DSS
Decision Sciences & Systems

Technische Universität München

Tutorial Business Analytics

Tutorial 9: Ensemble Methods and Clustering

Decision Sciences & Systems (DSS)

Department of Informatics

TU München

Tutorial 9 Business Analytics: Clustering and Ensemble Methods

Today's Agenda

1. Ensemble Methods

1.1 Theory: What are **Ensemble Methods**?

1.2 Theory: **Bagging**

1.3 Theory: **Boosting**

1.4 Theory: **Stacking**

2. Clustering

2.1 Theory: Difference Between **Classification and Clustering**

2.2 Theory: Partitional Clustering: **K-Means**
Practise: **Exercise 9.1**

2.3 Theory Probabilistic Clustering: **EM Algorithm**
Practise: **Exercise 9.2**

Tutorial and Homework

- **Exercise 9.1**
- **Exercise 9.2**
- **Exercise 9.3**
- **Exercise 9.4**
- **Exercise 9.5**

Tutorial 9 Business Analytics: Ensemble Methods

2.1 What are Ensemble Methods?

Ensemble Methods

- Ensemble methods can be used for **classification**
- Consider the **output of different expert models** for decision making
- Hope: **Increase of predictive performance** over a single model

Witten, Frank, Hall (2011), Data Mining, p.351-352.

Recall Learning Theory: Bias-Variance Tradeoff

Q: "Why can a combination of several bad models achieve better results than a good single model?"
A: Ensemble different models with **low bias** and **high variance** may reduce overall variance.

Three types of ensemble methods:

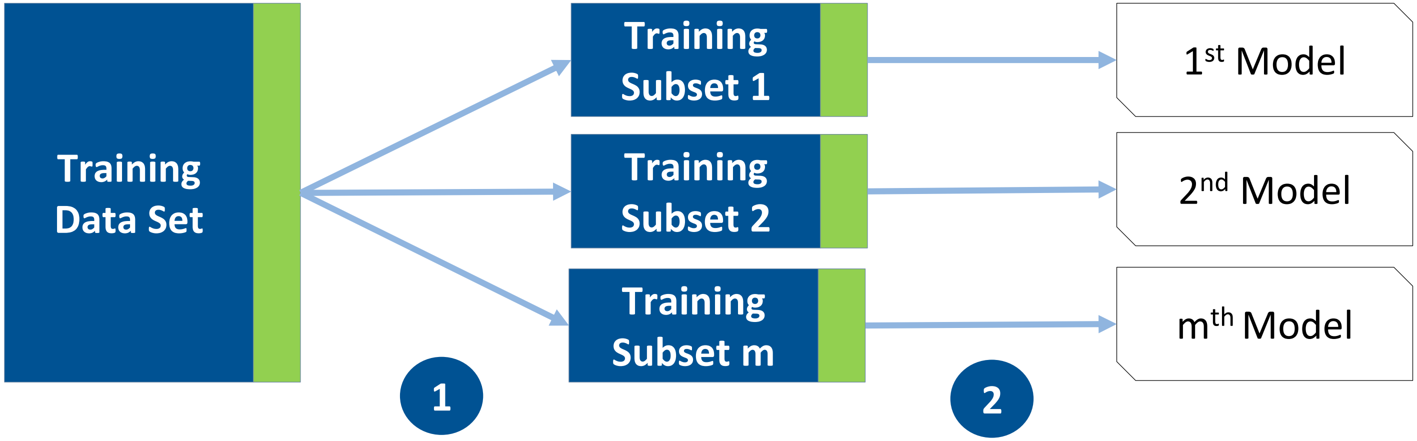
- Bagging
- Boosting
- Stacking

Tutorial 9 Business Analytics: Ensemble Methods

2.2 Bagging

1. Training Models 2. Classifying Instances

1. Sample **m training subsets** of size **n** from training data (of size **n**)
2. Train one model **for each training subset**

