Name : Pawan Bathe Roll No.:CS18M519

CS6745: Mining Massive DataSets

Tutorial 2

September 13, 2019

- Write your name and roll number in the space provided
- *'Think out of the box, but write within the box.'*
- Be neat, and use the space judiciously.
- **Rough sheets won't be evaluated.**

1. (2 marks) Can a single decision tree return the same decisions as a given random forest classifier? If yes, why do we then use random forest classifiers? If no, explain why it cannot be done.

No, single decision tree can not return the same decisions as a given random forest classifier, whereas we can train and build more accurate model on a given dataset using decision tree, the decision tree will not perform well on unseen data, whereas random forest being the ensemble method uses different samples to train estimators used in combined random forest model which will reduce variance so it will give better decisions for unseen data than a single decision tree.

2. (1 mark) What is the key idea of boosting?

Key idea of boosting is converting weak learners to strong learners by iteratively training a classifier on a given dataset and reweighing each of data items in dataset after each iteration such that misclassified data gets more weights and correctly classified data items loose weights at each iteration of classification. so future learners focus more on the misclassified data. weak learners are trained sequentially such that each learner tries to correct it's predecessors.

3. (4 marks) You've studied Gradient Boosting for Decision Trees in class. Suggest some methods of regularization to prevent overfitting.

In Gradient Boosting, overfitting can be regularised by controlling the number of boosting iterations i.e early stopping . Beside number of boosting iterations, we can use shrinkange and subsampling for regularization. Shrinkage can improve accuracy by reducing variance using bagging. Subsampling along with shrinkage also act as good regularization strategy to improve accuracy of model. Subsampling without shrinkage doesn't help much to avoid overfitting. Gradient can also be regularise by penalize model complexity of the learned model

4. (3 marks) Among bagging and boosting, state the scenarios in which you would pick one over the other. Also explain why.

If data set we have is prone to overfitting i.e resulting into low bias but high variance for the single model then bagging would be better which will help to avoid over fitting. If data set we have is under performing with single model then boosting can iteratively help to boost the performance by assigning more weights to misclassified data items which is leading lower performance. Bagging is better when single model results into high variance and low bias, by training model on more samples we could reduce variance. Boosting is better when single model results into low variance and high bias, by boosting iteratively we can reduce bias.