Rahul Matta

1. How did you represent the decision tree in your code?
    a. We represented the decision tree with a nested dictionary. The key of a simple tree object is its root node, and the value is the root node's subtree.
2. How did you represent examples (instances) in your code?
    a. Examples were represented as a list of lists. The outer list contains the records in the file. Each record is a list of each attribute.
3. How did you choose the attribute for each node?
    a. The attribute with the highest info gain was chosen as the best attribute.
4. How did you handle missing attributes in examples?
    a. Missing attributes were transformed into the mode for discrete data, and into the median of the column for numeric data.
5. What is the termination criterion for your learning process?
    a. We have two base cases for our decision tree: if the number of attributes in the inputted attributes list is less than or equal to 1 (including the target), if the target entropy is equal to 0 (homogenous data).
6. Apply your algorithm to the training set, without pruning. Print out a Boolean formula in disjunctive normal form that corresponds to the unpruned tree learned from the training set. For the DNF assume that group label "1" refers to the positive examples. NOTE: if you find your tree is cumbersome to print in full, you may restrict your print-out to only 16 leaf nodes.

    oppstartingpitcher3.0 AND startingpitcher1.0 AND numinjured2.0 AND oppnuminjured0.0 AND oppwinpercent5.0 AND 1.0 AND OR oppstartingpitcher5.0 AND startingpitcher3.0 AND dayssincegame2.0 AND oppwinpercent5.0 AND numinjured1.0 AND oppdayssincegame4.0 AND 1.0 AND OR oppstartingpitcher1.0 AND startingpitcher3.0 AND numinjured3.0 AND oppwinpercent5.0 AND 1.0 AND OR oppstartingpitcher2.0 AND startingpitcher2.0 AND numinjured5.0 AND 1.0 AND OR oppstartingpitcher3.0 AND startingpitcher3.0 AND numinjured2.0 AND oppnuminjured2.0 AND oppwinpercent2.0 AND oppdayssincegame6.0 AND 1.0 OR oppstartingpitcher2.0 AND startingpitcher2.0 AND numinjured4.0 AND oppwinpercent4.0 AND 1.0 AND oppstartingpitcher3.0 AND startingpitcher4.0 AND oppnuminjured0.0 AND oppwinpercent5.0 AND 1.0 OR oppstartingpitcher4.0 AND startingpitcher3.0 AND oppnuminjured1.0 AND numinjured3.0 AND oppwinpercent4.0 AND 1.0 OR oppstartingpitcher2.0 AND startingpitcher1.0 AND oppwinpercent2.0 AND dayssincegame2.0 AND oppnuminjured0.0 AND oppdayssincegame1.0 AND winpercent2.0 AND 1.0 AND oppstartingpitcher4.0 AND startingpitcher1.0 AND numinjured4.0 AND oppwinpercent2.0 AND weather0.0 AND 1.0 OR oppstartingpitcher4.0 AND startingpitcher4.0 AND numinjured5.0 AND 1.0 OR oppstartingpitcher1.0 AND startingpitcher2.0 AND oppnuminjured2.0 AND numinjured2.0 AND oppwinpercent2.0 AND winpercent3.0 AND oppdayssincegame3.0 AND weather0.0 AND temperature3.0 AND dayssincegame3.0 AND 1.0 AND oppstartingpitcher3.0 AND startingpitcher1.0 AND numinjured1.0 AND oppnuminjured3.0 AND opprundifferential4.0 AND winpercent1.0 AND 1.0 OR oppstartingpitcher4.0 AND startingpitcher3.0 AND oppnuminjured4.0 AND oppwinpercent3.0 AND oppdayssincegame1.0 AND 1.0 OR oppstartingpitcher3.0 AND startingpitcher4.0 AND oppnuminjured2.0 AND numinjured3.0 AND oppwinpercent4.0 AND 1.0 OR oppstartingpitcher3.0 AND startingpitcher1.0 AND numinjured4.0 AND oppnuminjured4.0 AND 1.0 OR oppstartingpitcher2.0 AND startingpitcher5.0 AND oppnuminjured0.0 AND oppwinpercent2.0 AND winpercent1.0 AND 1.0

7. Explain in English one of the rules in this (unpruned) tree.
    a. If the second opposing team starting pitcher is playing, the fifth starting pitcher on the home team is playing, there are 4 players injured on the home team and 4 people injured on the opposing team, then the home team will win.
8. How did you implement pruning?
    a. We took a randomly selected node, deleted it, and attempted to check the accuracy.
9. Apply your algorithm to the training set, with pruning. Print out a Boolean formula in disjunctive normal form that corresponds to the pruned tree learned from the training set.

