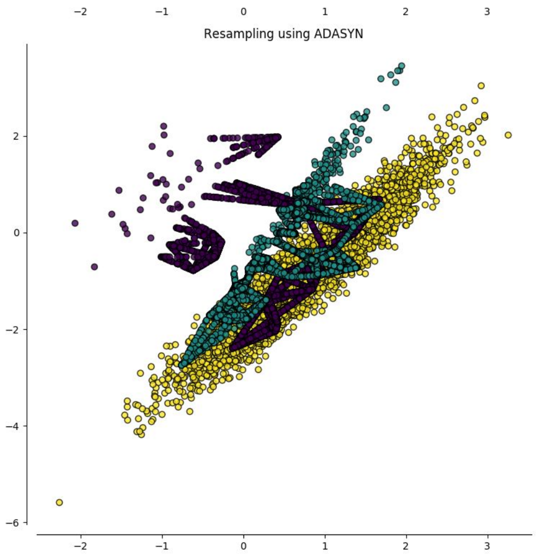
- Find "k-nearest neighbors" of an anchor point x_i from minority class
- Randomly select one of the nearest neighbors and interpolate randomly between the two



Adaptive Synthetic Sampling (ADASYN)

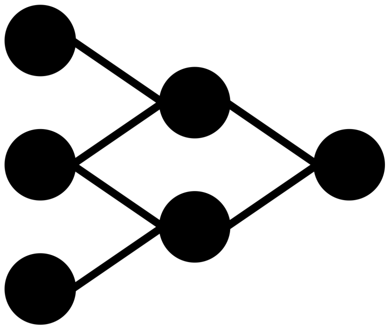
- Generate minority data samples according to their distributions
- more synthetic data for minority class samples that are harder to learn
- less synthetic data for minority samples that are easier to learn.





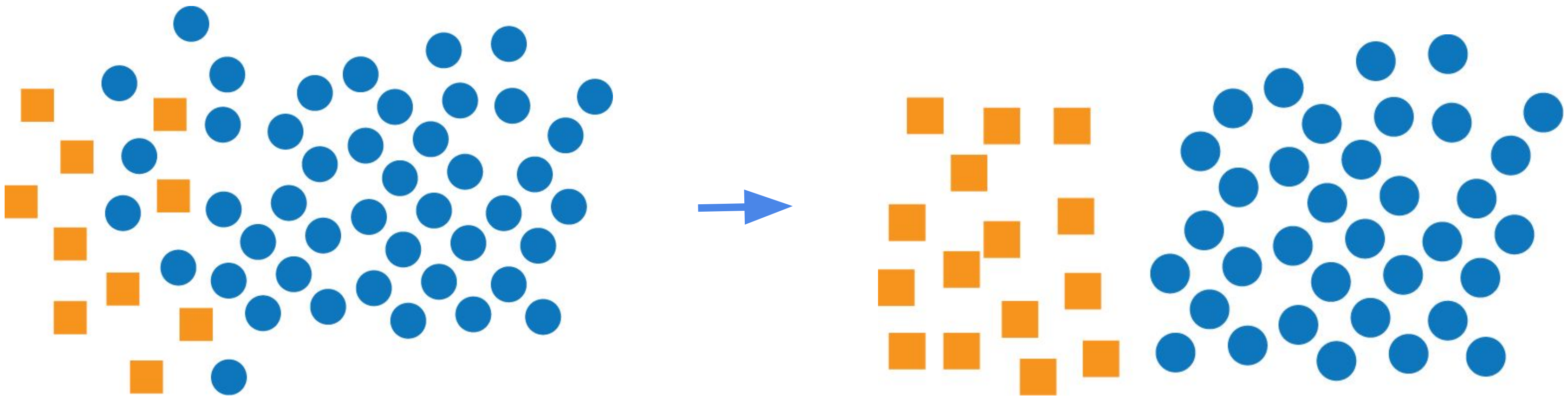
Ensemble

- Ensemble multiple sampling methods
- Ex: SMOTE + Token Links



Ensemble

- Ensemble multiple sampling methods
- Ex: SMOTE + Token Links
- Bagging and Bootstrapping (Ex. Subset-SMOTE)



Imbalanced-learn

- Easy sklearn-like API
- Can be used in sklearn Pipelines
- Supports all major resampling methods

