DATA SCIENCE WORKFLOW

Data Science Workflow
Overview

Data Exploration
- Exploratory analysis
- Visualization

Modelling
- Train an appropriate model
- Tune hyperparameters

Deployment
- Implement your model in a production system

Data pre-processing
- Feature (predictor) selection
- Outlier treatment
- Dimensionality reduction
- Model-specific processing

Validation
- Performance metrics
- Cross-validation
- Test model on an independent dataset

Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011

Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science
https://archive.ics.uci.edu/ml/index.php

Exploratory Analysis
Descriptive Statistics

Overall

- Missing values
  - Maximal share
  - Replacement
    - Average
    - Zero
    - Advanced
- Statistical tests for assumptions (for regression)

Quantitative

	alcohol
count	178.000000
mean	13.000618
std	0.811827
min	11.030000
25%	12.362500
50%	13.050000
75%	13.677500
max	14.830000

Std vs. Min/Max

Plausible Ranges

Median vs. Mean

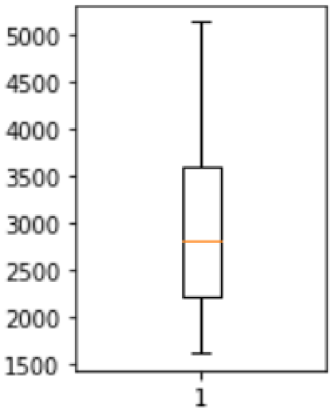
Qualitative

cylinders	
3	4
4	199
5	3
6	83
8	103

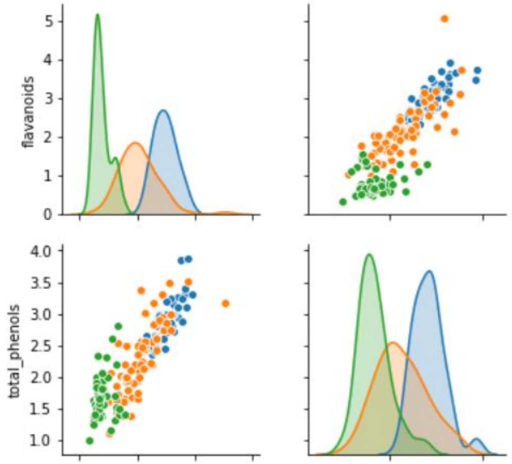
Rare Categories

Exploratory Analysis
Visualization

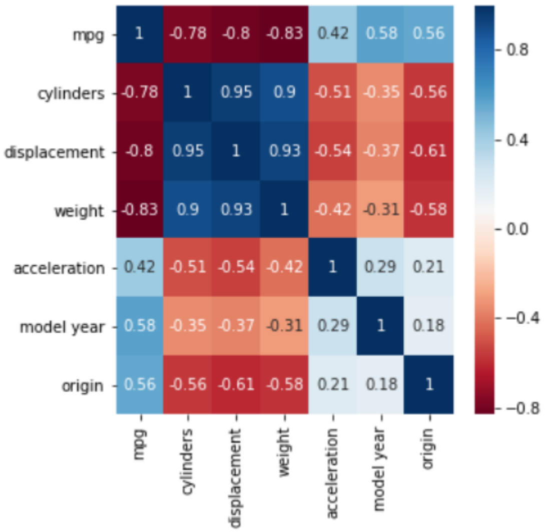
Box-Plot



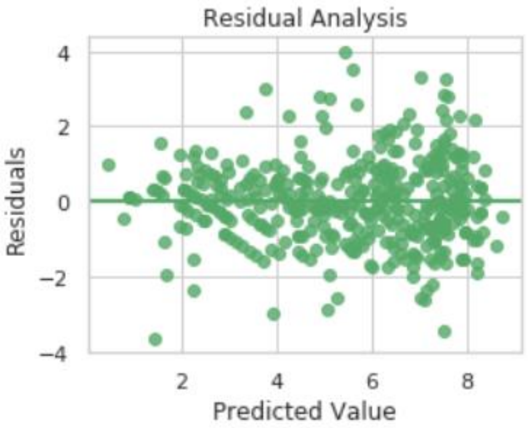
Scatter Plots



Correlation Table



Residual Analysis



FEATURE SELECTION

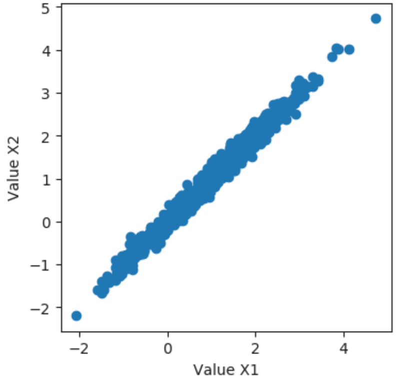
Feature Selection

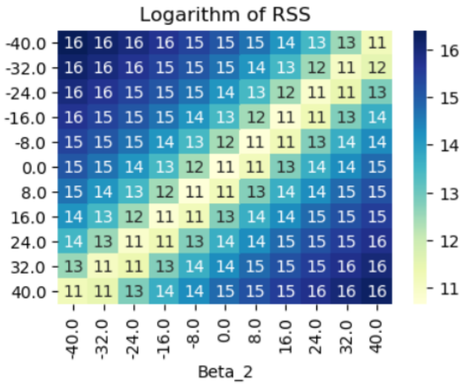
Data Pre-processing – Motivation

Dimensionality Reduction

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 499 entries, 0 to 498
Data columns (total 24 columns):
Q_INCA         499 non-null float64
nEng           499 non-null float64
pRail          499 non-null int64
p_verlauf_1    499 non-null float64
p_verlauf_2    499 non-null float64
p_verlauf_3    499 non-null float64
p_verlauf_4    499 non-null float64
p_verlauf_5    499 non-null float64
p_verlauf_6    499 non-null float64
p_verlauf_7    499 non-null float64
