6.872/HST950 Problem Set 3
Handed out: March 24, 2003
Due: Thu., April 3, 2003

1. Identification Problem

In an earlier lecture, we had outlined a "theory of record linkage" (the full paper is linked from our class schedule page) that tells us, in principle, how to do probabilistic matching of various features of two objects in order to decide whether they are likely to be the same object. Briefly, the theory is as follows. I have interspersed questions for you to answer with the description.

1. Given two purported objects (e.g., patients), o_1 and o_2, it is either the case that $o_1=o_2$ or that they are distinct individuals. For example, our records contain a patient file for Raul P. Szolovits of 123 Main Street, Boston, MA 02131; a new patient arrives claiming to be Peter Szolovits of 123 Main Street, Boston, MA 02113.
2. Among all the observations we might make of o_1 and o_2, we select a certain set of features $f_i(o)$ that we agree will be of interest. For example, we might choose last name, first and middle names, street address, city, and ZIP code.
3. For each pair of features $f_i(o_1)$, $f_i(o_2)$, we can compare the probability that one would observe $f_i(o_1)$, $f_i(o_2)$ in either of the two cases of step 1.
   For example, assuming that half the hospital's patient population have home addresses in Boston, then $P(f_{city}(\text{Raul}), f_{city}(\text{Peter})|\sim\text{same})$ is ½ * ½. = ¼. By contrast, if these two records belong to the same person, then we would just expect that the probability that that person lives in Boston is ½. Thus, the likelihood ratio
   $\frac{p(Boston, Boston \mid same)}{p(Boston, Boston \mid \sim same)} = \frac{1/2}{1/4} = 2$. Further, if 1% of people in the city live on Main St, then $\frac{p(Main, Main \mid same)}{p(Main, Main \mid \sim same)} = \frac{0.01}{0.01^2} = 100$. We may get an additional likelihood ratio of 1000 (say) for the address, 123, and another factor of, say, 75, for both states being MA. These are both estimates, and answer the question what fraction of all addresses is 123, or what fraction of individuals like in MA. If our initial database contains records on 1M individuals, then we might argue that the *a priori* odds are essentially $\frac{p(same)}{p(\sim same)} = 1/1M = 10^{-6}$. If we assume conditional independence of each of the feature pairs from each other e.g., if we believe that you are no more likely to get matching street numbers on Main Street than on Sunset Boulevard, then the posterior estimate is $10^{-6} \cdot 2 \cdot 100 \cdot 75 = 0.015$.

   I have added tables **lastnames**, **malenames** and **femalenames** to the cwsscrubbed database on Singapore in case you are interested in playing with name frequency tables downloaded from the census bureau last year. Each table has four columns:
   - Name, in all-caps
   - Freq, the percent of the (relevant) population that has this name
   - Cum, the cumulative percent of the relevant population that has this or a more frequent name

- N, the ordinal index of this name in its table.

E.g., SMITH is the most common last name, accounting for 1.006% of the population. JOHNSON is next, being 0.810%. The two most common names thus account for 1.816% of the population. AALDERINK is the 88799th most common last name (and the last one in this table), and accounts, at this resolution, for 0% of the population. A little over 9.5% of the population have last names (like mine) that are even less common, and not listed. JAMES, JOHN, ROBERT, MICHAEL and WILLIAM are the five most common male names, being almost 15% of males, and MARY is the most common female name (2.629%), followed by PATRICIA, LINDA, BARBARA and ELIZABETH. The top five account for less than 7%, suggesting more diversity in female names. Indeed it takes 1219 male names to cover 90% of the males, but 4275 female names to cover 90% of the females.

4. Treating the mismatches on first name and ZIP code correctly are challenging problems. For example, the fact that Raul P.'s middle initial is the same as the starting letter of Peter might suggest that his official name is Raul Peter, but he sometimes goes by Peter. One needs a theory of how to estimate the likelihood of this, compared to the obvious possibility that these are simply different people.

   Q1: Explain how you would come up with an estimate of this likelihood ratio.

   Similarly, though the ZIP codes are different, one might turn into the other by a simple transposition, which is probably a common transcription error.

   Q2: Give an estimate of the likelihood ratio for this pair of ZIP codes, and explain how you would estimate that ratio from data you might have available to you or data you might be able to acquire in a small study.

   In any case, we would expect that both likelihood ratios are smaller than 1.0; i.e., the mismatches are more likely if these are different individuals than if they are the same. Therefore, if we aggregate these with the previous calculation, the result is less than 1.5% likelihood (really odds) that these are the same individual.

The assumption of conditional independence among pairs of (mis)matching features is not really appropriate under some circumstances, no matter how convenient it may be. For example, different ethnic groups tend to have different last names (e.g., you might be more tempted to look for my ancestry among Central Europeans than Chinese, Hawaiians, Welsh or Hispanics). But the distribution of first names often follows similar ethnic patterns. Therefore, if you compare two records each with the name "Raul Gonzales", the likelihood ratio should almost certainly not be as high as the product of the likelihood ratios for "Raul" and for "Gonzales". Intuitively, once I learn that two records have a Hispanic last name, then a further match on a Hispanic first name should be less impressive than that same further match would be in conjunction with a Slavic last name (because that combination is much more rare). The census bureau (www.census.gov) does not, to my knowledge, publish statistics on name distributions in different ethnic groups or on the correlations between first and last names.

Q3: What kind of mathematical model would you make to allow you a first-order adjustment for the non-independence of matches among first and last names based on ethnicity?

Q4: List three other potential sources of non-independence that might be appropriate to take into account in implementing a patient identification system based on probabilistic matching.

Q5. How to be a CIO

You have heard two different CIO's, both at large Boston area integrated delivery networks, talk about their conception of their jobs (either implicitly or explicitly). Please compare and contrast the views of Drs. John Glaser and John Halamka as they presented their ideas in class. *Note: Your answer should be brief (no more than a page), but interesting and insightful.*