Log for Record Linkage Project on NC Voter Data:

Outline can be viewed as Tools > Document Outline

1. **6/23/2017 - Second Iteration:**

* Used R library e1071 for svm, stringdist & PGRdup for features.
* Used NC voter databases of district 15, years April 2013 & March 2017.
  + About 8,000 entities in the data.
* A total of 20 features have been used.
  + Four for each of first, middle and last name.
    - Edit distance, simple distance after string alignment.
    - Jaro\_Winkler distance, distance that gives more weights to similar characters is starting of a name.
    - Edit distance b/w Double Metaphone: Gives two alternate pronunciation of a name.
  + If age difference is inside (3,5) then 1 else 0.
  + Edit distance between address
  + If “edit distance” matches 1, else 0 for following parameters.
    - Resident\_City
    - Birth\_State
    - Race
    - Ethnicity
    - Party
    - Gender
* Generation of positive pairs:
  + Data is filtered only to required parameters.
  + Pairs were matched on the basis of voter registration number.
  + A total of 6,000 matches were found.
  + 4,000 of them were having all values. (Note: Most of the NA were present for people who have been removed from registration list, change of state, death etc)
  + In total we had 4,000 positive pairs.
* Generation of Negative pairs:
  + Out of 8000 observations, 6000 don’t have missing values.
  + Those 6000 values were cross matched to generate negative pairs.
  + 4 negative pairs for each observation.
  + Blocking was done at three levels: First name, address and age.
  + If voter reg no is not same, block on first name, select two pairs with lowest dissimilarity. Select next two by blocking on address. If 4 matches were not found, fill it by blocking on age.
  + Dissimilarity score:
    - Normalized edit distance for first, middle, last name and address.
    - If age difference is inside (3,5) then 0 else 1.
    - If “edit distance” matches 0, else 1 for following parameters.
      * Resident\_City
      * Birth\_State
      * Race
      * Ethnicity
      * Party
      * Gender
* Feature table was calculated for both of positive and negative pairs.
* Initial run of **SVM**:
  + Radial SVM classifier, cost = 10, gamma = 1
  + Randomly split data in training to test ratio of 80:20
  + Results:
  + true
  + test 0 1
  + 0 4629 28
  + 1 15 792
* Cross validation:
  + 10 fold cross validation
  + Says best performance is with gamma 0.5. Notice confidence intervals are overlapping.
  + gamma error dispersion (std. dev)
  + 1 0.5 0.005271561 0.001234456
  + 2 1.0 0.006589397 0.001750807
  + 3 1.5 0.007211706 0.001913536
  + 4 2.0 0.007321515 0.001928591
  + 5 2.5 0.007358132 0.001869401
* Specific results are present in log in round2 folder.
* In general, most of the misses are happening for females, who have their last/middle name changed. Probably due to marital status change.
* A few instances are also being observed, when it is extremely likely that the same person has two different voter registration numbers. Therefore, resulting in an error.
* A few of these instances are being recoded.
* **Random Forests:**
  + A random forest is also performed with 4 predictors at each split and 200 trees.
  + Its performance is better than SVM.
  + For specific results, including important predictors, see files inside round2.

1. **6/9/2017 - First Iteration:**

* Used R library e1071 for svm
* Used NC voter databases of district 15, years April 2013 & March 2017
* Only two features have been used. Namely, Levenshtein distance b/w first name and last name across the two data sets.
* Steps:
* Calculated Levenshtein distance for all entities.
* Selection of positive examples: Total of 100. 50 among those with highest distance, 50 with zero distance.
* Selection of negative examples: The 50 entities with highest distance were crossed with each other to create 50\*49 negative cases.
* A total of 2,550 cases were created.
* Out of that, 600 were used as a test set, and rest for training randomly.
* Radial Classifier was used.
* Results:

table(test = yhat\_test, true = nc\_run.test$Res)

True

test 0 1

0 561 11

1 6 22