p_verlauf_8    499 non-null float64
p_verlauf_9    499 non-null float64
p_verlauf_10   499 non-null float64
p_verlauf_11   499 non-null float64
p_verlauf_12   499 non-null float64
p_verlauf_13   499 non-null float64
p_verlauf_14   499 non-null float64
p_verlauf_15   499 non-null float64
p_verlauf_16   499 non-null float64
p_verlauf_17   499 non-null float64
p_verlauf_18   499 non-null float64
p_verlauf_19   499 non-null float64
p_verlauf_20   499 non-null float64
phiMI          499 non-null float64
dtypes: float64(23), int64(1)
memory usage: 93.6 KB
```

Correlated Variables





Model Quality

Number of Data

Trade-Off

Number of Predictors

- Training time
- Overfitting

Interpretation (Feature Importance)

Feature	Feature Importance
proline	0.250901
color_intensity	0.117607
alcohol	0.096448
flavanoids	0.088399
total_phenols	0.085046
hue	0.083909
malic_acid	0.062438
od_280_315	0.058619
proanthocyanins	0.039775
alcalinity_of_ash	0.036927
magnesium	0.036225

Feature Selection

Feature Selection vs. Feature Engineering

Feature Selection

- Given the current set, choose a certain number due to a threshold

Feature 1
Feature 2
Feature 3

↓ Selection Criterion

Feature 2
Feature 3

- For example: best subset selection

Feature Engineering

- Creation of new features based on existing ones

Feature 1
Feature 2
Feature 3

↓

Feature 1+2/3

- For example: principal component analysis

DIMENSIONALITY REDUCTION

Dimensionality Reduction
Principal Components Analysis – General Idea

- Idea:
  - Replace a set of correlated variables by a set of fewer orthogonal ones
- How (intuitively):
  - Define a new orthogonal coordinate system through the mean where the first axis captures the most variance

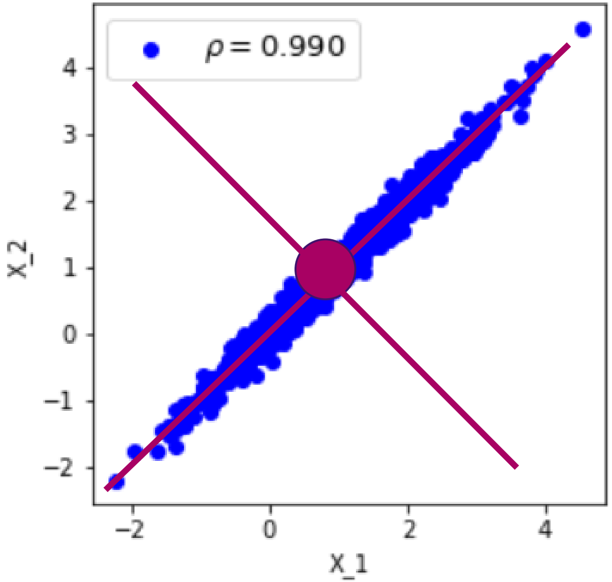
Define new orthogonal system

↓

Keep “meaningful” axes

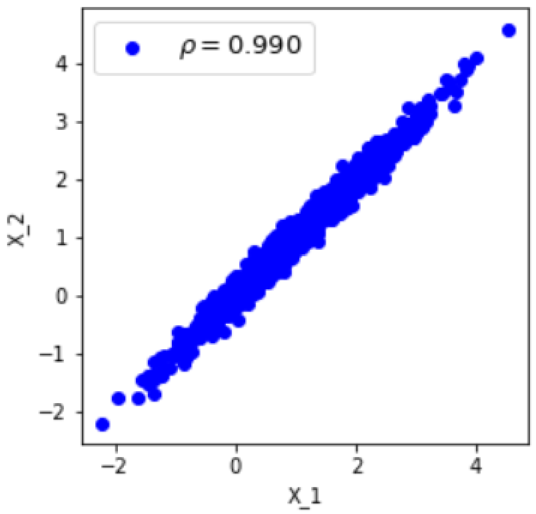
↓

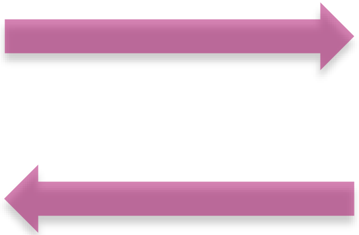
Define new coordinate values

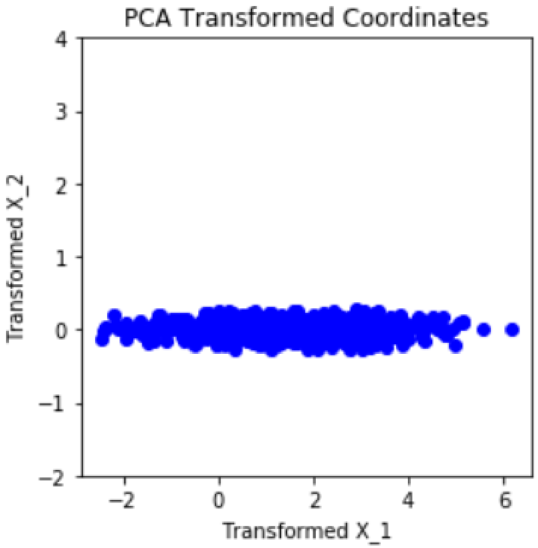


