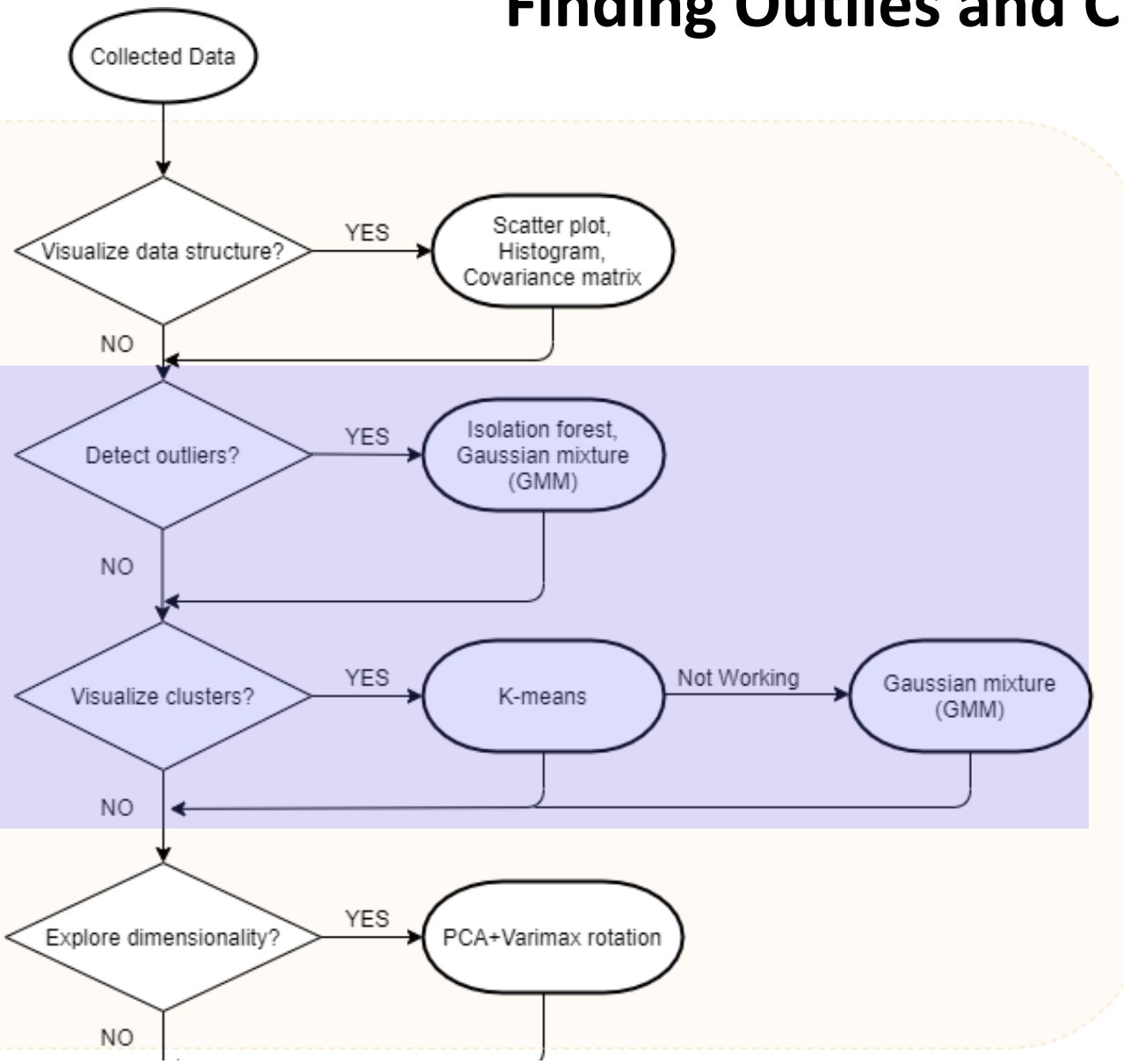
Clustering and Outlier Analysis

Himaghna Bhattacharjee

"To be or not to be..."

