6.872/HST950 Problem Set 7
Due 10/28/2004

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.

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.

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.

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 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, 1.5, 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

$$\frac{p(same)}{p(\sim same)} = \frac{1}{1M} = 10^{-6}$$

.

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 $10^{-6} \bullet 2 \bullet 100 \bullet 1000 \bullet 1.5 = 0.3$. (This estimation is based on the information including city, street name, street address, and states, but not including other information such as names.)

Q1:
We have two records as follows. Please estimate the likelihood ratio that these are actually of the same person. You can assume any probabilities you need for the calculation, but please suggest (you don't have to actually do it) a practical way to estimate these probabilities, from any resource available to the hospital, or from a small study. The way by which you estimate these probabilities is as important as, if not more than, the likelihood ratio calculation. Assume independency between the variables for now.

Variables	*Record1*	*Record2*
First name	*Jim*	*Jim*
Last name	*Smith*	*Smith*
Gender	*M*	*M*
Address	*123 Main St.*	*123 Main St.*
City	*Newton*	*Newton*
State	*MA*	*MA*

In reality, we must also consider the effect of typos and errors. For instance, a zip code 02113 can be miswritten as 02131, a common transcription error. Another possible problem is that the same person may use different first names, e.g. Dave and David, or people may be referred to by their middle names. Also sometimes people put middle names/initials on the registration form, but sometimes they don't. Treating such mismatches can be challenging problems.

Q2:
With addition information of date of birth as following, what will be the likelihood ratio? Ignore the possibilities of typos and errors for now.

Variables	*Record1*	*Record2*
Date of birth	*Jan. 2, 1956*	*Jan. 2, 1965*

Q3:
Consider the possible typo that the dates of birth can actually be the same but with a transposition error. Please re-estimate the likelihood ratio of the above two records. You can assume any probabilities you need for the calculation, but please suggest a practical way to estimate these probabilities ,including the possibility of transposition errors, from any resource available to the hospital, or from a small study. The way by which you estimate these probabilities is as important as, if not more than, the likelihood ratio calculation. Assume independency between the variables.

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.

If such distributions are available, however, we can make a first-order adjustment of such dependencies with a simple Bayesian model. Assuming P (first | ethnic) and P (last | ethnic) are conditionally independent given the ethnic group, we have

$$P(first, last) = \sum_{Ethnic} \mathrm{P}(\mathrm{first} \mid \mathrm{ethnic}) * \mathrm{P}(\mathrm{last} \mid \mathrm{ethnic}) * \mathrm{P}(\mathrm{ethnic})$$

.

Q4:

There are quite a few possible dependencies among the variables used in our example, besides that of first names and last names based on ethnics. Please list at least three of them, and suggest for each one a mathematical model to make a first-order adjustment. You can use similar model as in the ethnic group example, but other reasonable and valid models will be as much appreciated.