

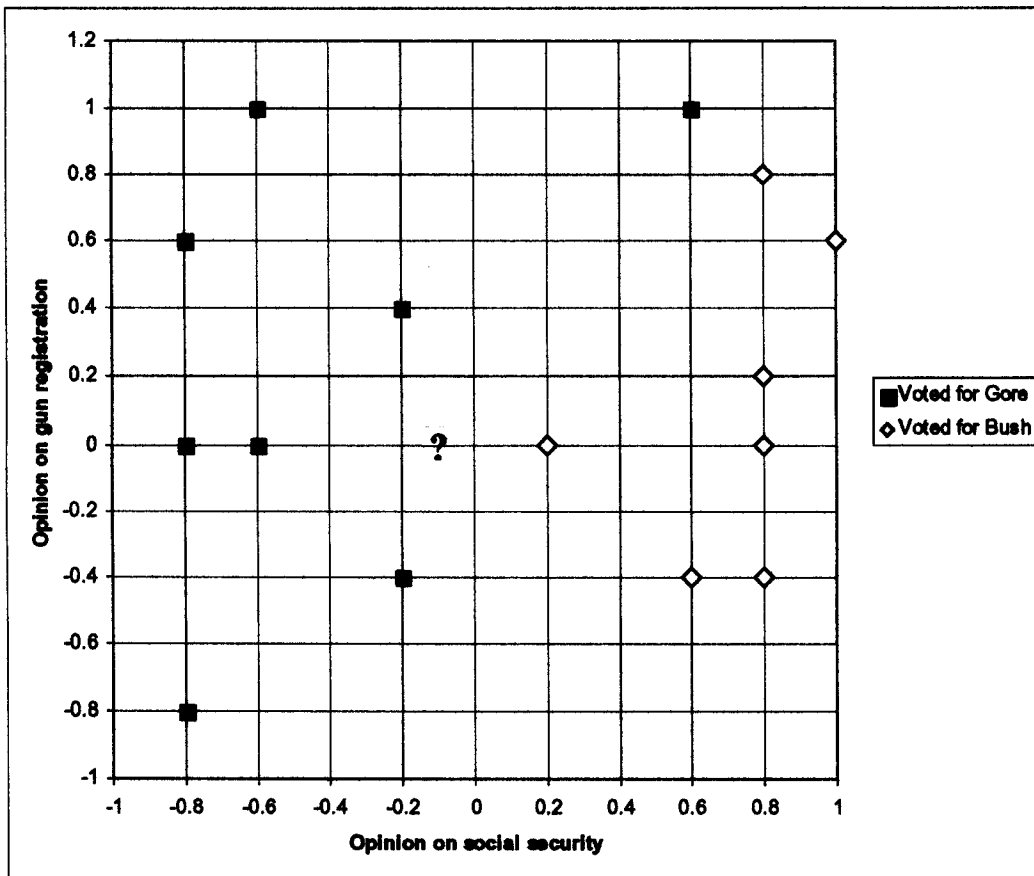
Recitation 9: Nearest Neighbor, ID Trees Problem (from Exam 2, Fall, 2000) KKoile

Congress has decided to ask each voter a few key questions so as to predict how each will vote. This will, of course, save everyone the troublesome and time-consuming practice of actually having to examine the ballots to figure out the election result.

They decide to start with just two questions:

- a) On a scale of -1 (strongly disagree) to $+1$ (strongly agree) how do you feel about privatizing social security?
- b) On a scale of -1 to $+1$, how do you feel about registering handguns?

The training sample is shown below, with 15 individuals plotted according to how they feel about these two issues, with a dark square (for Gore) and a light diamond (for Bush) indicating how they voted for president. There is also a question mark “?” on the plot, indicating one of the infamous undecided voters about which so much has been said during this election. We’ll call him Mr. Undecided.



Part A: Nearest Neighbor (10 points)