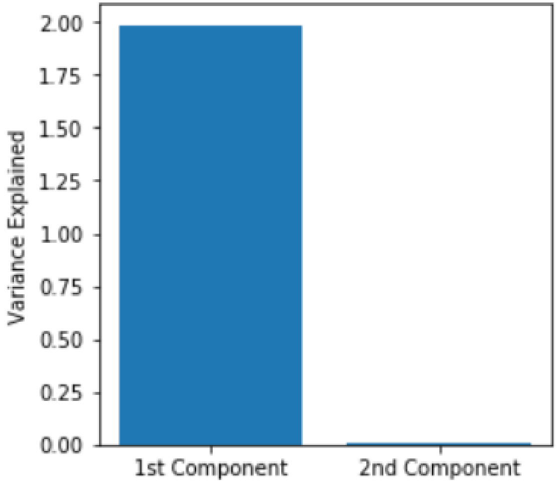
Dimensionality Reduction
Transformation of the Input Data – Meaningful Case

- Data transformation achieved by multiplying the data by the matrix of eigenvectors



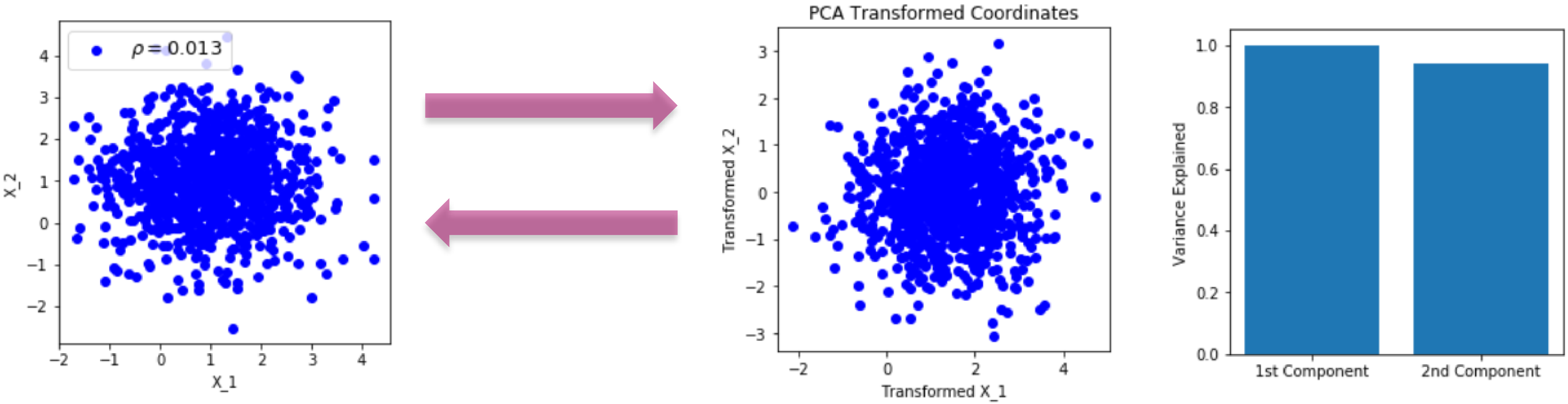






Eigenvalue - amount of the variance explained by eigenvector

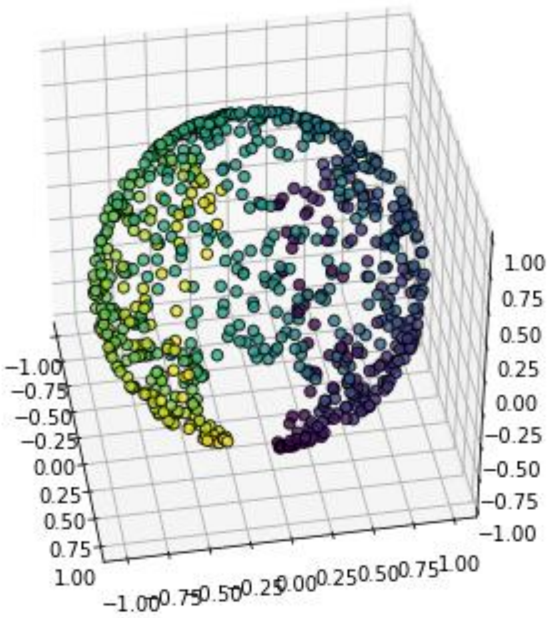
Dimensionality Reduction
Transformation of the Input Data – Non-Meaningful Case

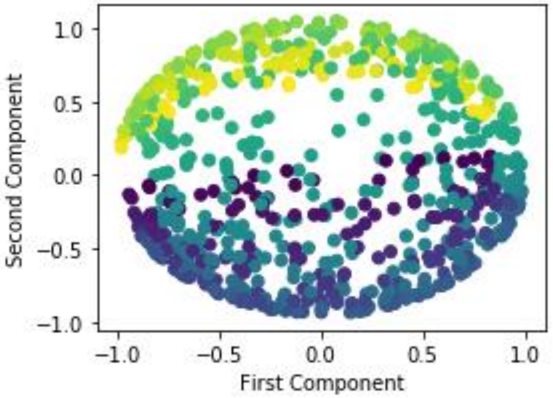


BOSCH

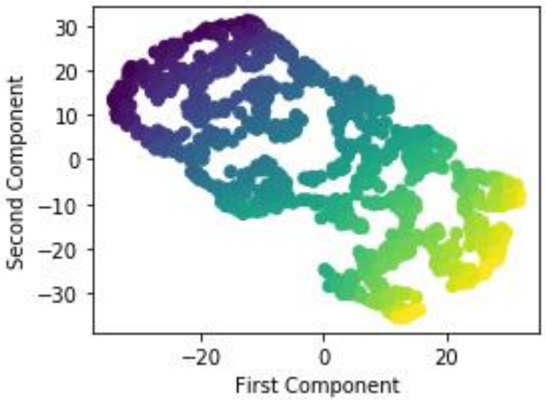
Dimensionality Reduction
Non-Linear Dimensionality Reduction – t-SNE Example

- T-SNE captures the local structure of the data better given the same number of dimensions





PCA is not able to capture local high-dim dependencies



t-SNE puts points together, which are locally close

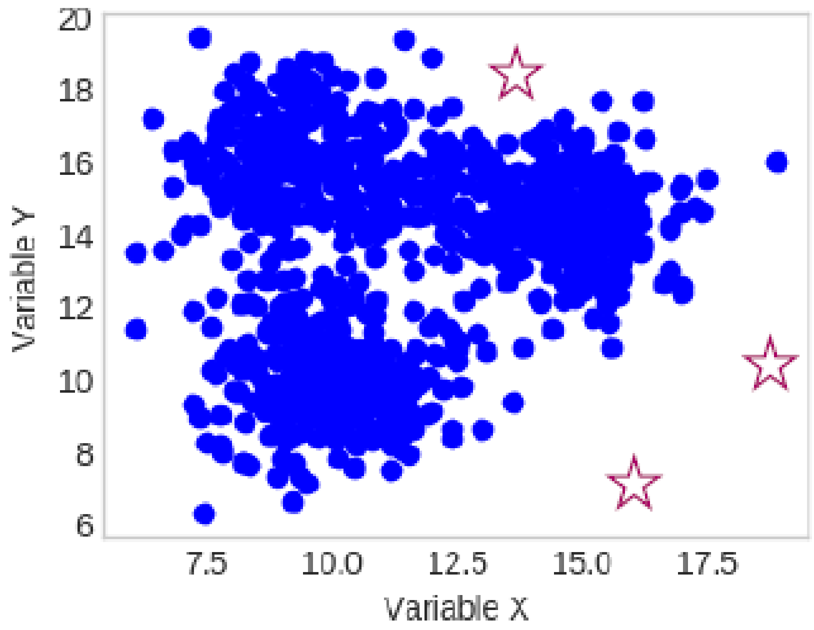
BOSCH

OUTLIER DETECTION

Outlier Detection
General Remarks and Application Areas

- Outlier (anomaly)
  - Significant deviation from the majority of points
  - Probably, generated by another process
  - Outlier vs. noise

- Application areas
  - Data pre-processing
  - Fraud detection/security
  - Medical care
  - System diagnosis



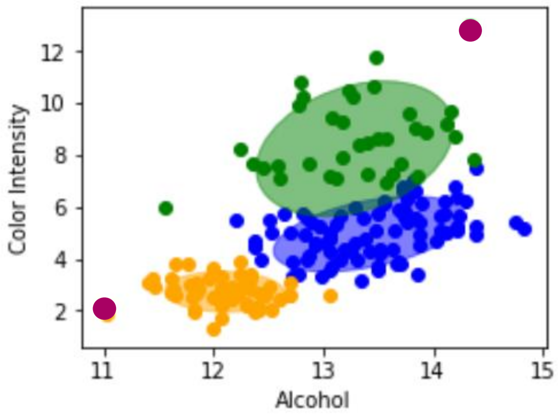
Aggarwal, Charu C. "Outlier analysis." Data mining. Springer, Cham, 2015

Link to Book Description

BOSCH

Outlier Detection
Categories of Outliers

Point outlier/anomaly



When is a point considered an outlier (metrics)?

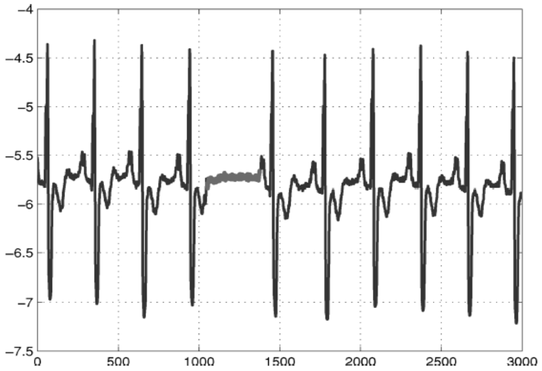
Contextual outlier

Contextual attributes
-28° in Spain on the 1st of July

Behavioural attributes
15° in Spain on the 1st of July

Adding a context, e.g. using an additional dimension

Collective outlier



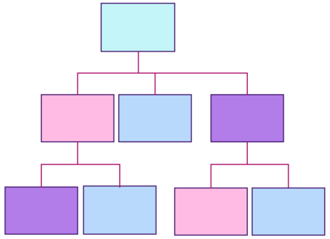
Chandola et. Al. 2009

Graph data
Time series

Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." ACM computing surveys (CSUR) 41.3 (2009): 15

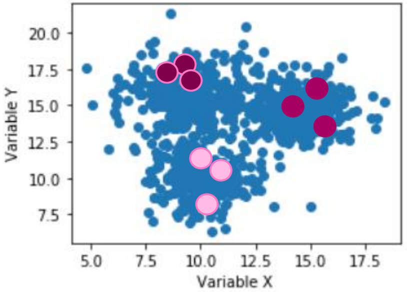
Outlier Detection
Methodical Overview

Supervised Methods



