Data Mining and Machine Learning

Ensemble learning

Gergely Horváth

November 24, 2021

Outline

1. Stacking

2. Bagging

3. Boosting

Stacking

- Level 0 models: arbitrary, heterogeneous
- Level 1 models: better to keep it simple
- Level 1 models are using labels and/or probabilities
- Separate training sets!

Bagging – Random forest

- Bootstrapping
- Homogeneous base classifiers
- Voting/averaging

Random forest:

1. Bootstrapping data
2. Choosing n features (randomly)
3. Train decision tree on the previously bootstrapped dataset and use only the n features we have selected randomly
4. Repeat the previous steps
5. Estimate the predictive power, e.g. OOB samples (different trees use voting scheme, BAGGING!)
6. An optional step is hyperparameter tuning

Boosting - Adaboost, Gradient boosting

Boosting: learning from the mistakes of others

- (Weighted) Voting/averaging
- Homogeneous base classifiers
- Iterative!

Adaboost:

1. Tree training (root node + leaves)
2. Sample weight updates
3. Model weight update
4. Repeat

Gradient boosting:

- Boosting: unlimited decision trees
- XGBoost: Gradient boosting with stronger regularization capabilities