Workshop 8

COMP90051 Machine Learning

Semester 2, 2018

Learning Outcomes

At the end of this workshop you should be able to:

1. explain the pros and cons of ensemble methods

Discussion

2. implement a Random Forest using the DecisionTreeClassifier in sklearn
3. apply Gradient Boosting Trees with hyperparameter tuning

Worksheet 8

Ensembles of Trees

Common characteristics for ensembles of trees

Advantages

- Non-linear
- Can provide variable importance measures
- Natural support for categorical features

Disadvantages

- Poor interpretability

Ensembles of Trees

Characteristics specific to Random Forests

Advantages

- Validation "for free" using OOB estimates
- Variance reduction
- Resilience to correlated features (feature bagging)
- Scalable to large data sets (parallelisable)

Disadvantages

- May underfit (e.g. on rare cases)

Ensembles of Trees

Characteristics specific to Gradient Boosting Trees

Advantages

- Potential for high accuracy

Disadvantages

- Susceptible to overfitting
- Training is not parallelisable
- Hyperparameters should be tuned