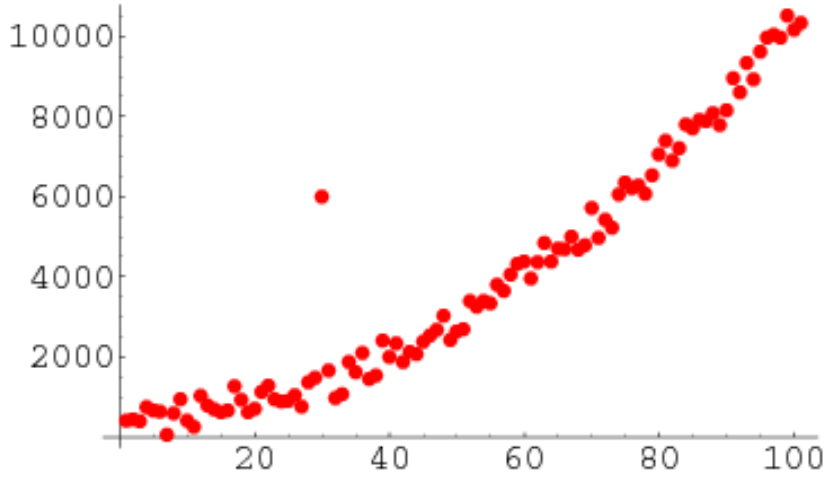
Rapid Advancement in Process Intensification Deployment

Finding Outlies and Clustering Data



Downstream ML

Outliers



- Outliers are points in the input space which are “far away” from the other points
  - Errors in measurement
  - Regions of input space with scarce data

How are they found?

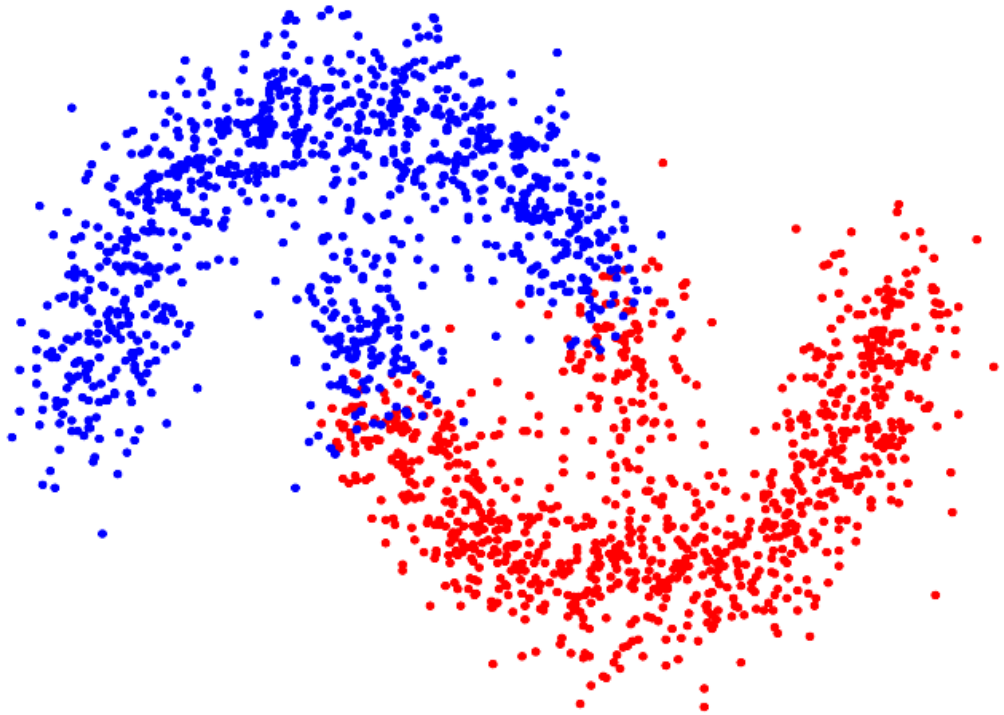
- Points easiest to separate from others