- Any classification method
- Lack of labelled outlier data
- Imbalanced class distribution

Semi-Supervised Methods



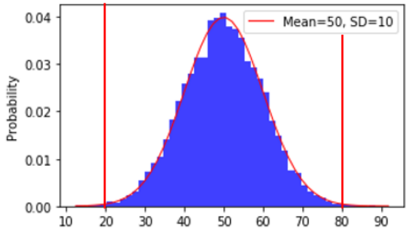
- Also called novelty detection
- Semi-supervised clustering

See Chapter 6.3-6.4 of Aggarwal, Charu C. "Outlier analysis."

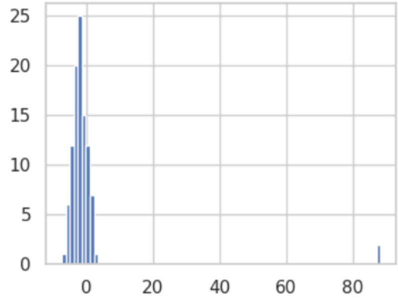
Unsupervised Methods

Statistical

- Parametric (model)

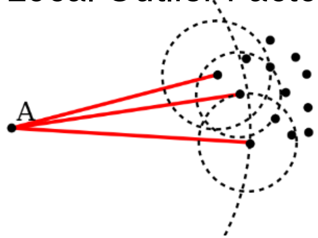


- Non-parametric

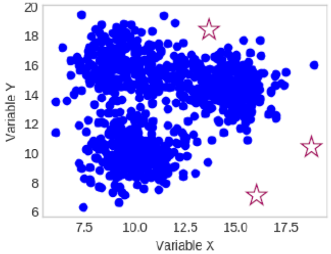


Proximity-based

- Local Outlier Factor



Clustering-based



MODELING – CLUSTERING

Clustering
Motivation - Typical Clustering Applications

Data Analysis

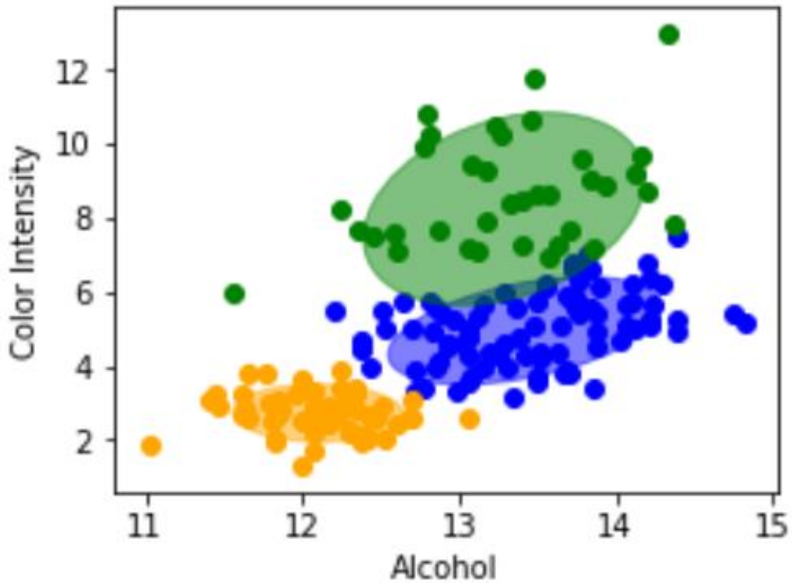
- Presence of data structure and distribution

Knowledge Discovery

- Finding and interpreting patterns and dependencies

Data Segmentation

- Partitioning and profiling of the segments



Outlier Detection

- Filtering for “abnormal” data
- Anomaly detection

Data Reduction/Pre-Processing

- Replacing high-dimensional data by representatives

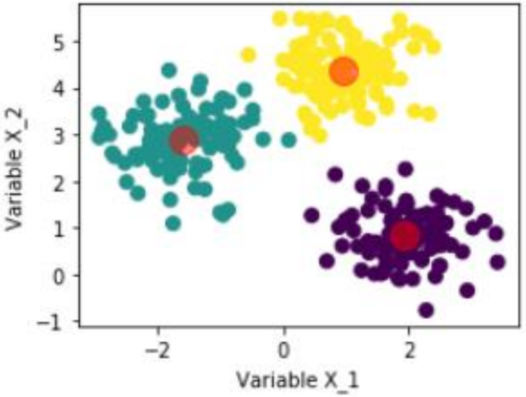
Classification

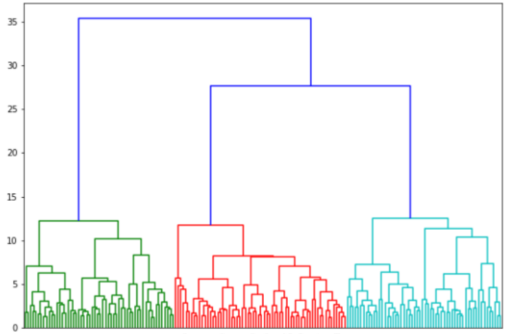
- Mapping to known groups (clusters)

Clustering

Overview of Clustering Approaches

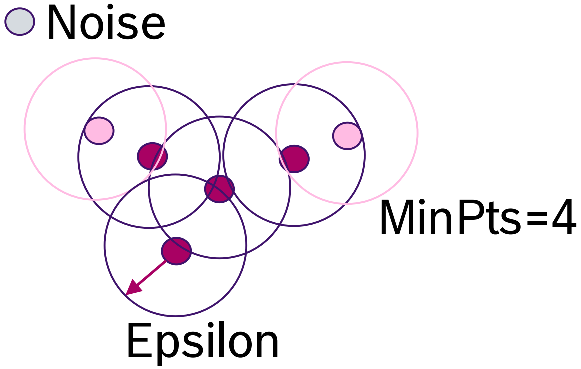
Partitioning Methods

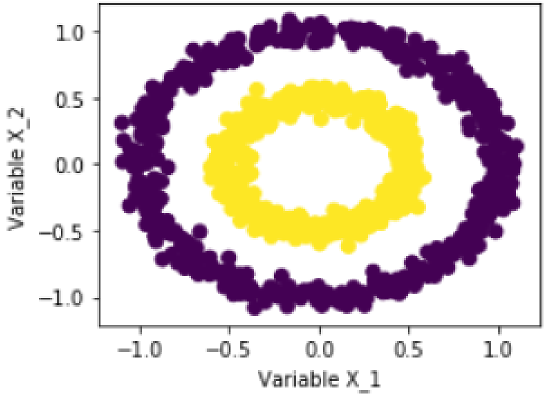




Hierarchical Methods

Density-Based Methods





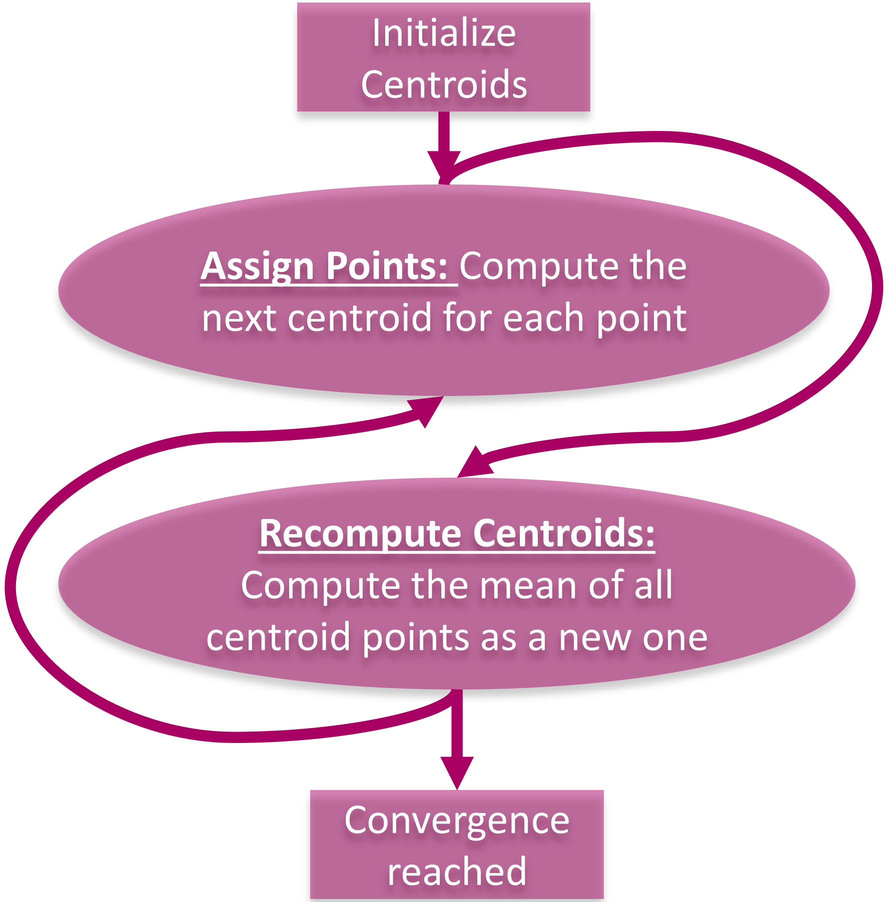
Advanced Methods

Clustering
K-Means - General Idea

- Distance-based representation clustering
- Result: set of centroids

- Advantages
  - Simple (understanding and implementation)
  - Relatively scalable

- Disadvantages
  - Sensitive to outliers
  - Is not guaranteed to converge



Clustering
K-Means - Algorithm

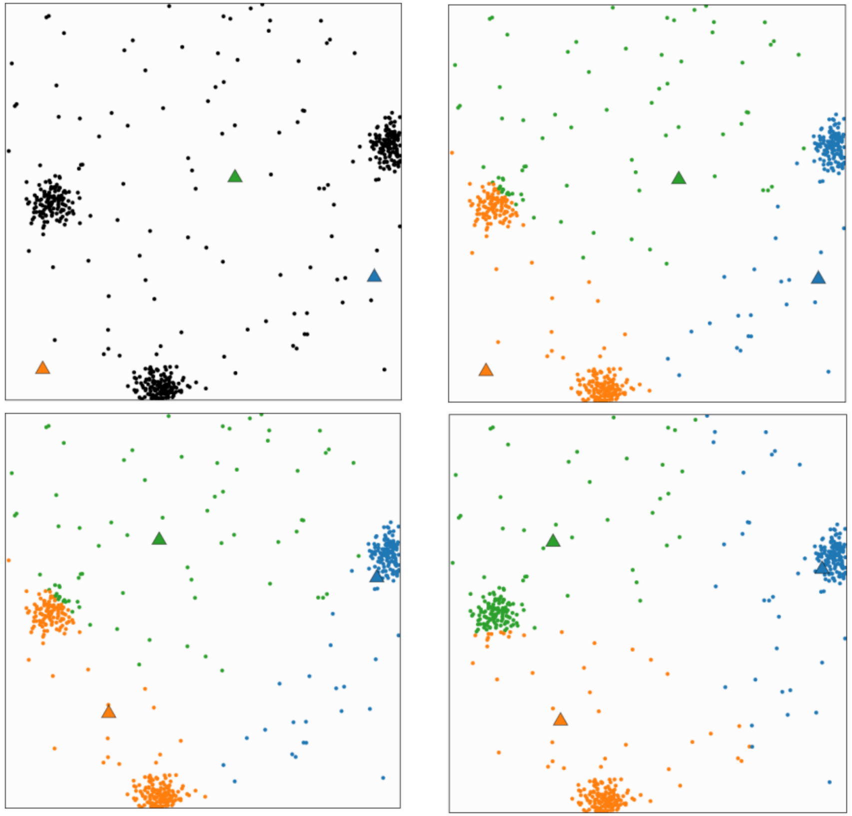
- **Input:** *N* data points, *K* centroids (cluster centers)
- **Initialize:** several possibilities
  - random centroids
  - *K* points from data
- **Iterate**
  - **Assign** each point from *N* to the closest centroid in *K*

