Project in Data Intensive Systems

4DV652
Lab Lecture 5b
Welf Löwe

1

Agenda

- Ensemble techniques in regression and classification
- Lab 5b task descriptions

2

Ensemble techniques

- Bagging
- Random Forest
- Boosting

- Heterogenous ensembles with different estimators
- Stacking

3

Heterogenous ensembles with different estimators

Idea:
- Train different models using different approaches
- Independent approaches (regression-, probability-distribution-, tree-based)
- Aggregate the results using
  - averaging (regression) and
  - majority voting (classification)
- Can be combined with bootstrap for quasi-independent training datasets

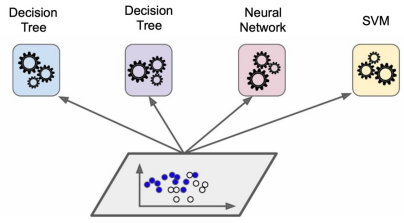
Assumption:
- gives independent predictors,
- hence, better predictions with lower error

4

Bootstrap: getting quasi-independent training datasets

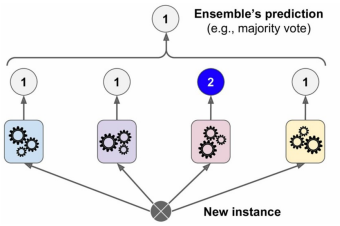
5

Heterogenous ensembles training



6

Heterogenous ensembles prediction



7

Other ensemble techniques

- Bagging
- Random Forest
- Boosting

- Heterogenous ensembles with different estimators
- Stacking

8

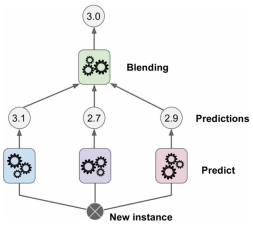
Stacking

Idea
- Replace simple aggregation using average or mode with yet another machine learning model also called blender or meta-learner
- Maps output of the ensemble models to the final prediction
- It is trained on the output of the individual models and the actual response
- Can be combined with both bootstrap and heterogenous ensembles

Assumption:
- can learn complex aggregation functions
- hence, better predictions with lower error

9

Stacking prediction



10

Resources

- Tutorial by Pádraig Cunningham: https://towardsdatascience.com/ensembles-in-machine-learning-9128215629d1
- Comes with Jupyter notebooks in Python: https://github.com/PadraigC/EnsemblesTutorial

11

Agenda

- Ensemble techniques in regression and classification
- Lab 5b task descriptions

12

Lab assignment 5b: ensemble techniques for regression and classification

- ML
  - Challenge the current champion regression with an ensemble approach
  - Challenge the current champion classification with a ensemble approach
- Software development
  - If applicable, implement and deploy the new champion regression and classification
- Reporting:
  - In a next notebook, document the iteration(s) over the ML process steps
- Deadline: 2023-03-08

13