Assignment 9 (Extra Credit)

36-350, Data Mining

Due Monday, 15 December, at 10:30 am

1. *Can you predict all of the people some of the time?* The data set for this assignment on Blackboard (`popular-vote-margins-vs-predictions.txt`) lists, for each state[1], the popular votes received by the two major party candidates for president in 2008, the number of votes for minor parties, and the predictions of three different forecasters (`538.com`, EV = `election.princeton.edu`, and `pollster.com`). — This data was kindly provided by "washerdryer" of `Unfogged.com`.

   (a) Calculate the mean squared error of each forecaster's predictions for the vote share. What is the ranking of forecasters?

   (b) Calculate the median absolute error. Does the ranking change?

   (c) Calculate the mis-classification rate, i.e., the fraction of states where the candidate predicted to get a majority actually got only a minority. What is the ranking?

   (d) Calculate the weighted mean squared error, with each state weighted by population. What happens to the ranking?

   (e) Plot error versus predicted margin for each forecaster, and describe any patterns.

   (f) Plot error versus *actual* margins, and describe any patterns.

   (g) Plot error versus the total number of votes cast, by state. Describe any patterns.

   (h) Which forecaster would you trust in 2012? (Explain.)

2. *Rates and costs* The old test for disease X had an error rate of 0% false positives and 10% false negatives. A new test, based on extensive data mining of patients' medical records, has an error rate of 1% for both false positives and false negatives. The fraction of patients who have X is p.

   (a) Patient who do not have X but are falsely diagnosed with it must undergo painful and embarrassing follow-up tests, equivalent to a cost to them of $5,000. Patients who have X but are not diagnosed with it die, which for the purposes of this problem is worth $10,000,000

[1] and the District of Columbia; but I'll keep saying "state".

to avoid. Find the expected cost to the patient of: taking no test, taking the old test, and taking the new test, all as functions of p.

(b) When, as a function of p, would the patient want to take the new test, the old test, or no test at all?

(c) For an insurance company, the cost of follow-up testing for someone who does not have X but is falsely diagnosed with it is \$5,000, and the cost of treating someone who really does have X is \$100,000. Untreated X costs the insurance company \$250 in pain-killers. Find the expected cost to the insurance company of offering the new test, the old test, and no test at all, as functions of p.

(d) For what values of p do patients and insurance companies agree on which test to take?