$$k = \arg\min_k ||x_n - c_k||^2$$

  - **Recompute** new centroids based on the new assignments

$$c_k = \frac{1}{|\mathcal{C}_k|} \sum_{n \in \mathcal{C}_k} x_n$$

  - **Repeat** until convergence



Karanveer Mohan, MIT License

http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

BOSCH

Clustering

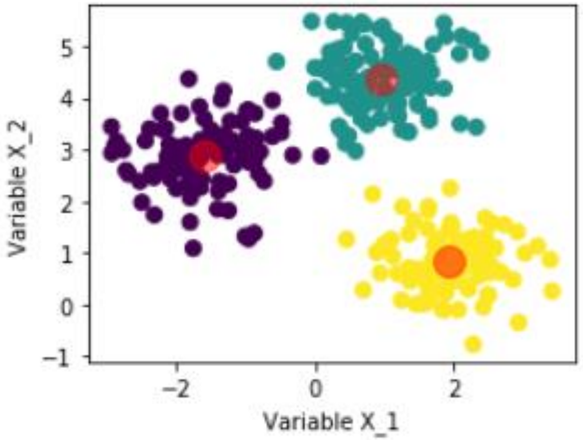
K-Means – Evaluation Criteria – Elbow Method

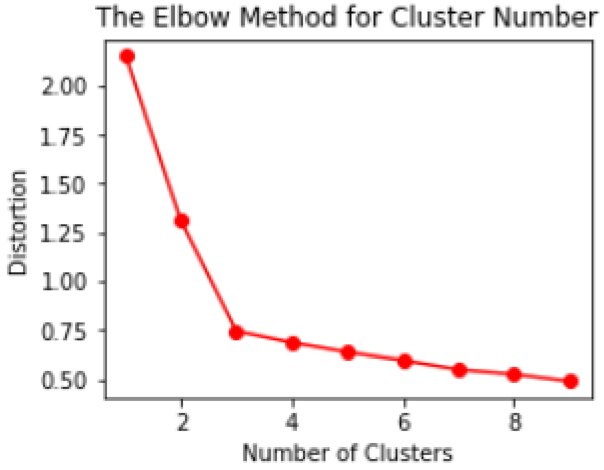
- A good clustering is given by
  - High within-cluster similarity
  - Low inter-cluster similarity

- Optimization criteria – minimize squared sum of distances to centroids (distortion)

$$J(c, r_{nk}) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - c_k||^2$$

Binary variable If a point belongs to a cluster

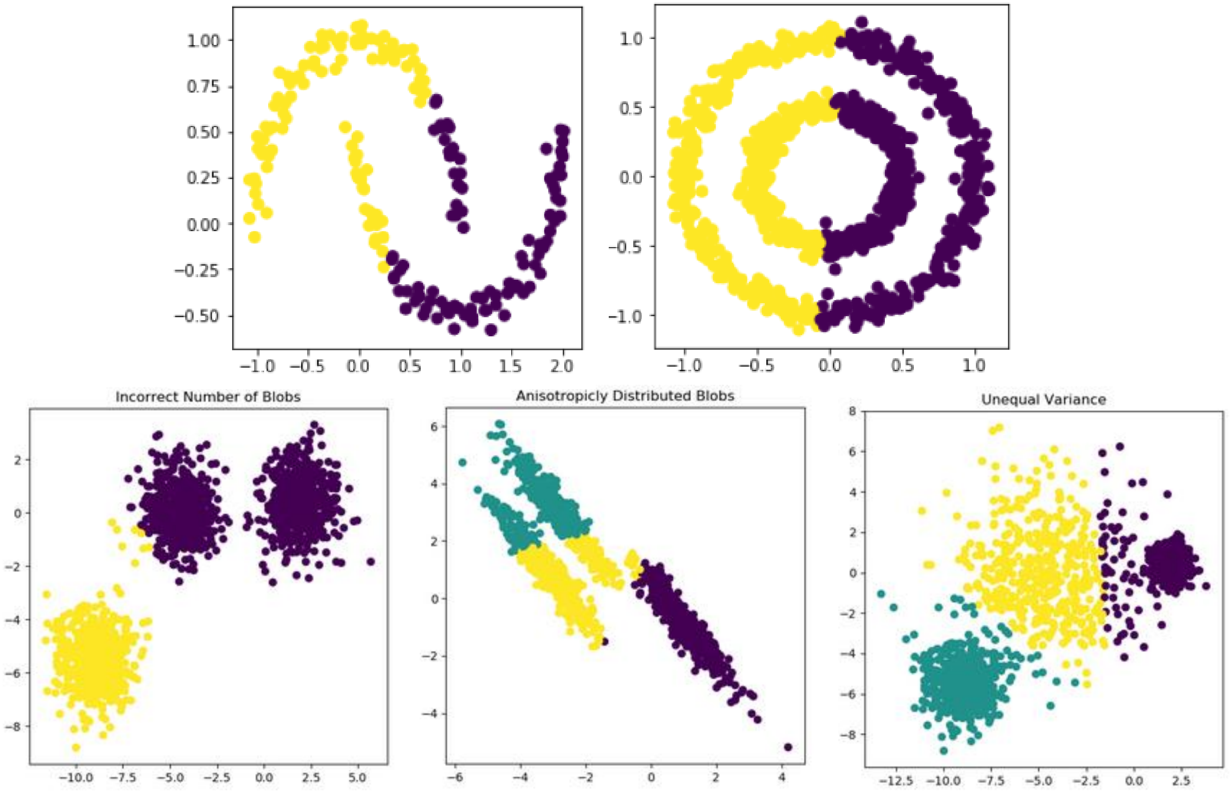




Clustering
K-Means – What Can Go Wrong

- Issues to address
  - Centroid initialization
  - Outlier
  - Normalization/standardization
  - Similarity measures (categorical data)
  - Number of clusters

- Know your data!

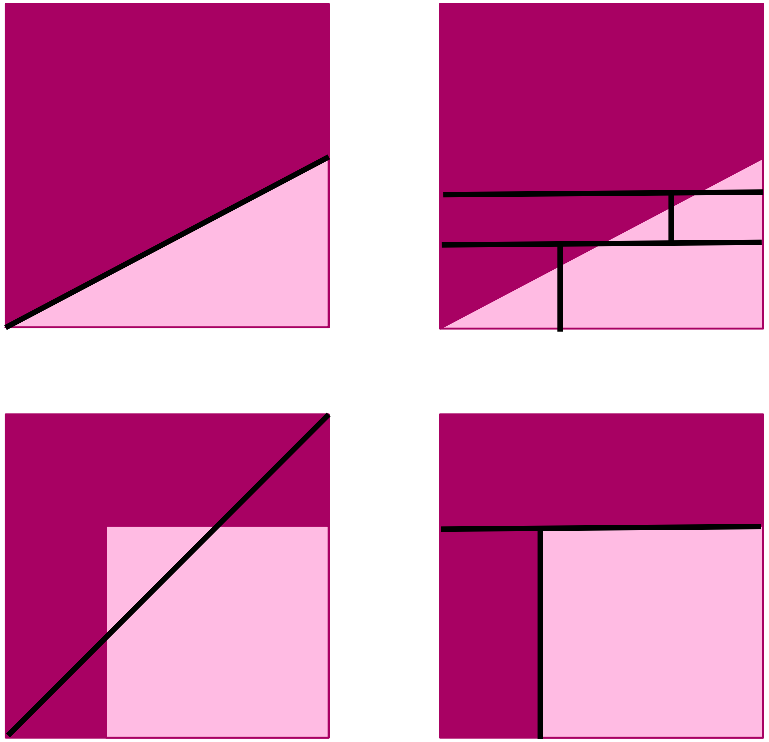


[Link to the Source with Python Code]

MODELING – CLASSIFICATION- DECISION TREES

Decision Trees
Linear Classifier vs. Decision Tree

- Different “cutting” technique

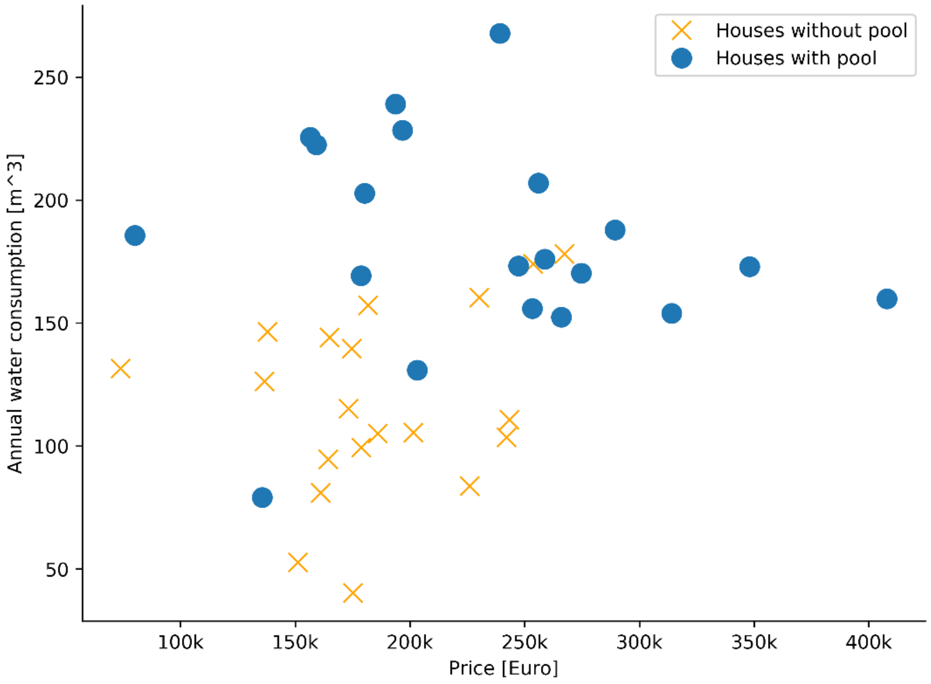


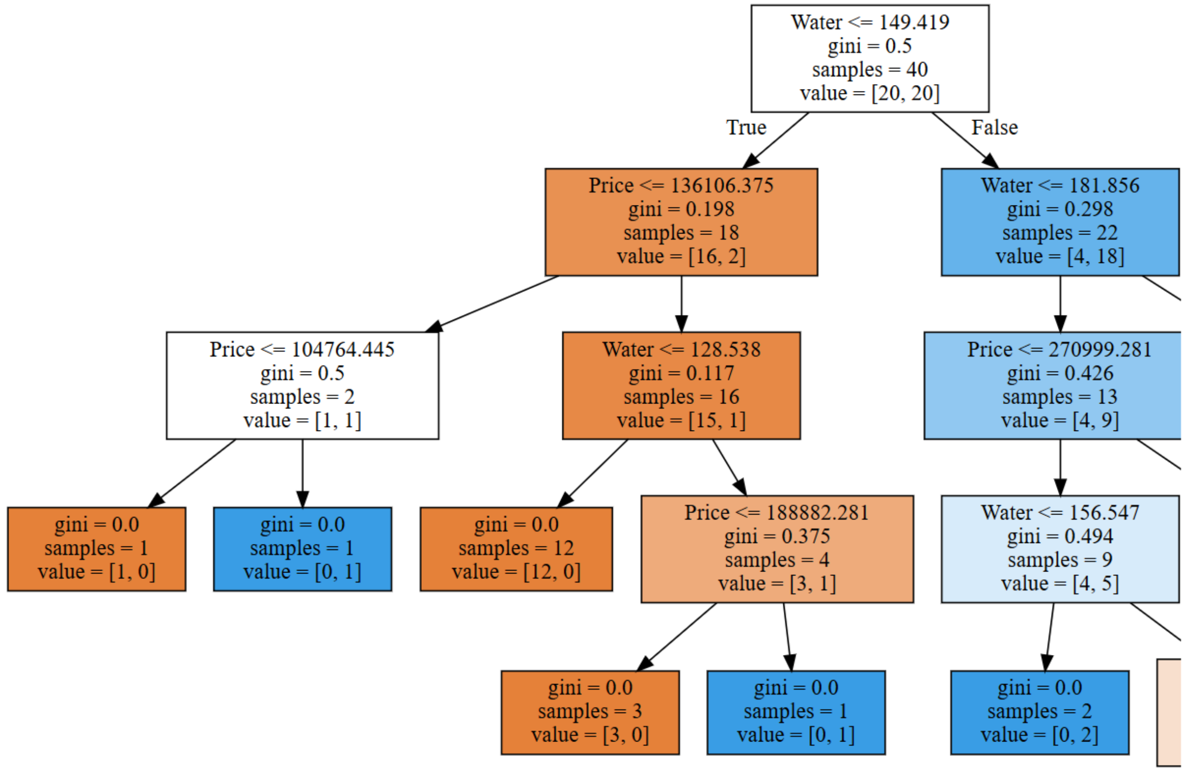
AI-Enabling CR/PJ-AI-E | March 2019

BOSCH

Decision Trees

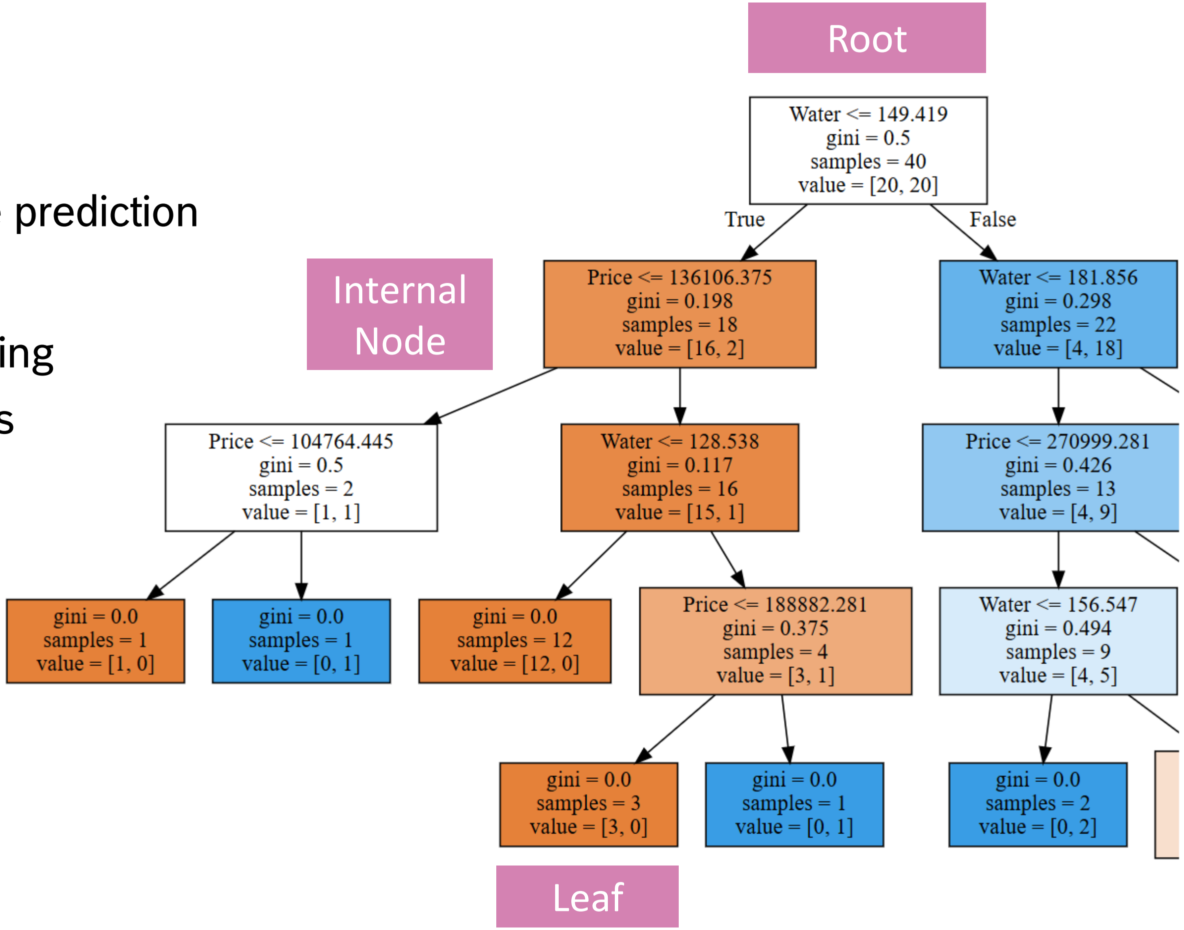
Basic Idea – Tree Visualization



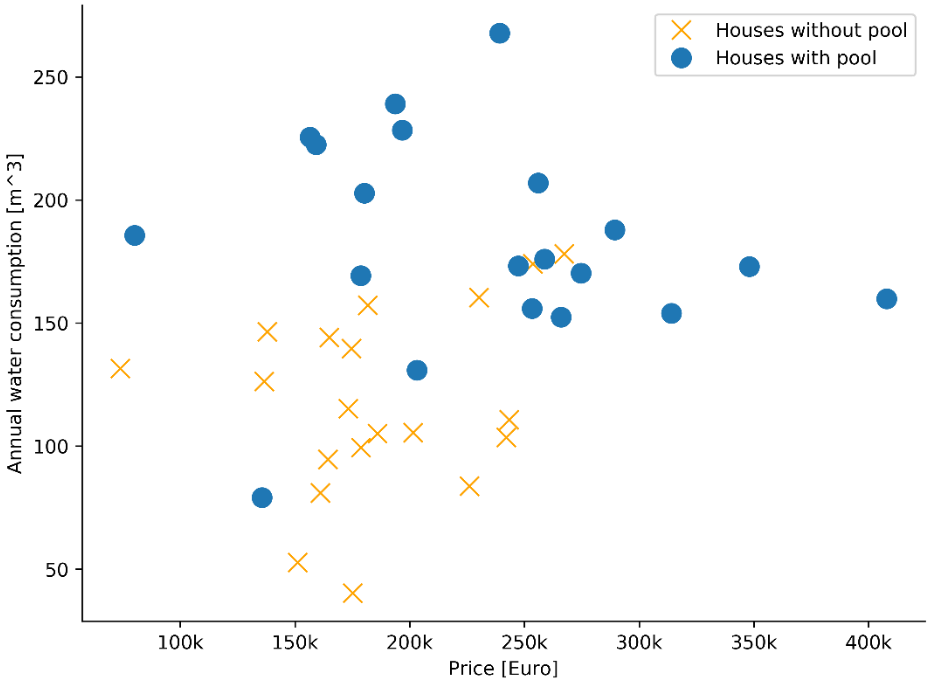


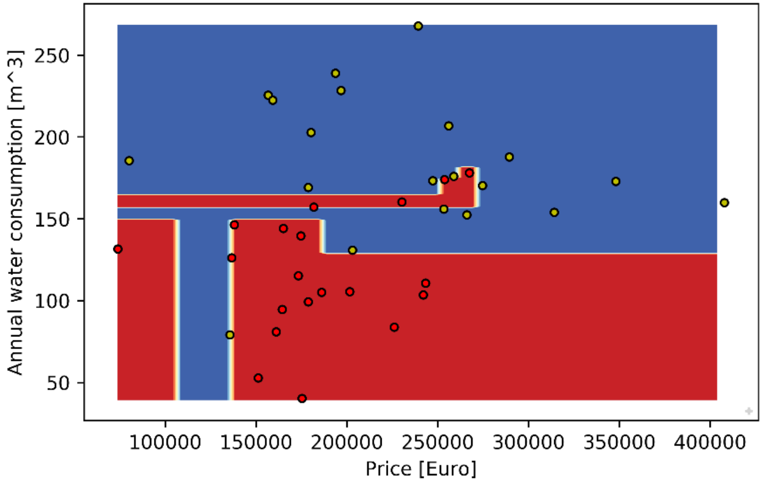
Decision Trees
General Remarks

- Greedy rule-based prediction mechanism:
  - Divide predictor space into regions
  - For each observation of a region make the same prediction
- Pro:
  - Interpretability, as close to peoples decision making
  - Handles categorical variables and missing values
- Contra:
  - Not the most accurate
  - Sensitive to variation in data – high variance



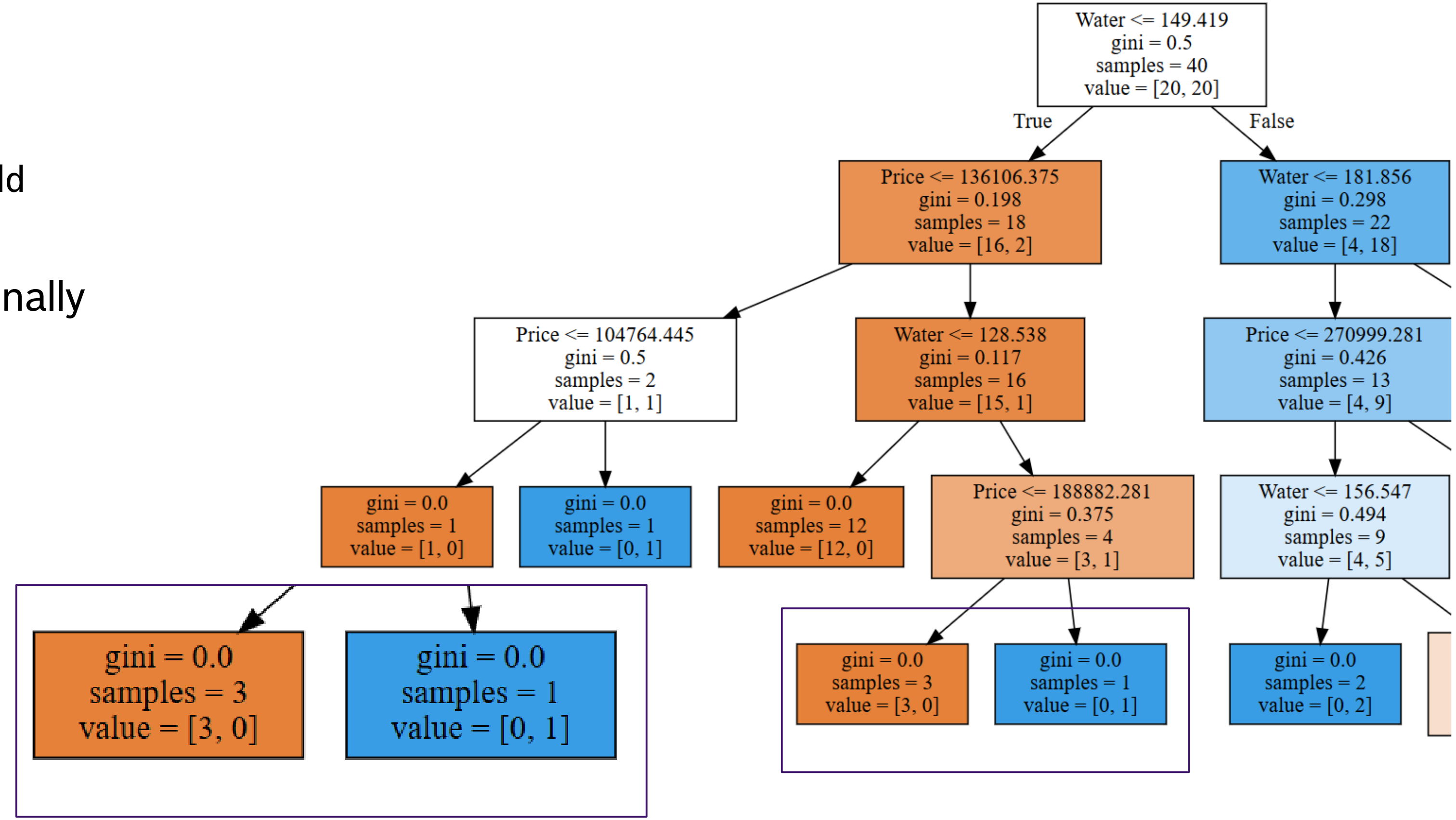
