36-491 HOMEWORK 3

Due: Friday 9/27/13, 5pm, Prof Nugent's mailbox

For this homework, you'll again need to download/install the `RecordLinkage` library.

Our goal is to explore the Fellegi-Sunter probabilistic record linkage model. First download *list1.txt* and *list2.txt* from Blackboard. These are your List A, List B. The records are set up as `RLdata500` where you have name components and birthday components. The last column is an ID number. Our goal is to link pairs of records across the lists. Each record does not necessarily link to a record in the other list.

1. First create a Jaro-Winkler score matrix as done in the second HW. Each row is a pair of records. Use all fields except `fname_c2, lname_c2`, five in total. The last column should indicate match/non-match. Treat all fields as text strings, even the birthdate information. Your final matrix should have 10000 rows and six columns.

2. Defining agreement as having a JW score above 0.90, use the first three fields (first name, last name, and birth year) to find the probabilities of the eight agreement patterns in the group of matches and in the group of non-matches. *Note that you are looking at the probability of the joint vector of 0s and 1s for all fields at once. Also that you are making use of known labels at this time.*

   Calculate the corresponding R values for each of the eight agreement patterns. Without looking at the actual labels, suggest possible upper and lower threshold values for determining match, non-match, and clerical review.

   Now use your thresholds to label your record pairs matches, non-matches, and clerical review. Compare your assigned labels to the true labels (looking at the matches and non-matches only, ignore clerical review).
   How many mistakes did you make? What kind were they?

   Describe the clerical review record pairs (if any).
   Can you tell why they were the "in-between" ones?

3. Repeat (2) using conditional independence instead to find the R values (using appropriate upper, lower thresholds). Also compare your results to (2).
   How did your results change by assuming conditional independence (if anything)?

4. We can use our estimated probabilities, R values, and upper/lower thresholds to predict links for other data sets if we believe that they are appropriate.

   Download *list3.txt* and *list4.txt* from Blackboard. These two lists are similar in form including an id variable. For each set of R values from (2) and (3), label these record pairs as matches, non-matches or clerical review.

   Compare your assigned labels (the matches, non-matches) to the true labels. How did your previously estimated probabilities, etc do with these two lists? Did conditional independence hurt or help?

   Were your clerical review record pairs similar in nature to your previous clerical review record pairs?

Optional/Just for Fun:
In the previous questions, we just worked with three fields (eight agreement patterns). Try using four fields (can assume conditional independence).
Did you improve your classification?

Continuing the Optional Fun:
The E-M algorithm described in class on Monday for a binary agreement/conditional independence model is actually fairly straightforward to code. See 9.4 of Herzog for details, specifically p.99.

For example, let's look at Lists 1,2 and the first three fields.

- Start by computing initial estimates for $\{m_1, m_2, m_3, u_1, u_2, u_3, P(M)\}$
  What could you choose? Jaro suggests that your m estimates should be greater than your u estimates.
- For each record pair, calculate an estimate of $\hat{g}$ (formula on p.99)
- Then given your $\hat{g}$ vector, re-calculate $\{\hat{m}_1, \hat{m}_2, \hat{m}_3, \hat{u}_1, \hat{u}_2, \hat{u}_3, \hat{P}(M)\}$
  (formulas also on p.99)
- Alternate the previous two steps until you see almost no change in the $\{\hat{m}_1, \hat{m}_2, \hat{m}_3, \hat{u}_1, \hat{u}_2, \hat{u}_3, \hat{P}(M)\}$ values *(coding this is probably easier than sitting and watching the values change)*

Now use your final $\hat{g}$ vector to decide what is a link and what's not.
Given that you know the true labels, how did the E-M do?