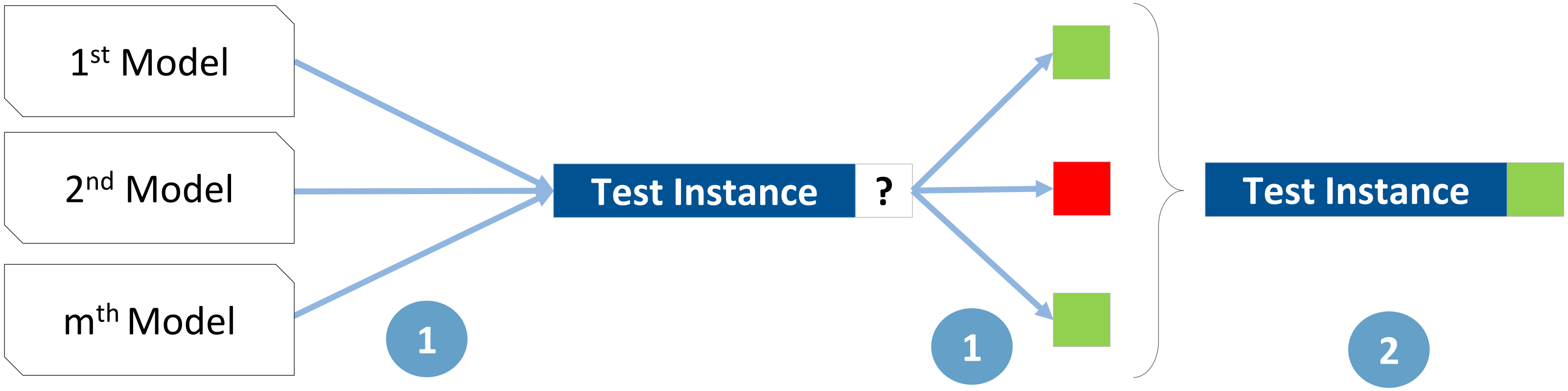


Tutorial 9 Business Analytics: Ensemble Methods

2.2 Bagging

1. Training Models	2. Classifying Instances

1. Each model **gives a vote** for a class
2. Chose class which received **the majority of the votes**

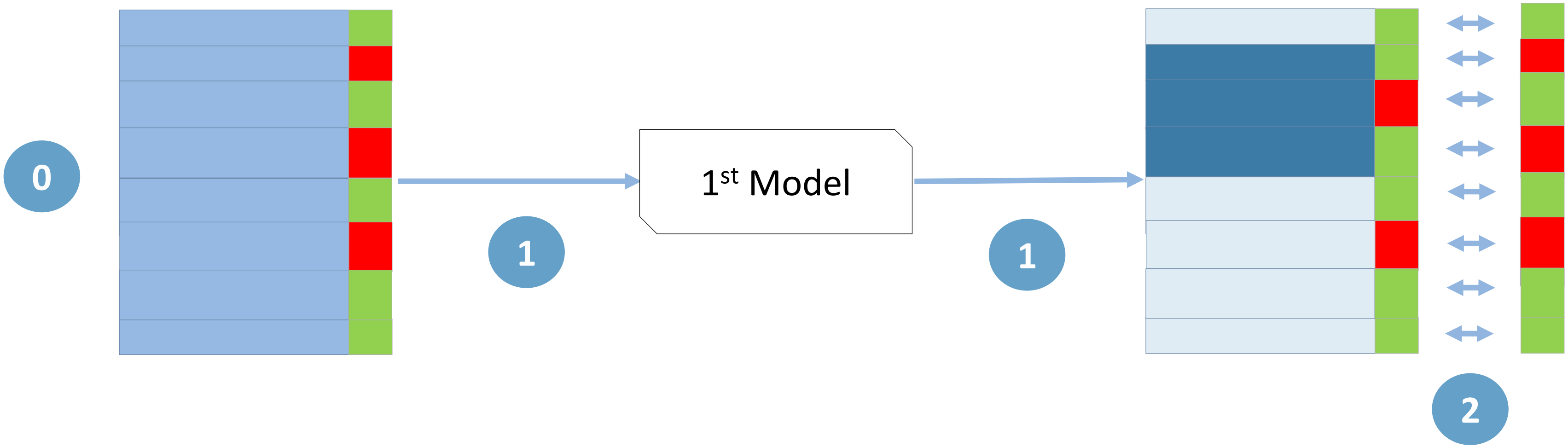


Tutorial 9 Business Analytics: Ensemble Methods

2.3 Boosting

1. Training Models – 1st round | 2. Classifying Instances

0. **Initialization**: All training data set's instances have **equal weights**
1. The **first model** is trained and **predicts the training data's instances**