Decision Trees
Basic Idea – Tree Visualization



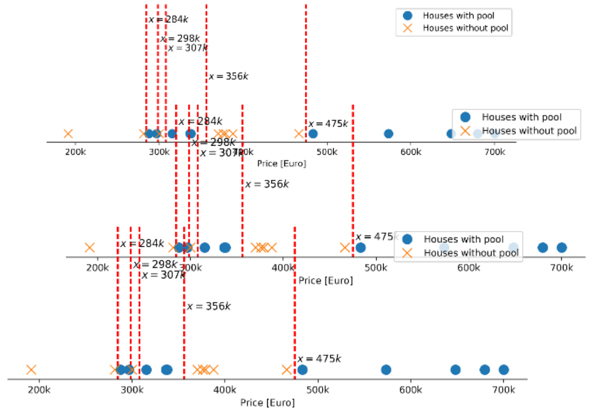


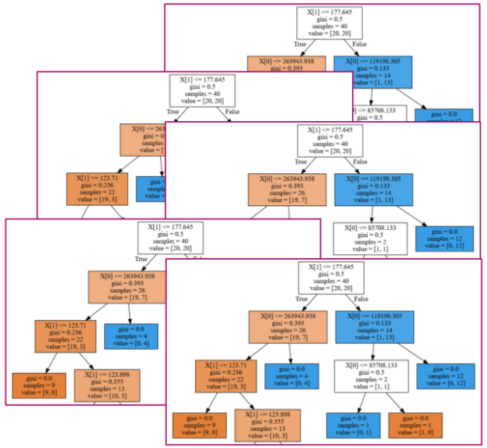
Decision Trees
Tree Pruning Against Overfitting

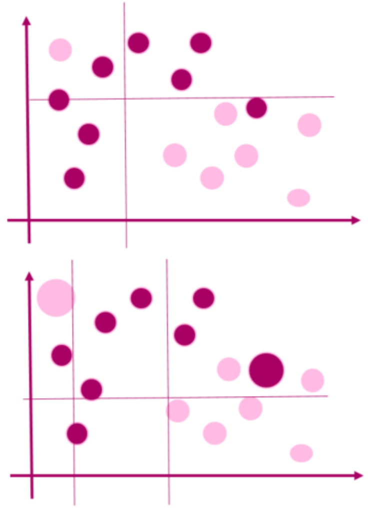
- Decision trees are likely to overfit
- Two strategies
  - Prepruning - fit until
    - Reduction in error above threshold
    - Minimal number of leaves
  - Postpruning – more computationally intensive



Decision Trees
Bagging – Random Forests -Boosting







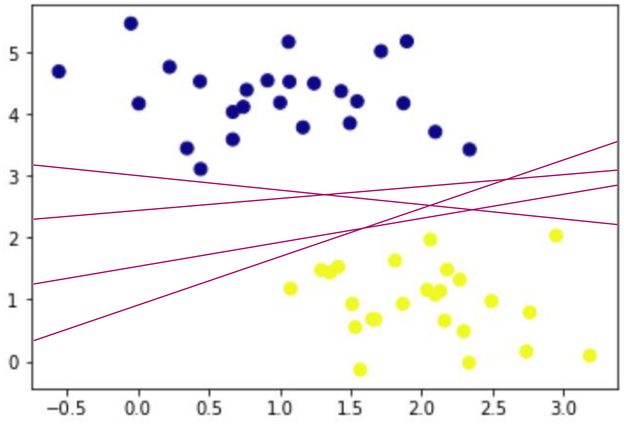
$$F(x) = \sum_m \gamma_m f_m(x)$$

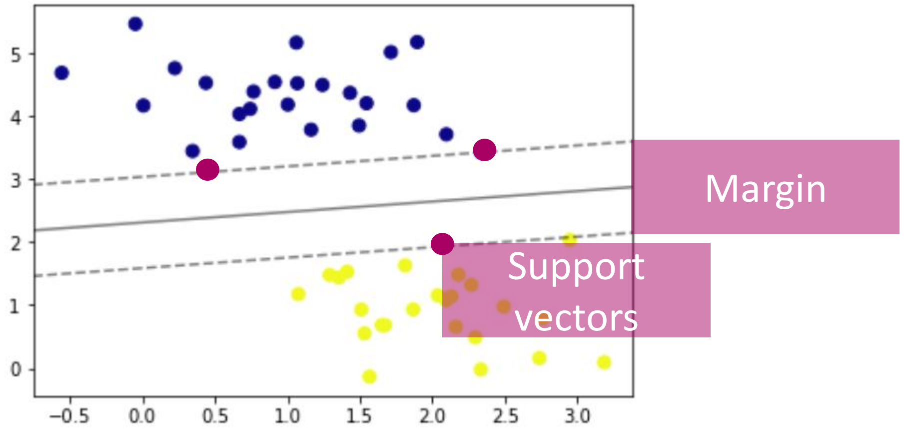
- Bagging - reduces the variance component of error using average of complex models
- Random forest – bagging with sampling of data and predictors
- Boosting – reduces bias with simple models and re-weighting of errors

MODELING-CLASSIFICATION-SUPPORT VECTOR MACHINES (SVM)

Support Vector Machines
Example of the Linear Case – Margin and Support Vectors

- Support vectors
  - (Very small) set of training samples
  - Only support vectors define decision function



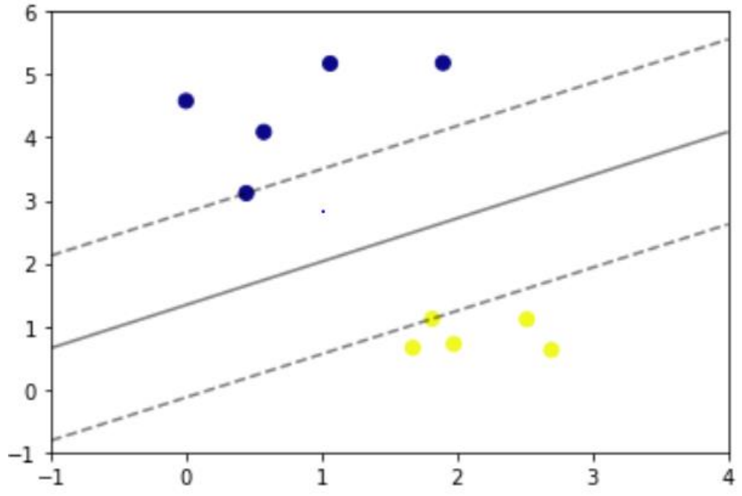


Goal: maximize margin

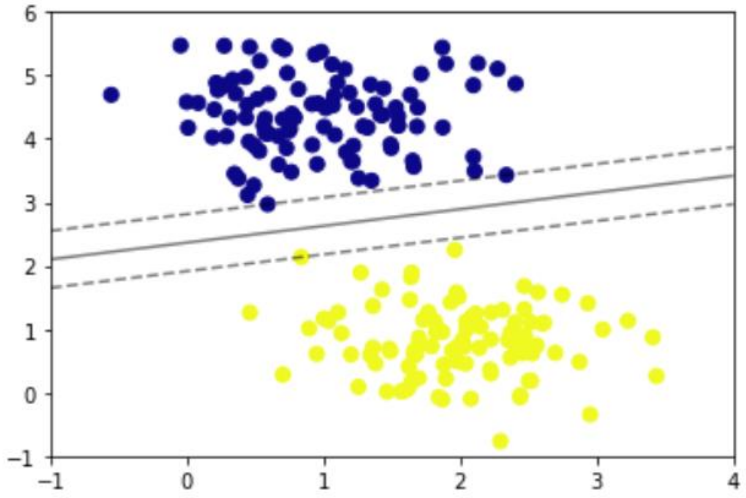
BOSCH

Support Vector Machines
Meaning of Support Vectors

- Support vectors constitute the classification decision boundary
- Sign of overfitting: too many support vectors



10 points – 2 support vectors

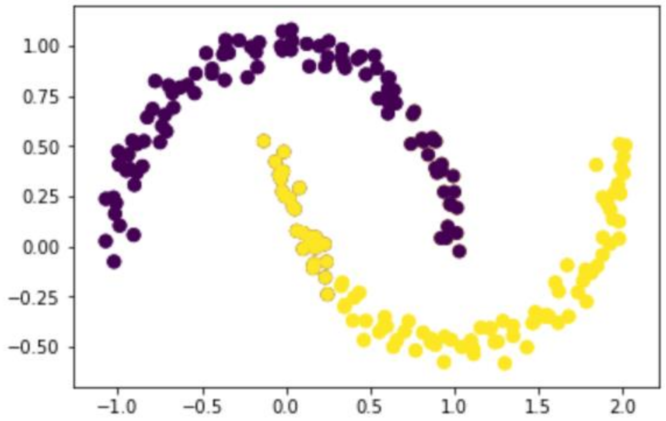


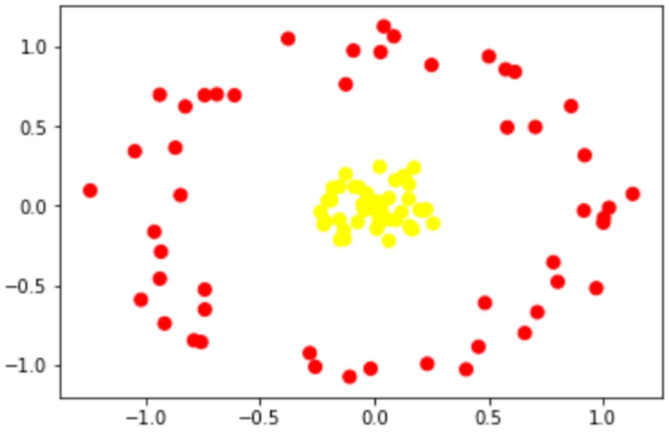
200 points – 3 support vectors

BOSCH

Support Vector Machines
Non-Linear Case

- Problem: some cases are linearly non-separable

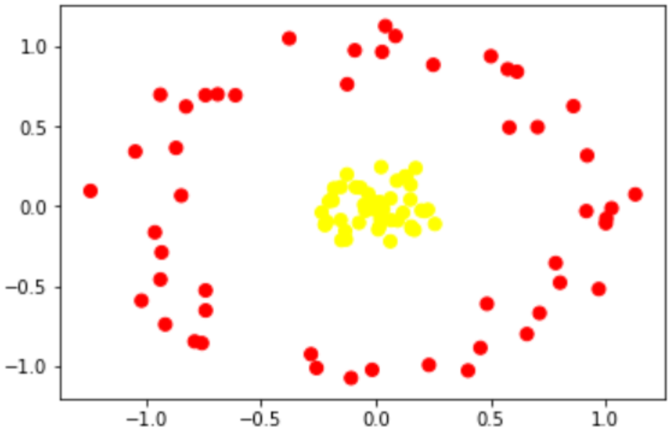




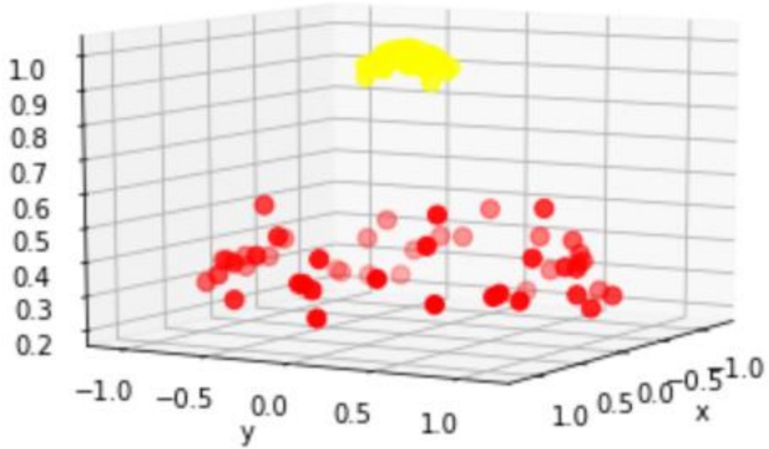
BOSCH

Support Vector Machines
Non-Linear Case

- Idea: Non-linear transformation into an alternative input space
- Result: points are linearly separable



$$k(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2l}\right)$$



Support Vector Machines
Some Typical Kernel Functions

Linear (Dot Product)

