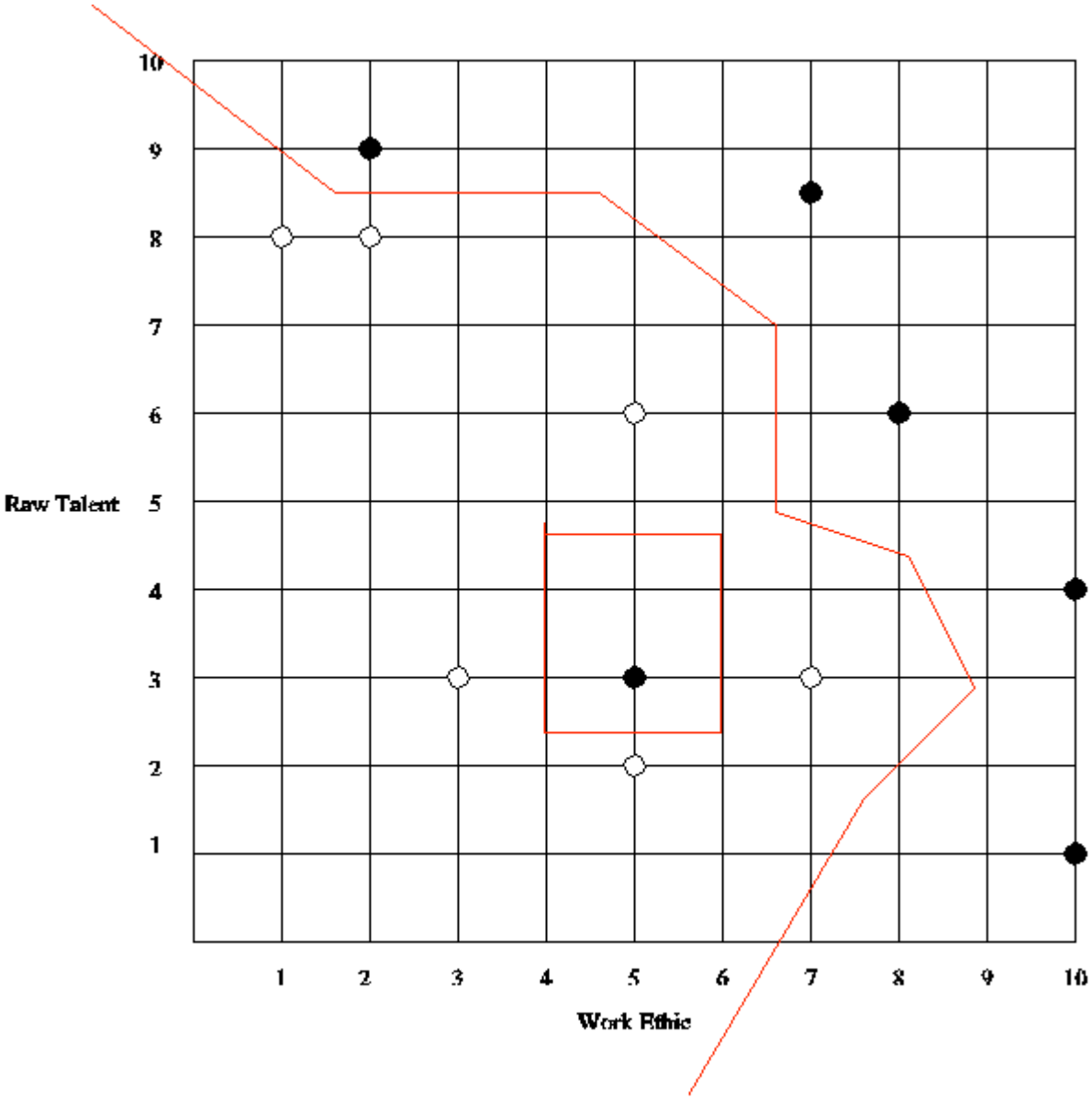
Final Exam 2002 Problem 3: Classification (14 Points)

Part A: Nearest Neighbors (6 Points)

The 6.034 staff has decided to launch a search for the newest AI superstar by hosting a television show that will make one aspiring student an *MIT Idol*. The staff has judged two criteria important in choosing successful candidates: work ethic (W) and raw talent (R). The staff will classify candidates into either potential superstar (black dot) or normal student (open circle) using a nearest-neighbors classifier.

On the graph below, draw the decision boundaries that a 1-nearest-neighbor classifier would find in the R-W plane.



