Feature Engineering

Feature Engineering

- Outlier
- https://scikit-learn.org/stable/modules/feature_selection.html#feature-selection
- PCA / ICA
- Imbalance data
  - Class weight
  - Sample weight
  - SMOTE
  -

Outlier Detection

- Box plot
- IQR
- One Class SVM
- Localoutlier
- Elliptical
- let see code ...

Imbalance Dataset

Imagine you are trying to classify **two groups of people** based on where in a room they are standing.

People with **green shirts** tend to stand in the south part of the room, but there's a lot of them (let's say **1,000**—it's a big room), so they fill up the south wall all the way to the middle of the room.

People with **yellow shirts** tend to stand in the middle and towards the northwest corner of the room, but there's few of them **(let's say 5).**

The regions where these two groups of people prefer to stand overlap, so there are a couple yellow shirted folks who have green shirted neighbors around them.

Now if we train our classifier on this data, it will simply tell us everyone is wearing green shirts, and it will be right 99.5% of the time. Not useful.

Imbalance Dataset

- Using class wt in SVM
- Using sample wt in SVM

- Let see code..

Imbalance Dataset

In this case, you can try resampling the data, either by **under-sampling your majority class or over-sampling your minority class** .

Over-sampling consists of either sampling each member of the minority class with replacement, or creating synthetic members by randomly sampling from the feature set.

This is what SMOTE—Synthetic Minority Over-sampling Technique—does.

conda install -c conda-forge imbalanced-learn