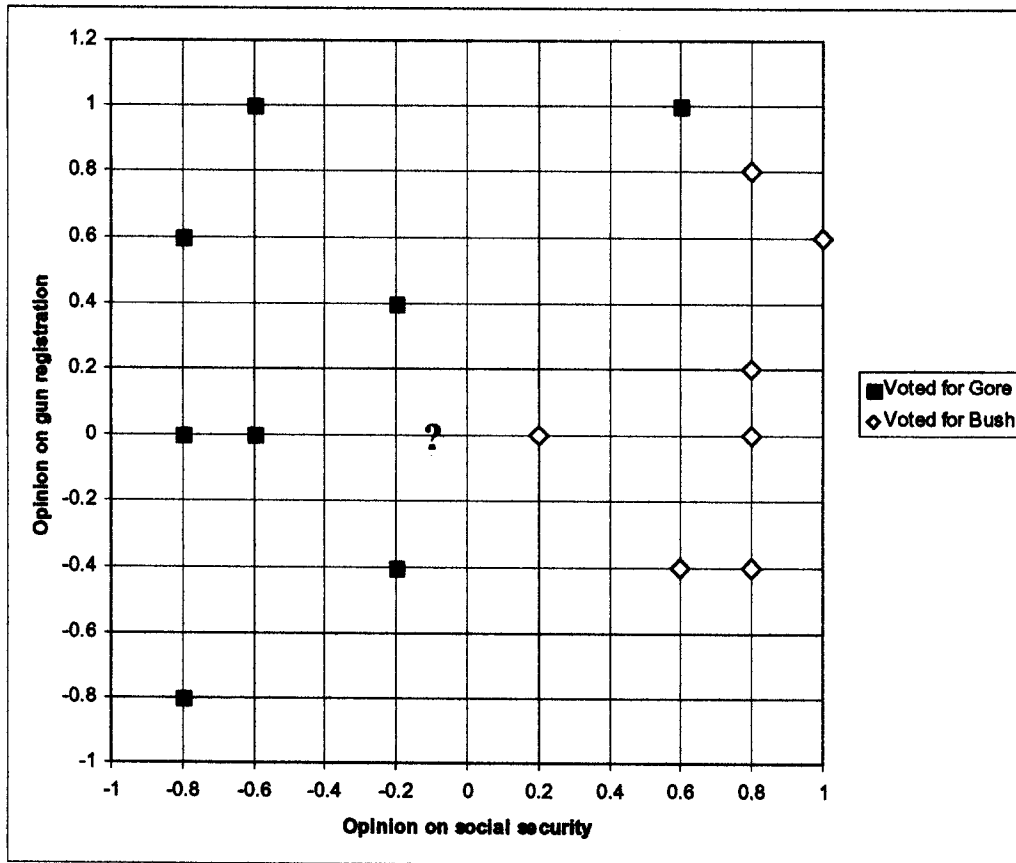
- 1) What would nearest neighbor predict about the vote of Mr. Undecided, assuming the use of the standard Euclidean distance as the metric? (Your answer should be either Gore or Bush.)

- 2) On the plot above, carefully draw the precise boundary lines that nearest neighbor would indicate as separating the Gore part of the sample space from the Bush part. Do not include the ? in your analysis.

- 3) What would 3-nearest neighbor predict about the vote of Mr. Undecided (using the same Euclidean metric)?
- 4) It turns out there was one other question that voters had been asked: “How do you feel about lowering the pay of Congressmen/women?” The question was not included in the publicly released data because, (according to the politicians who controlled the release), the data will not be useful in making a decision. When digging in, you find out that the actual problem was that all the answers were strongly clustered near the +1 end of the scale. You are brought in as a consultant and suggest that:
- a) The politicians are correct; the data will not be useful.
 - b) The data can still be useful, you just need to **divide** all the values by the mean of the value.
 - c) The data can still be useful, you just need to **subtract** from each value the mean of all the values, then **multiply** all the values by the standard deviation of the values
 - d) The data can still be useful, you just need to **divide** by the standard deviation.
 - e) The data can still be useful, but none of the choices offered above are correct.

B: Identification Trees (8 points)

(We repeat the same data here for your convenience.)



Things seem to be going along well, when suddenly, Ralph Nader appears on the scene and suggests that nearest neighbor is wrecking the environment by wasting precious time and space. He suggests using ID trees instead.

- 1) You decide to try as your first test opinion on social security < 0 . But as you know, you need to determine the average disorder of the sets produced by this test to see whether it's any good. What is the average disorder? (Your answer can include the \log_2 operator, you need not simplify your expression.)

- 2) You decide that it looks good, so you decide to complete the decision tree. Draw the ID tree and specify all tests. Do not include the ? in your analysis.

- 3) What does your tree predict about how Mr. Undecided will vote for president? Circle the corresponding node on your ID tree.

Part C: More Voter Questions (4 points)

Mr. Gore, having invented the Internet, claims to know a thing or two about technology. He says that we're asking way too few questions of the voters, and indicates that to get a decent predictive ability we should ask them at least 100 questions. You come up with 100 questions, and while you worked hard at it, you don't think all 100 questions are going to give you predictive information about the voter. Nevertheless, you forge ahead and try using both nearest neighbors and identification trees. Your initial experiments indicate (circle the most likely result):

- a) Both techniques work well and work about equally well.
- b) Nearest neighbors works much better than identification trees.
- c) Identification trees work much better than nearest neighbors.
- d) Neither works well.