Part B: Identification Trees (4 Points)

Part B1 (2 Points)

Now, leaving nearest neighbors behind, you decide to try an identification-tree approach. In the space below, you have two possible initial tests for the data. Calculate the average disorder for each test. Your answer may contain $\log_2$ expressions, but no variables. The graph is repeated below.

Test A: R > 5:

(+ (* _ (* 2 (- (* _ log _))))
(* _ (* 2 (- (* _ log _))))) = 1

Test B: W > 6:

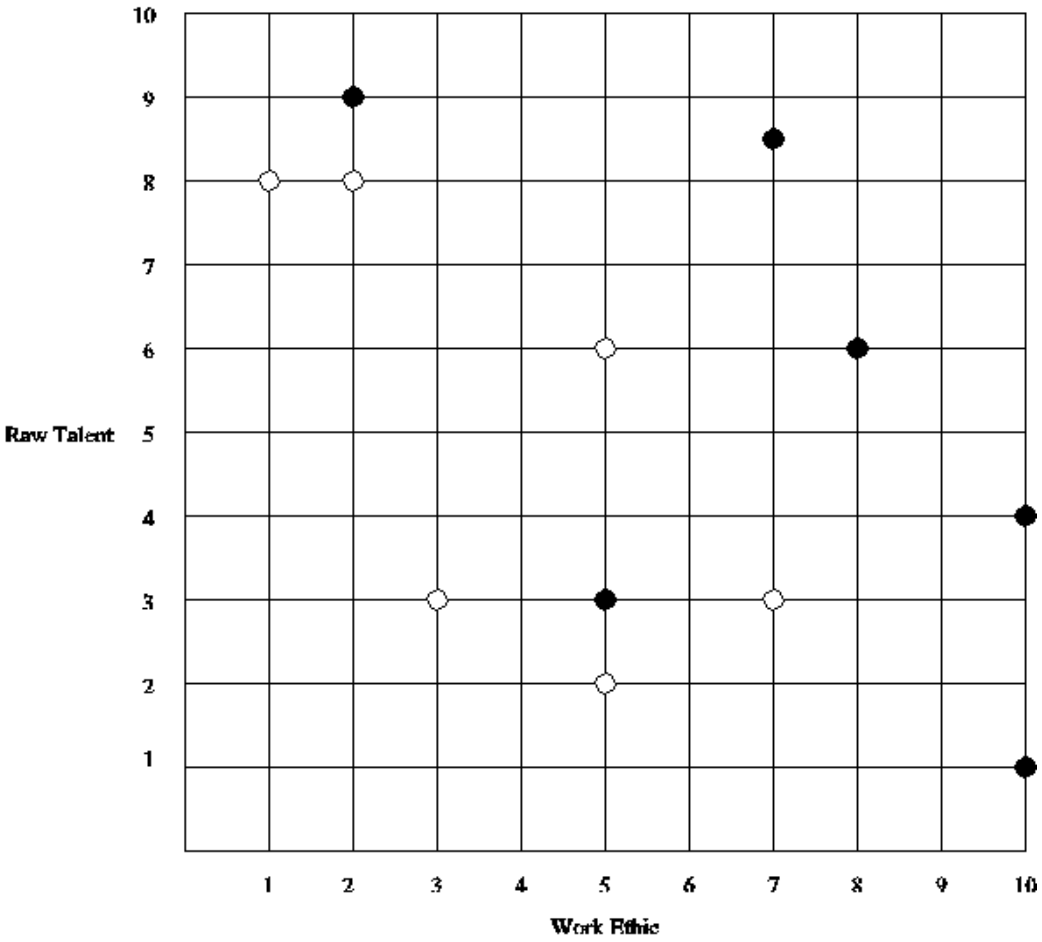
(+ (* 5/12 (+ (* 1/5 log 1/5) (* 4/5 log 4/5))
(* 7/12 (+ (* 2/7 log 2/7) (* 5/7 log 5/7)) < 1

Part B2 (2 Points)

Now, indicate which of the two tests is chosen first by the greedy algorithm for building identification trees.

B

We include a copy of the graph below for your scratch work.



Part C: Identification Trees (4 Points)

Now, assume R > 5 is the first test selected by the identification-tree builder (which may or may not be correct). Then, draw in all the rest of the decision boundaries that would be placed (correctly) by the identification-tree builder:

Note: other solutions do just as good a job at dividing up the bottom portion of the graph.