$$k(x_i, x_j) = x_i \cdot x_j$$

Polynomial

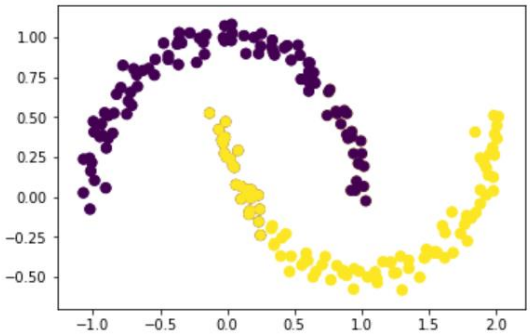
$$k(x_i, x_j) = (x_i \cdot x_j + 1)^p$$

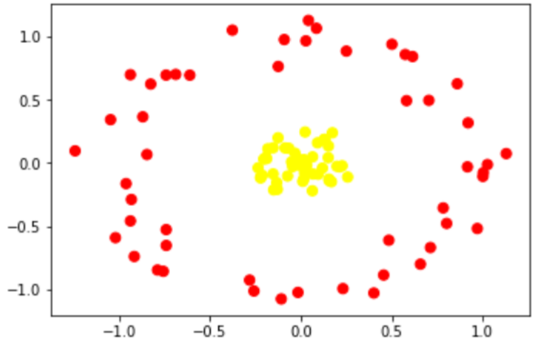
Radial-basis Function

$$k(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2l}\right)$$

Neural Network

$$k(x_i, x_j) = tanh(\beta_0 x_i \cdot x_j + \beta_1)$$





MODEL-SPECIFIC PRE-PROCESSING AND VALIDATION

Model-Specific Pre-processing and Validation
Confusion Matrix and Derived Performance Metrics

True Class

Predicted Class	Class="Yes"	Class="No"
Class="Yes"	True Positive (TP)	False Positive (FP)
Class="No"	False Negative (FN)	True Negative (TN)

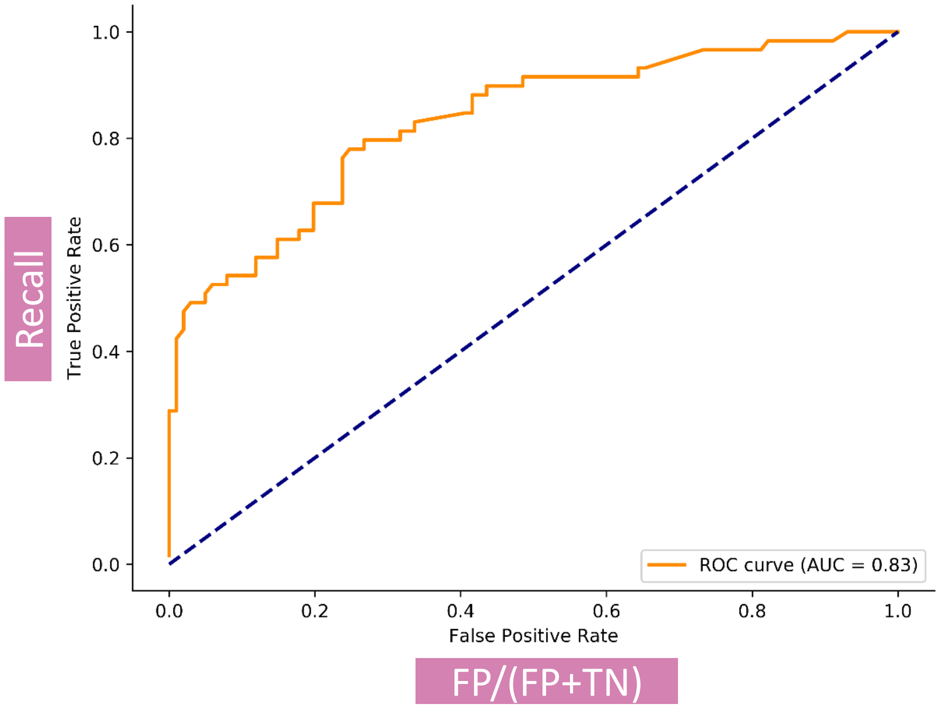
$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

Model-Specific Pre-processing and Validation
Confusion Matrix and ROC Curve

		True Class	
		Class="Yes"	Class="No"
Predicted Class	Class="Yes"	True Positive (TP)	False Positive (FP)
	Class="No"	False Negative (FN)	True Negative (TN)



CR/PJ-AI-E | March 2019

Model-Specific Pre-processing and Validation
Model Selection and Validation

Training Set

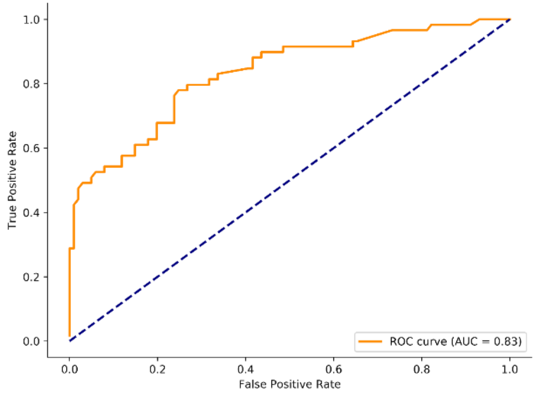
- In-sample training and interpretation

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.693
Model:                            OLS   Adj. R-squared:                  0.692
Method:                 Least Squares   F-statistic:                     878.8
Date:                Thu, 18 Oct 2018   Prob (F-statistic):          6.02e-102
Time:                        08:18:58   Log-Likelihood:                -1130.0
No. Observations:                 392   AIC:                             2264.
Df Residuals:                     390   BIC:                             2272.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     46.2165      0.799     57.867      0.000      44.646      47.787
weight        -0.0076      0.000    -29.645      0.000      -0.008      -0.007
==============================================================================
```

