MA5832 Data Mining and Machine Learning

Week 3, Tree-based Methods

Hong-Bin Liu

James Cook University

21 May 2020

Admistration

- Assessment 2: 20%, Due date: Week 4 - Sunday, 31st May 2020, 11:59pm AEST.
- Future sessions will be held on Thursday, 6:00pm AEST.
- Ask any questions during the session.
- Let me know if you are lost.

Outline

1. Review Gradient Decent
2. Tree-based Methods
3. Demo
4. Questions?

Gradient Decent on Assessment 1

- What problems did you face?
- How did you solve it?

Outline

1 Review Gradient Decent

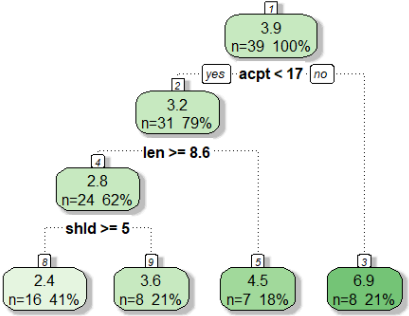
2 Tree-based Methods

3 Demo

4 Questions?

Motivation of Classification and Regression Trees (CARTs)

- It is one type of decision trees.
- Works on both classification and regression.
- Binary tree.
- Underline is probability theory.



Applications of CARTS

- Classification: predicting if a patient has a cancer or not, predicting an outcome of a treatment based on treatment type, predict types of emails (spam versus non spam).
- Regression: predicting housing price based on hedonic characteristics, predicting wage based on type of degree, working experience, and age.

Pros and Cons of Tree-based Methods

Advantages:

- Do not need to scale or normalise data
- Easy to explain

Disadvantages:

- Instability
- Lack of smoothness in the prediction of regression trees

Bagging

- Pros: bagging reduces the instability and improves the predictability
- Cons: bagging bad classifiers can produce worse results

Random Forest

- Enhanced version of Bagging.
- Subset of size m out of P (size of predictor) is used when splitting a node.
- Typically $m = \sqrt{P}$, when $m = P$, Random Forest is identical to Bagging.

Boosted Trees

- Same as Bagging, combining multiple "weak" learner to a "strong" learner.
- Sequential learner based on previous learner.
- Adjust weight of data based on previous learner.

Outline

1 Review Gradient Decent

2 Tree-based Methods

3 Demo

4 Questions?

Tree-based methods on Titanic Dataset

- What is Titanic Dataset?
- Our goal predict who would survive in Titanic Disaster.
- How do I train the model?
- How do I evaluate the performance of the model?
- How do I interpret the results?

Outline

1. Review Gradient Decent

2. Tree-based Methods

3. Demo

4. Questions?

Questions?

Thank You.