COMPSCI 762 2022 S1 Week 11 Questions

Luke Chang

May 27, 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?

   (a) Each model must perform better than random guess;

   (b) Must be uncorrelated;

2. Bagging changes two thins in a dataset. What are they?

   (a) The choice of data instances;

   (b) The distribution over them (with replacement);

3. What is the main differences between *random forest* (RF), bagging and XGBoost?

   XG Boost chooses without replacement so does not change the distribution of the dataset; Also Boosting will have to use a smaller training set by definition; RF uses democratic voting and XG Boost uses weighted voting.

4. Which of the "methods for constructing ensembles" do RF and XGBoost use?

   They are both manipulating the training set, manipulating the input features (columns), injecting Randomness.

5. Will variable importance in RF always give you the the "correct" answer? Why?

   No, because if you have correlated attributes Random forest will say neither are important even if they are the most important.
   This is because it randomizes the variables one at a time, thereby relying on the correlated variable when each is randomized.
   Example:
   Given a logic Function: $A \vee (B \wedge C)$ - B and C are correlated. If we train a tree with only A and B, or A and C, we will not have the correct logic.

6. Between RF and bagging, what will the effect be of having a data set with a larger or smaller number of instances?

   The effect on bagging and random forest will be the same. They both sample the instance space with replacement.

7. Between RF and bagging, what will the effect be of having a data set with a larger or smaller number of features?

   Since random forests sample the features, you might get better results when there are a lot of features because you got rid of a lot of noise, but with a data set with only a few features you might do worse because you are not left with enough features to make a good classifier.

Question 2

No standard solution for Question 2. Any reasonable analysis is a valid solution.