2. **Evaluation** of prediction: Weights of correctly classified instances are reduced (and vice versa)

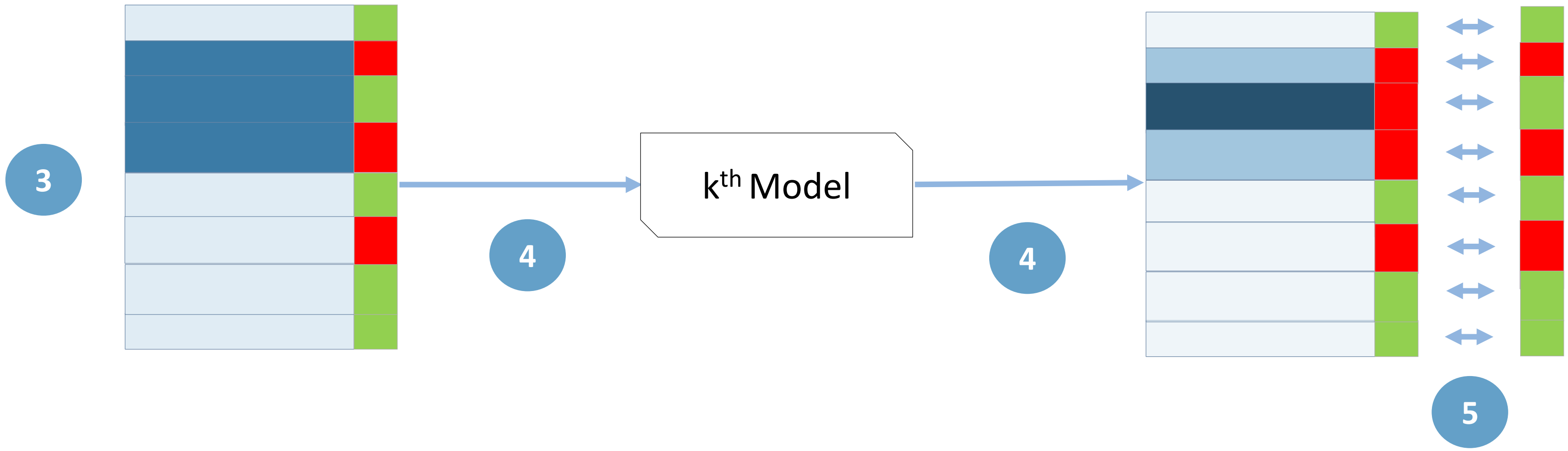


Tutorial 9 Business Analytics: Ensemble Methods

2.3 Boosting

1. Training Models – kth round | 2. Classifying Instances

3. **Precondition**: All training data set's instances have **different weights**
4. The **kth model** is trained on the training data set **focusing on high weights**
5. The prediction is **evaluated**: Weights of correctly classified instances are reduced (and vice versa)