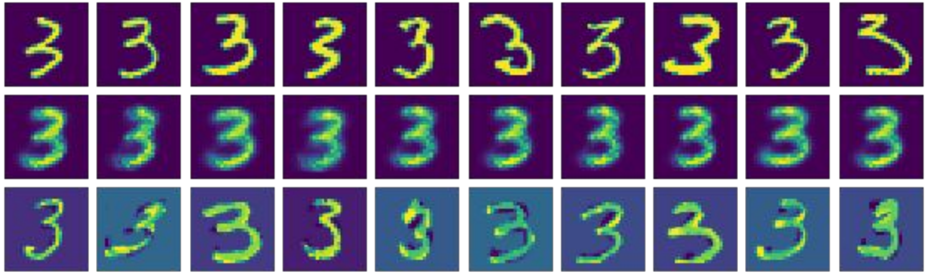
```
from sklearn.svm import LinearSVC
from imblearn.under_sampling import NearMiss
from imblearn.pipeline import make_pipeline


pipeline = make_pipeline(NearMiss(version=2),
    LinearSVC())

pipeline.fit(X_train, y_train)
```

Feature Learning

- Create generative model on the minority class
- If done well, it can outperforms resampling
- Generative model have more parameters to tune.
- May take longer due to training. (Neural Network)

Synthetic Minority Reconstruction Technique (SMRT)

- Train variational autoencoder on the data
- oversampling -> sampling the autoencoder
- Performs very well
- Requires lot of data and training epochs