Isolation Forest

- Points in low probability density regions of input space

Gaussian Mixture Models

Clustering

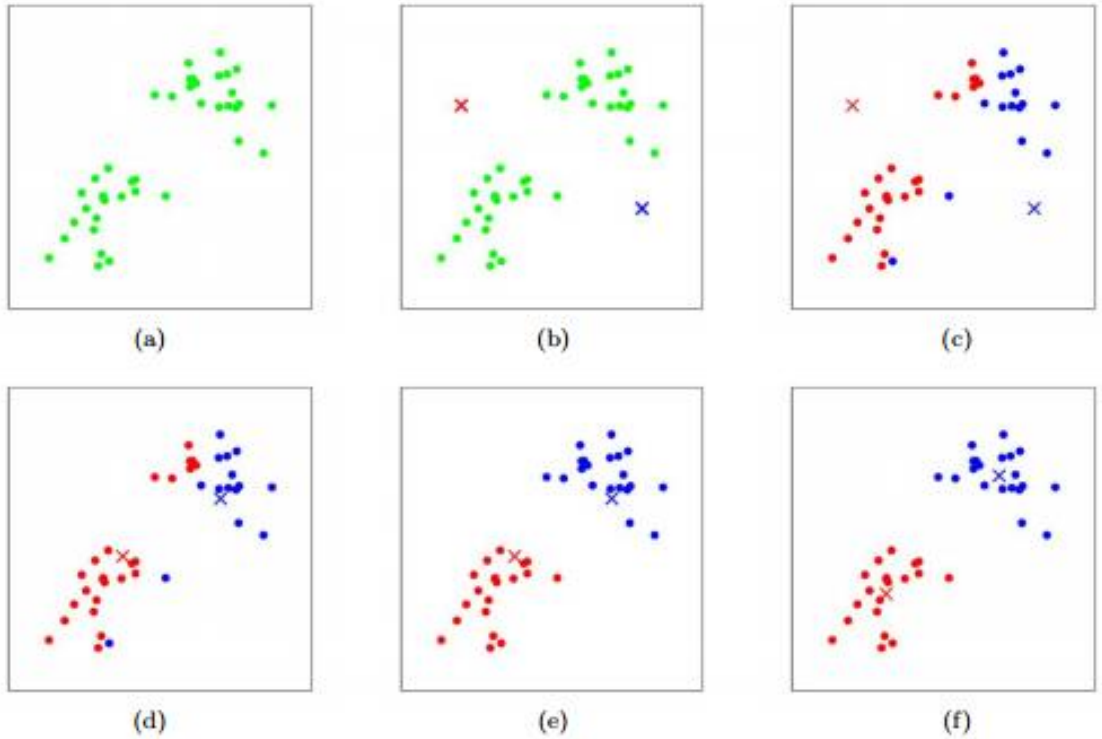


- Clustering is the unsupervised technique where we find clusters of similar data in our data set
- This may lead to insights such as identifying outliers and also making separate models for separate clusters

Examples

Identify different types of catalysts in your data set based on binding energy, crystal structure, heats of formation etc.

