COMPSCI 762 2022 S1 Week 11 Questions

Luke Chang

May 13, 2022

Question 1

Answer the following questions regard to ensemble methods. The explanation for each question should be 200 words or less. Using figures to illustrate your solution is highly encouraged.

1. What are the two key factors an ensemble must have?
2. Bagging changes two thins in a dataset. What are they?
3. What is the main differences between *random forest* (RF), bagging and XGBoost?
4. Which of the "methods for constructing ensembles" do RF and XGBoost use?
5. Will variable importance in RF always give you the the "correct" answer? Why?
6. Between RF and bagging, what will the effect be of having a data set with a larger or smaller number of instances?
7. Between RF and bagging, what will the effect be of having a data set with a larger or smaller number of features?

Question 2

sklearn includes a collection of commonly used machine learning methods, including ensembles. You task is to benchmark decision tree (can be pruned or unpruned), SVM, RF, bagging, AdaBoost and XGBoost on the breast cancer wisconsin dataset (*sklearn.datasets.load_breast_cancer()*). Analyze your results. You need to install *XGBoost* separately. The installation guide can be found here: https://xgboost.readthedocs.io/en/stable/install.html#python