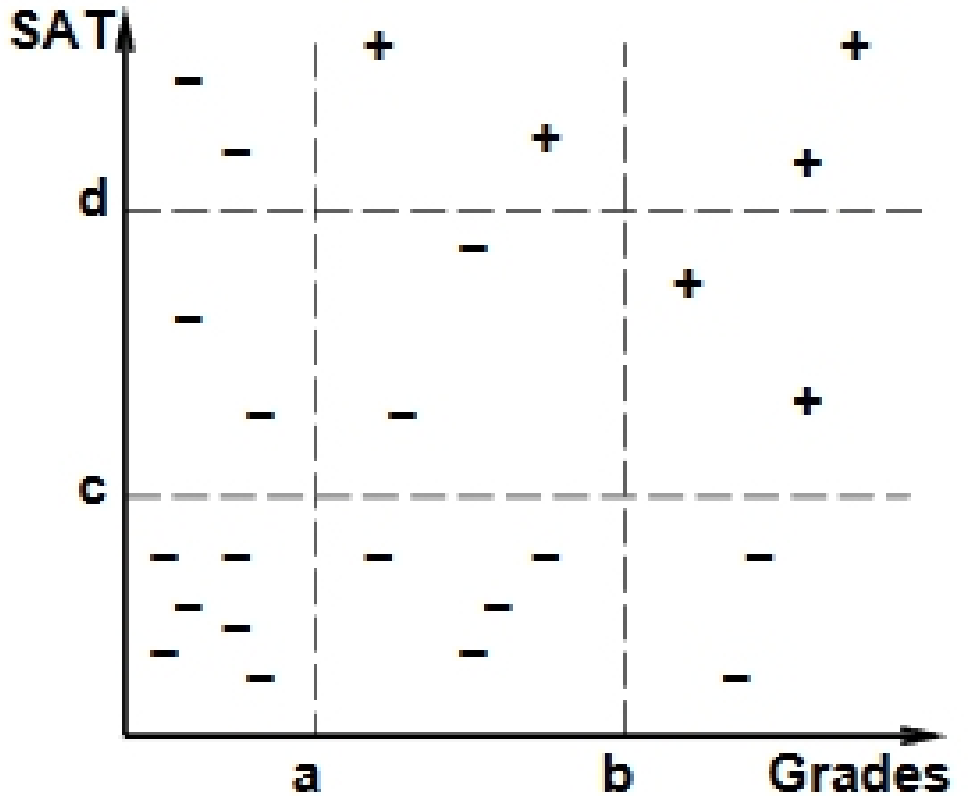
Problem 3 Identification Trees (20 points)

This problem contains independent parts, so do not give up on the problem as a whole just because you give up on part of the problem.

You are given the following admissions data from the MIT admissions office. A "+" indicates someone admitted and a "−" indicates someone not admitted. The x axis is grades (on a 0-100 scale) and the other is average SAT score (on a 0-800 scale):



In building an identification tree for this data, we will consider only the following four tests as candidates for the nodes:

1. **SAT** $>$ **c**
2. **SAT** $>$ **d**
3. **Grades** $>$ **a**
4. **Grades** $>$ **b**

Part A (4 points)

Compute the value of the average disorder measure for the test **SAT** $>$ **c**:

Part B (4 points)

Assume that the test **SAT > c** is chosen as the test at the top of the identification tree. Consider the remaining tests that could be used at the next level of the tree:

- **SAT > d**
- **Grades > a**
- **Grades > b**

1. Which one of these tests, if any, should be used at the next level of the tree for the branch in which the top-level test (**SAT > c**) is *false*. Indicate in one sentence the reason for your answer.

2. Now, for the branch in which the top level test (**SAT > c**) evaluates to *true*. Next to each test below, write its rank based on average disorder. That is, write a 1 next to the test with least average disorder, a 2 next to the next best, etc. If there is a tie, use the same rank for the tied tests. **Hint: you need not do detailed calculations to determine the correct answer.**

Test	Rank
SAT > d	
Grades > a	
Grades > b	

Part C (8 points)

Given your success so far, you decide to turn your horizons past college, to the financial future, and use ID trees to work on bankruptcy prediction.

From 10 features (synonym for tests), measured surreptitiously from credit card usage patterns, you propose to a big bank to "guarantee" that you can predict consumer bankruptcies 18 months in advance.

For some reason, the tenth feature, which you thought would be used prominently in the identification tree built by your ID tree program, is not deployed at all for the sample data you provide. Naturally, you suspect a bug, but in fact, when you pay your 6.034 recitation instructor big bucks for consulting, he provides an alternative explanation.

Check all of the following that your recitation instructor might have said, without destroying your faith in his grip on the material. That is, check all of the following that could explain why the tenth feature is not used:

☐ The tenth feature usefully splits the sample data into groups, but it splits the sample data much like the ninth feature, but not quite as well. Therefore, the tenth feature is always dominated by the ninth feature.

☐ The tenth feature is useful, but only in combination with the ninth feature. Because ID tree construction does no look ahead or test combinations, the value of the tenth feature is overlooked.

☐ The values of the tenth feature are all tightly clustered around the mean. The feature can be useful, but you must first divide all values by the standard deviation of the values for the tenth feature in the sample set.

☐ The tenth feature is not useful. However, a new, eleventh feature can be derived from the tenth feature, and that new eleventh feature is useful.