How well is the model learning data by heart?

Validation Set

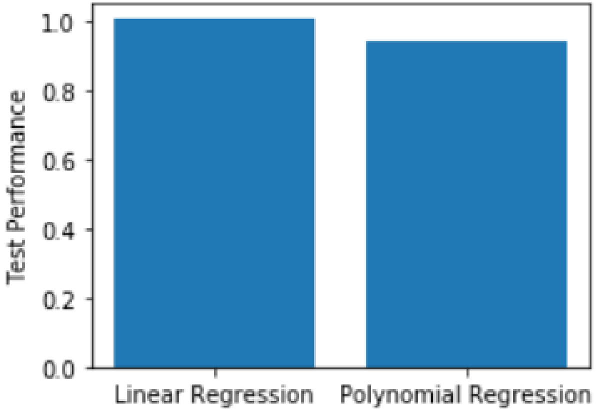
- Cross-validation



Which model type and hyperparameters are best?

Test Set

- Model assessment



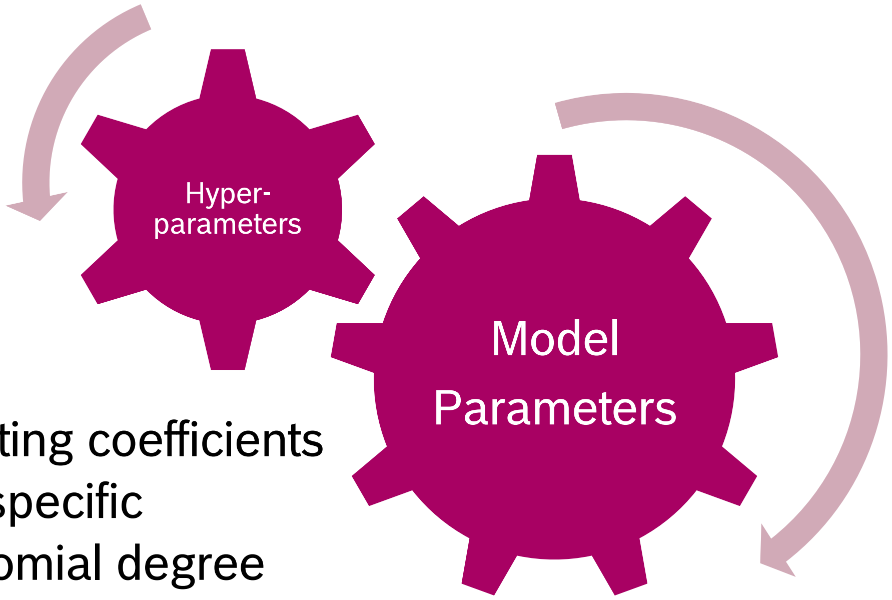
How well do models perform on unseen data?

BOSCH

Model-Specific Pre-processing and Validation
Hyperparameter Tuning

Example: Regression

- Polynomial degree



- Resulting coefficients for a specific polynomial degree

Purpose of the Validation Set – Find Best Performing Parameters

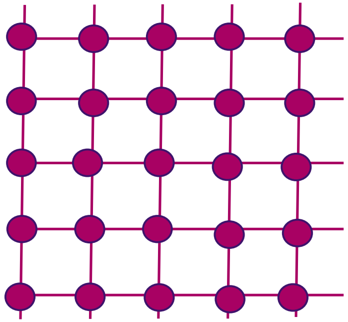
- Typical strategies see on the next slide
- Advanced strategies
  - Gradient-based optimization
  - Evolutionary Algorithm

BOSCH

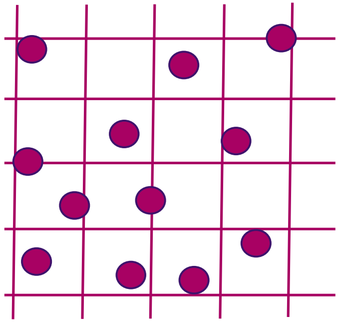
Model-Specific Pre-processing and Validation

Hyperparameter Tuning – Overview of Common Strategies

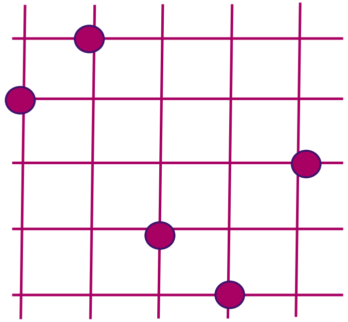
- Testing different parameter combinations



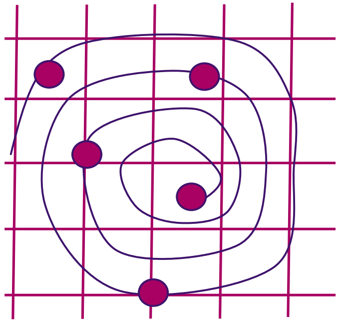
(Log)Grid Search



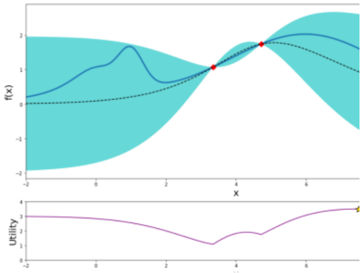
Random Search



Latin Hypercube



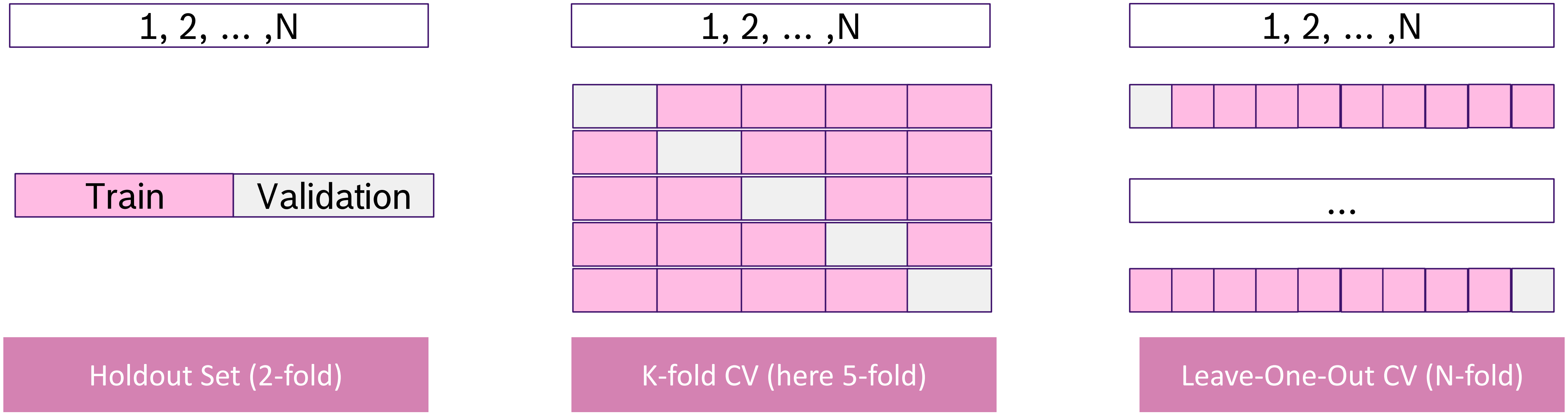
Spiral (In/Out)



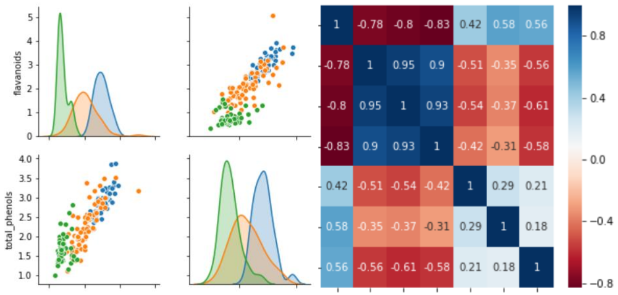
Bayesian Optimization

Model-Specific Pre-processing and Validation
Model Selection – Cross-Validation

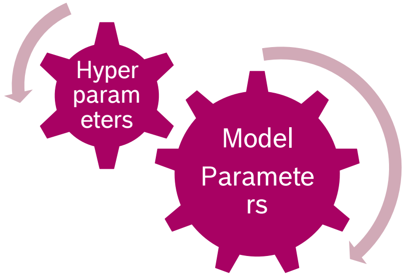
- Estimation of prediction error on validation data
- Splitting *N* data points into *k* folds and rotating the validation data set → average performance

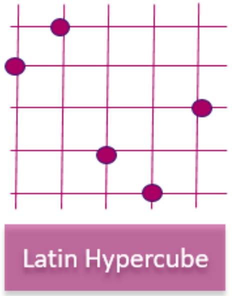


Data Science Workflow
Summary So Far



Data Exploration

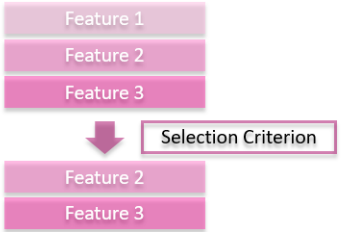


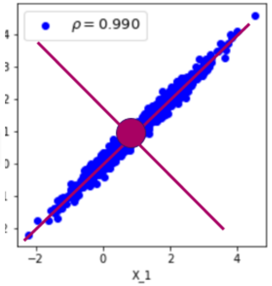


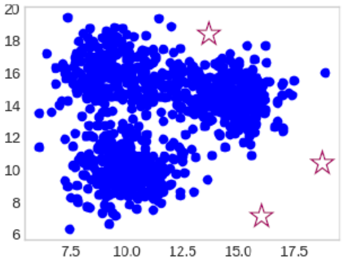
Modelling

Deployment

Data pre-processing







Validation

		True Class	
		Class="Yes"	Class="No"
Predicted Class	Class="Yes"	True Positive (TP)	False Positive (FP)
	Class="No"	False Negative (FN)	True Negative (TN)