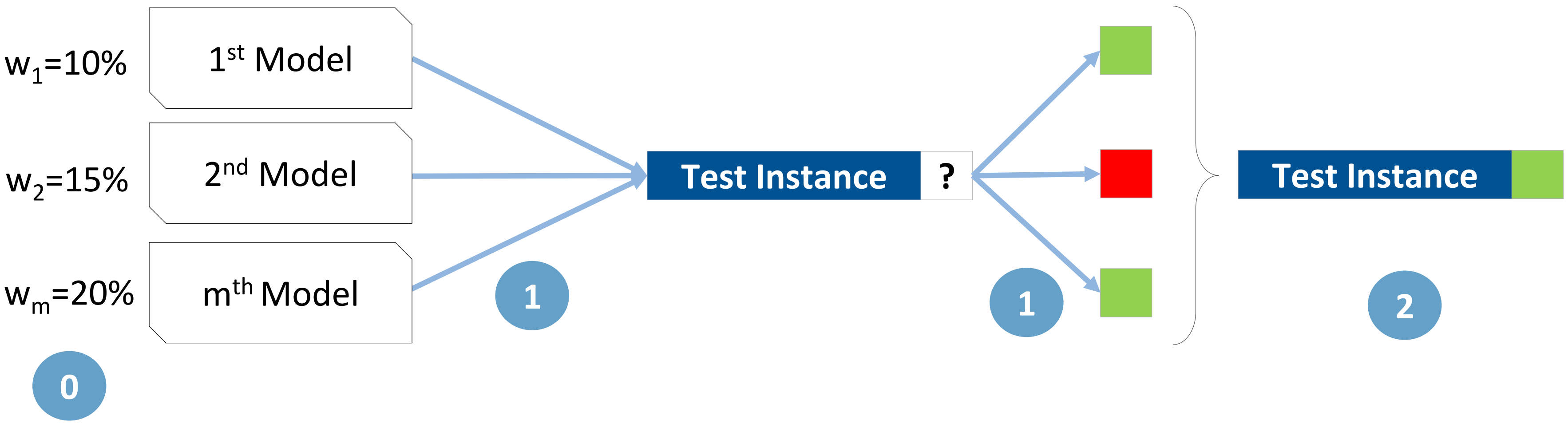


Tutorial 9 Business Analytics: Ensemble Methods

2.3 Boosting

1. Training Models – kth round | **2. Classifying Instances**

0. Each model is assigned a **weight** according to its error rate during training
1. Each model **gives a vote** for classifying the test data set
2. Prediction is based on a **weighted majority vote**