    oppstartingpitcher1.0 AND startingpitcher4.0 AND numinjured4.0 AND oppwinpercent1.0 AND temperature2.0 AND 1.0 AND OR oppstartingpitcher3.0 AND startingpitcher2.0 AND oppnuminjured2.0 AND numinjured0.0 AND oppdayssincegame5.0 AND winpercent5.0 AND 1.0 AND OR oppstartingpitcher3.0 AND startingpitcher4.0 AND oppnuminjured1.0 AND numinjured5.0 AND 1.0 AND

OR oppstartingpitcher5.0 AND startingpitcher5.0 AND oppnuminjured3.0 AND numinjured4.0 AND oppwinpercent3.0 AND 1.0 AND OR oppstartingpitcher2.0 AND startingpitcher2.0 AND numinjured5.0 AND 1.0 AND oppstartingpitcher5.0 AND startingpitcher1.0 AND numinjured5.0 AND 1.0 AND OR oppstartingpitcher4.0 AND startingpitcher1.0 AND numinjured4.0 AND oppwinpercent3.0 AND 1.0 AND OR oppstartingpitcher1.0 AND startingpitcher5.0 AND oppwinpercent2.0 AND oppdayssincegame2.0 AND dayssincegame1.0 AND oppnuminjured3.0 AND numinjured3.0 AND 1.0 AND OR oppstartingpitcher3.0 AND startingpitcher3.0 AND numinjured3.0 AND oppwinpercent5.0 AND oppnuminjured3.0 AND 1.0 AND OR oppstartingpitcher3.0 AND startingpitcher3.0 AND numinjured0.0 AND oppnuminjured1.0 AND oppdayssincegame0.0 AND 1.0 AND OR oppstartingpitcher2.0 AND startingpitcher3.0 AND numinjured2.0 AND oppnuminjured2.0 AND oppwinpercent3.0 AND weather0.0 AND dayssincegame1.0 AND 1.0 AND OR oppstartingpitcher4.0 AND startingpitcher2.0 AND oppnuminjured1.0 AND numinjured2.0 AND oppwinpercent4.0 AND oppdayssincegame3.0 AND 1.0 AND OR oppstartingpitcher4.0 AND startingpitcher5.0 AND oppwinpercent4.0 AND oppnuminjured1.0 AND numinjured4.0 AND 1.0 AND OR oppstartingpitcher3.0 AND startingpitcher3.0 AND numinjured5.0 AND 1.0 AND OR oppstartingpitcher3.0 AND startingpitcher2.0 AND oppnuminjured1.0 AND numinjured3.0 AND oppwinpercent4.0 AND dayssincegame1.0 AND 1.0 AND OR oppstartingpitcher2.0 AND startingpitcher1.0 AND oppwinpercent2.0 AND dayssincegame4.0 AND numinjured2.0 AND oppdayssincegame3.0 AND 1.0 AND OR oppstartingpitcher2.0 AND startingpitcher3.0 AND numinjured4.0 AND oppwinpercent3.0 AND dayssincegame2.0 AND 1.0

10. What is the difference in size (number of splits) between the pruned and unpruned trees?
    a. The pruned tree has 80 splits, whereas the unpruned tree has approximately 120 splits.
11. Test the unpruned and pruned trees on the validation set. What are the accuracies of each tree? Explain the difference, if any.
    a. Because our pruning strategy was unsophisticated, the accuracy for the pruned tree was lower than that of the unpruned tree. Pruned: 88%, Unpruned: 86%
12. Create learning curve graphs for both unpruned and pruned trees. Is there a difference between the two graphs?
    a. Please see ReadMe
13. Which tree do you think will perform better on the unlabeled test set? Why? Run this tree on the test file and submit your predictions as described in the submission instructions.
    a. Because the accuracies for the two trees were so similar, it is hard to say. Because pruning is a method for improving accuracy, I would think that its tree would perform better, but since our pruning strategy did not result in better results, I cannot make a confident determination. Please see ReadMe for part 2 of this question.
14. What aspects of the feature set (if any) are a good fit for decision trees, and what aspects aren't a good fit?
    a. The nominal features such as starting pitcher, and number of injured people tend to be better with decision trees because there is a set amount of splits that can be done as opposed to numeric data which, if not binned, would do a large amount of splits.
15. Which members of the group worked on which parts of the assignment?
    a. Sachin Lal worked on pre-processing, validation, and pruning. I worked on the create decision tree function as well as the disjunctive conjunctive normal form.