K-Means

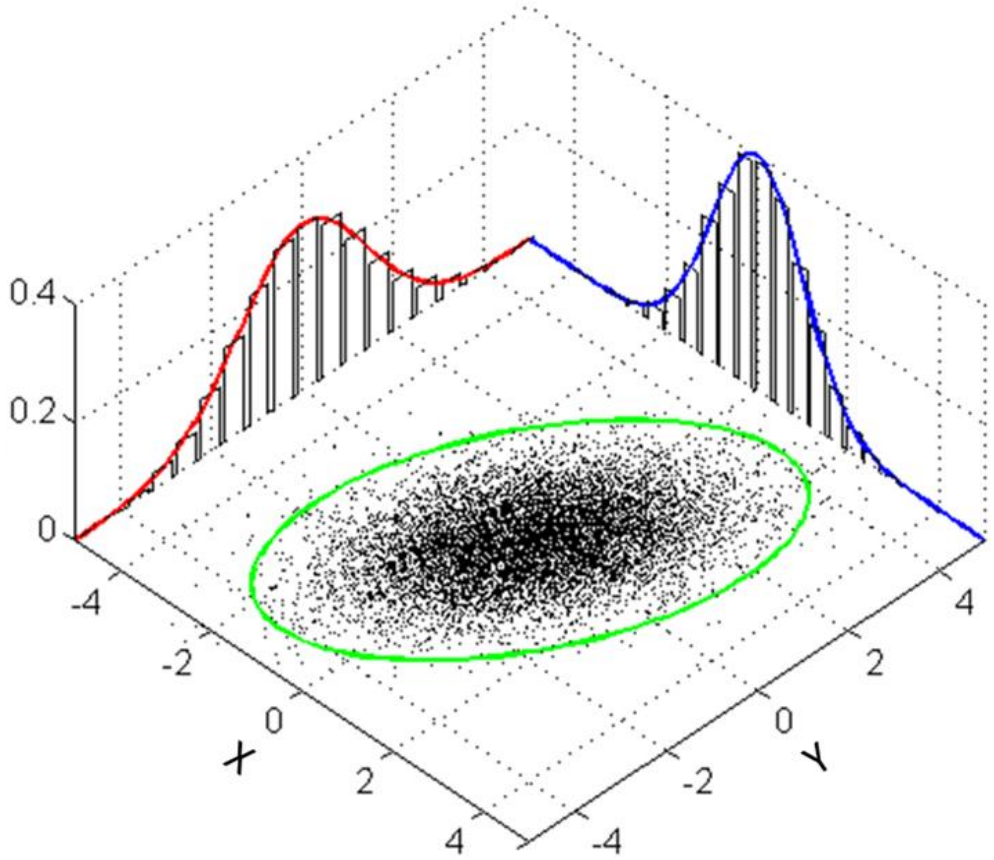


- Iterative classification
  - Set cluster centers randomly
  - Assign points to cluster with closest center
  - Check the centroid (mean) of all the points in cluster
  - If different from current cluster center, shift cluster center to mean
  - Re-assign points

K-Means does a good job of finding clusters which are roughly 'spherical' in shape

Gaussian Mixture Model (GMM)

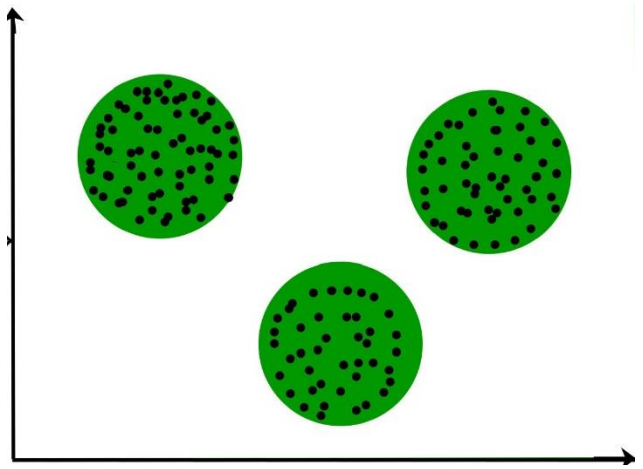
- Assumes that data is generated from several cluster parametrized by normal distributions
- Estimates the mean and std. dev of these normal distributions from data



K-Means vs GMM

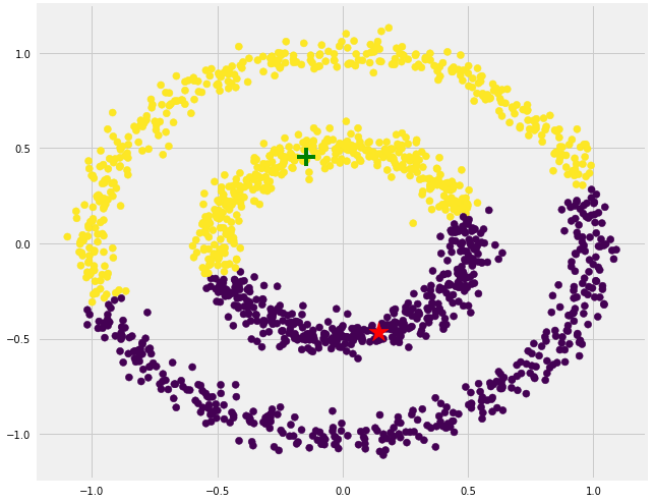
K – Means

- Computationally inexpensive
- Restricted to spherical clusters
- No probabilistic interpretation (hard clustering)



GMM

- Expensive convergence
- Arbitrary cluster shapes
- Probabilistic (soft clustering)



Measures and Metrics

Mean distance nearest cluster

Silhouette Coefficient: $\frac{b - a}{\max(a,b)}$ ← Mean distance to other points in same cluster

1: Point is well classified
0: Point is at cluster edge
-1: Point is misclassified

Bayesian Information Criterion: Probabilistic criterion that penalizes model complexity and ill-fit