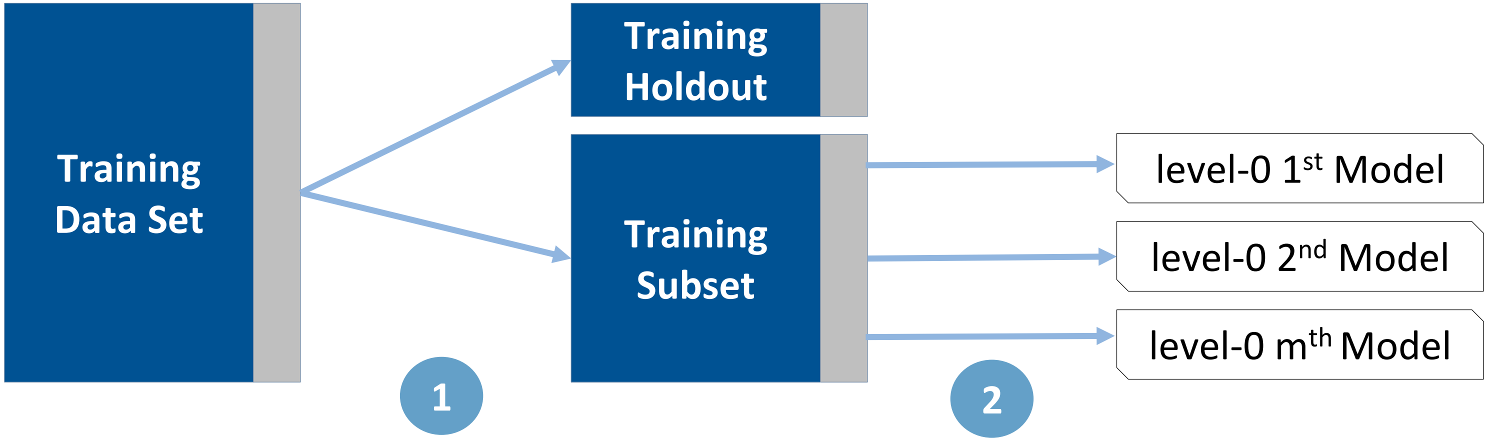


Tutorial 9 Business Analytics: Ensemble Methods

2.4 Stacking

1. Training Models – level-0 | 2. Classifying Instances – level-0

1. **Split** the training data set into a **training subset** and a **holdout subset**
2. **m models** are trained on the training subset. They are called level-0 classifiers

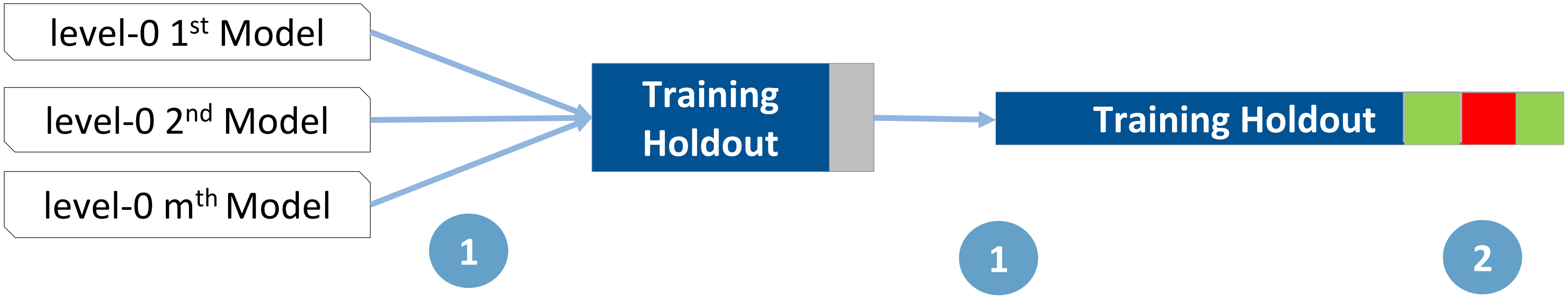


Tutorial 9 Business Analytics: Ensemble Methods

2.4 Stacking

1. Training Models – level-0	2. Classifying Instances – level-0

1. The level-0 models predict the **training holdout's label**
2. The training holdout data set contains the **predictions of the level-0 classifiers only**