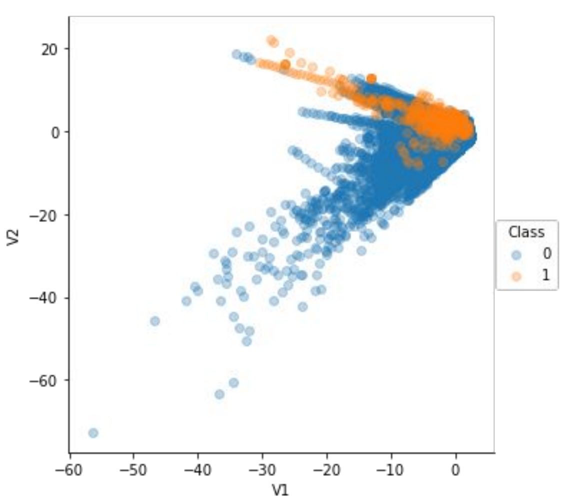


SMRT

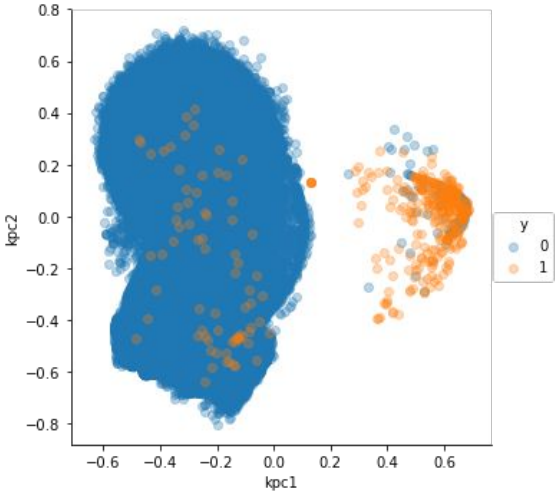
SMOTE

Feature Engineering

- Create Features which help separation of the classes



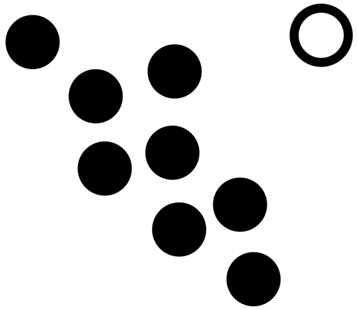
Simple PCA



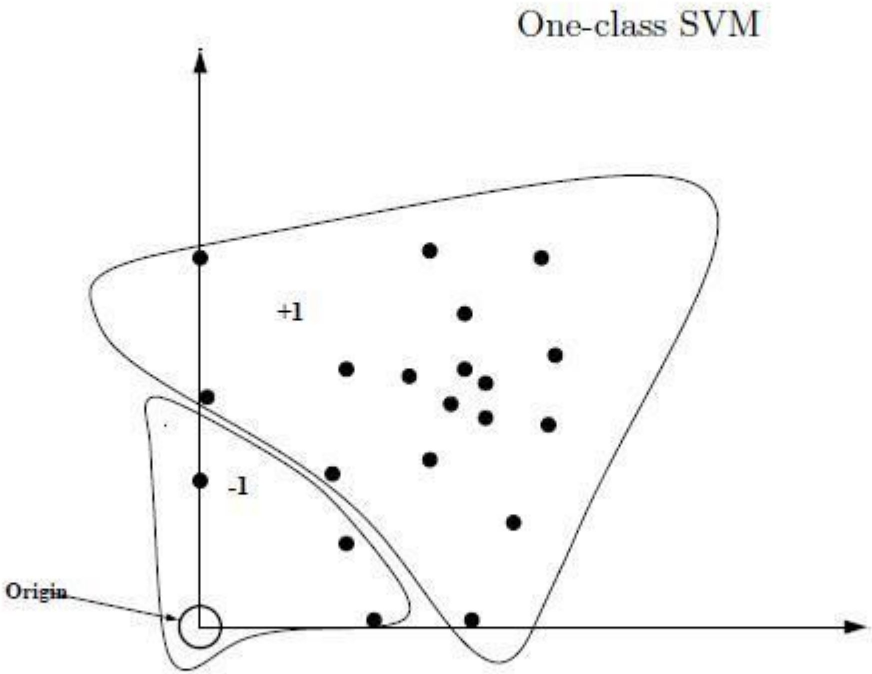
Cosine Kernel PCA

Anomaly Detection

- Treat problem as a anomaly detection problem
- Use ML Algorithms and methods dedicated for such tasks

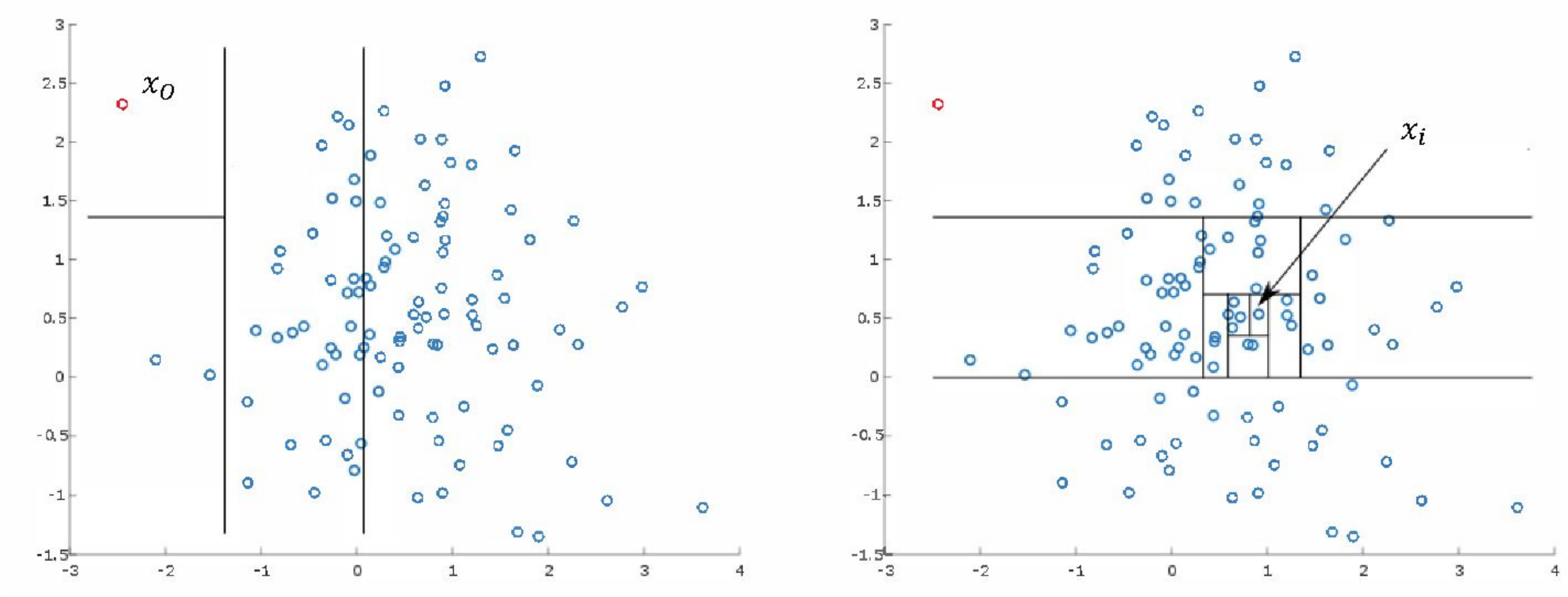


One Class SVM



Source:http://www.geocities.jp/mabonakai/sub/ex_oneCsvm.htm

Isolation Forest



Resources

- SMRT: https://github.com/tgsmith61591/smrt
- Imbalanced-learn: http://imbalanced-learn.org/

Kubat, Miroslav, Robert C. Holte, and Stan Matwin. "Machine learning for the detection of oil spills in satellite radar images." *Machine learning* 30.2-3 (1998): 195-215.

Thank you for your Attention

Adrian Spataru
Data Scientist at Know-Center GmbH
adrian@spataru.at
https://www.fb.me/adrian.spataru.5

DATA SCIENCE CHALLENGE 2018

https://goo.gl/v7wBso

KNOW Center