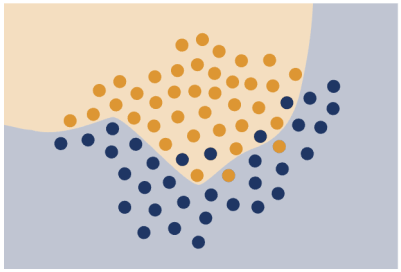
Problem Classes

Problem classes

Supervised Learning
(with labels)

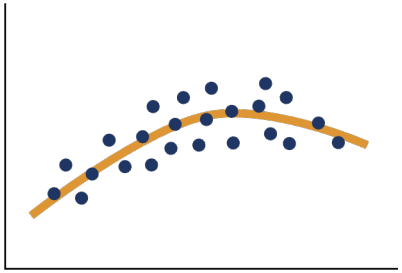
Unsupervised Learning
(without labels)

Classification



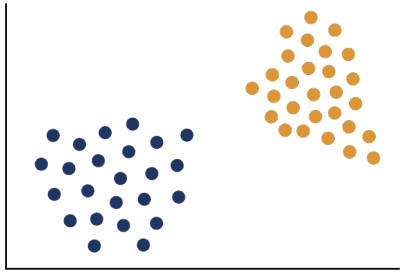
- Stability
- Reaction mechanisms
- ...

Regression



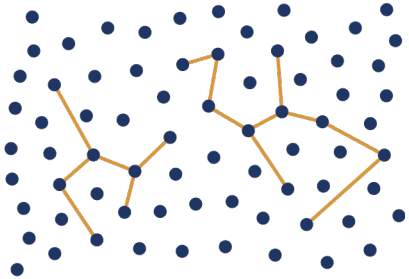
- Reaction barriers
- Geometries
- ...

Clustering



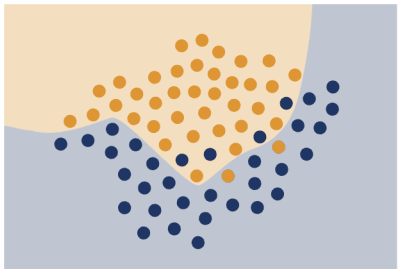
- Dimensionality reduction
- Find mechanisms
- ...

Association



- Find mechanisms
- Detect networks
- ...

Classification



Challenges

- Imbalanced frequencies
- Irrelevant features
- Overlapping classes
- Non-linear data
- High-dimensional data

Approaches

- One vs All: n classifiers
- One vs One: n*(n-1) classifiers

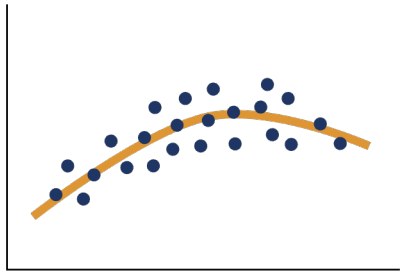
Common algorithms

- Decision trees / Random forest
- K-nearest neighbours
- Neural networks

Regression

Challenges

- Outliers
- Multicollinearity
- Non-normalised features
- Heteroscedasticity



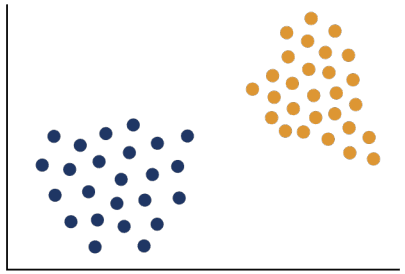
Approaches

- Regularisation
- Bootstrapping

Common algorithms

- Support vector machines
- Gaussian process regression
- Neural networks

Clustering



Challenges

- Outliers
- High-dimensional data
- Imbalanced data
- Non-globular clusters
- Large data sets

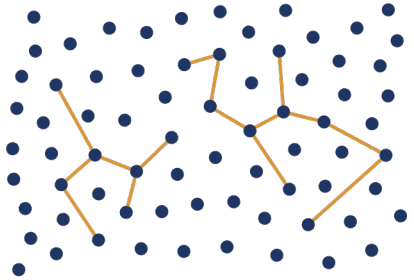
Approaches

- Projection to lower dimensions
- Iterative improvement

Common algorithms

- K-means
- Density-Based Spatial Clustering of Applications with Noise

Association



Challenges

- Rare/sparse interactions
- High-dimensional data
- Imbalanced data
- Large data sets

Approaches

- Restrictions on rule formulation

Common algorithms

- Apriori
- FP-Growth

Summary Problem Classes

- Data pre-processing crucial
- Most common classes are classification and regression
- Feature choices drive limiting capabilities of models
- Wide range of methods for each class
- Detailed understanding of methods required to choose
- No one-fits-all way

Learning Workflow