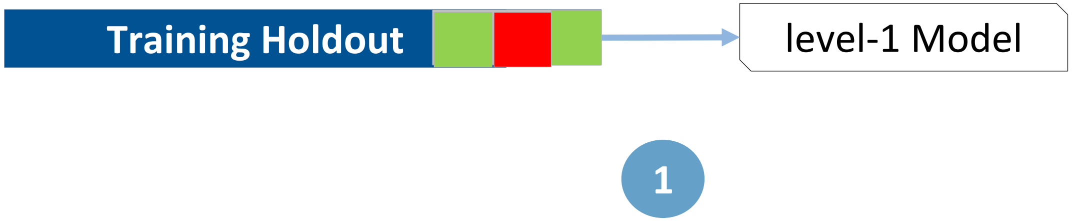


Tutorial 9 Business Analytics: Ensemble Methods

2.4 Stacking

1. Training Models – level-1	2. Classifying Instances – level-1

1. The training holdout now serves as the training data set for a **single level-1 model.**

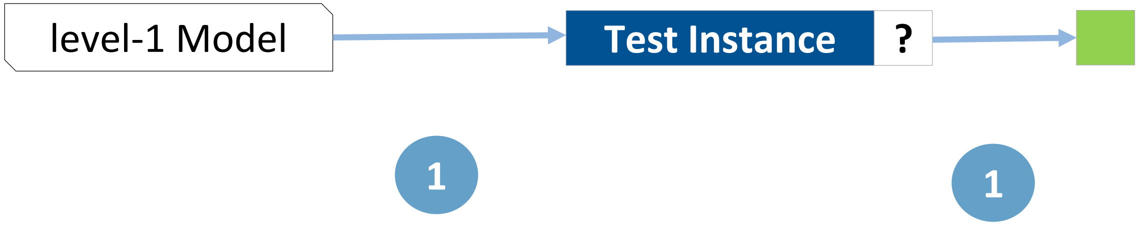


Tutorial 9 Business Analytics: Ensemble Methods

2.4 Stacking

1. Training Models – level-1	2. Classifying Instances – level-1

1. Finally, the level-1 model is used to classify test data set



Tutorial 9 Business Analytics: Clustering

1.1 Clustering Definition

Given: A p-dimensional data set with n instances.
Want: **Partition** data set into a number of clusters (k)

Clusters Characteristics:

- Items in same cluster are similar: **Intra-cluster similarity is maximized**
- Items from different clusters are different: **Inter-cluster similarity is minimized**

Tutorial 9 Business Analytics: Clustering

1.1 Difference between Classification and Clustering

Classification

Characteristics

- Supervised learning
- Target is known
- Training data

Examples

- Naïve Bayes
- Decision Trees
- Ensemble Methods

Clustering

Characteristics

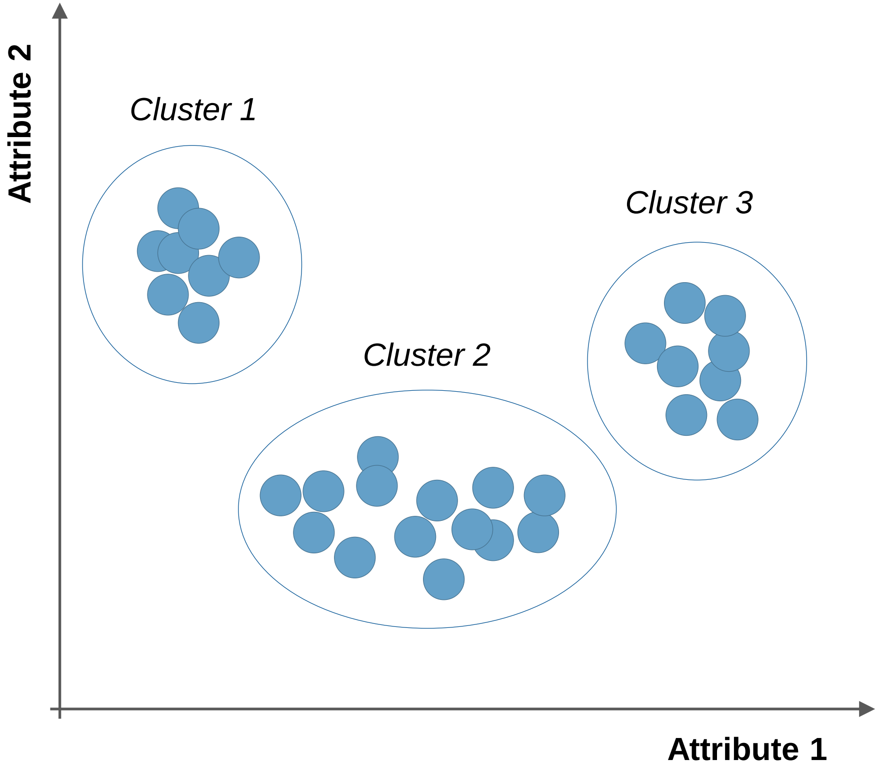
- Unsupervised learning
- Target is unknown
- No labels → no true class

Examples

- K-means
- Minimal Spanning Tree
- Expectation Maximization

Tutorial 9 Business Analytics: Clustering

1.1 Central Issues in Clustering



Clustering: Two Key Questions

Q1: “What is the right k (the number of clusters)?”

Q2: “How to identify class membership of instances?”