Basic theory of entity resolution

- Classical record linkage formalism,
- key simplifying assumptions,
- links to modern techniques,
- supremely bland presentation template.

A model for independent decisions

Classical Fellegi-Sunter (1969) record linkage model is as follows – consider two sets of records A and B, with elements denoted a and b respectively. Product set $A \times B$ is somehow partitioned into matches M and non-matches U.

Pairs (a, b) in M typically agree on attributes such as first name, last name, components of date of birth and address. Pairs in U may have isolated random agreements on some of these. The aim is then to recover M while only having access to record attributes.

It should be noted that this set-up allows multiple matches between elements of A and B with uncertain interpretation. Further restrictions on the structure of M, however, require modifications to the basic "theory".

Comparison space

We to denote the vector of "agreement codes" as:

$$\gamma(a,b) = \big(\gamma_1(a,b), \gamma_2(a,b), \ldots, \gamma_n(a,b)\big),$$

where $\gamma_i(a,b)$ might be an indicator corresponding to statements like "name is the same", "name is the same and is Brown", "name disagrees", "name missing on one record", "agreement on city part of address but not the street".

All possible realisations of agreement codes form the *comparison space* Γ, i.e. $\gamma(a,b) \in \Gamma$ for all $(a,b) \in A \times B$.

Local decision rules

A randomised *decision rule* or *linkage rule* is then the mapping

$$d\big(\gamma(a,b)\big) : \Gamma \to \big\{P\big(d_i \,|\, \gamma(a,b)\big) \,|\, i = 1,2,3\big\}$$

which assigns a distribution, i.e. $\sum_{i=1}^{3} P\big(d_i \,|\, \gamma(a,b)\big) = 1$, over three possible decisions $\{d_1, d_2, d_3\}$ to each element of Γ.

Here d_1 denotes $(a,b) \in M$ (a *positive link*), d_3 denotes $(a,b) \in U$ (a *positive non-link*) and d_2 is a *possible link*.

To reiterate this is not the right formalism if we want to impose restricitions on the structure of match set M (e.g. one to one), as in these cases decisions can no longer be taken for individual pairs (a,b) in isolation.

Constructing the decision rule

To construct the "optimal" decision rule we define the probabilities of observing γ for a match $(a, b) \in M$:

$$m\big(\gamma(a,b)\big) = P\big(\gamma(a,b) \,|\, (a,b) \in M\big)$$

as well as a non-match $(a, b) \in U$:

$$u\big(\gamma(a,b)\big) = P\big(\gamma(a,b) \,|\, (a,b) \in U\big).$$

It is a straightforward application of Neyman-Pearson theory to show that the "optimal" decision rule with respect to the usual objectives:

$$P(d_1 \,|\, U) = \sum_{(a,b) \in A \times B} u\big(\gamma(a,b)\big) P\big(d_1 \,|\, \gamma(a,b)\big) \text{ and}$$

$$P(d_3 \,|\, M) = \sum_{(a,b)\in A\times B} m\big(\gamma(a,b)\big) P(d_3 \,|\, \gamma(a,b)\big),$$

namely one that generates fewest expected false matches for a given expected number of undetected matches (as we are dealing with a bivariate objective), is a function of the likelihood ratio $\frac{m(\gamma)}{u(\gamma)}$:

$$d\big(\gamma(a,b)\big) = \begin{cases} (1,0,0), & T_1 \leq \frac{m(\gamma)}{u(\gamma)} \\ (0,1,0), & T_2 < \frac{m(\gamma)}{u(\gamma)} < T_1 \\ (0,0,1), & \frac{m(\gamma)}{u(\gamma)} \leq T_2. \end{cases}$$

Simplifying assumptions

This is almost entirely unhelpful as we don't know probabilities m or u. The original proposal of Newcombe et al. (1959) was to make some heroic independence assumptions about the structure of $P\big(\gamma(a,b) \,|\, (a,b) \in M\big)$:

$$m\big(\gamma(a,b)\big) = \prod_{i=1}^{n} P\big(\gamma_i(a,b) \,|\, (a,b) \in M\big),$$

where $\gamma_i(a,b)$ are individual components of the agreement vector and similarly for $u(\gamma)$.

Notice that the log likelihood ratio can then be written as:

$$\log\left(\frac{m(\gamma)}{u(\gamma)}\right) = \sum_{i=1}^{n} \log\big(m_i(\gamma_i(a,b))\big) - \sum_{i=1}^{n} \log\big(u_i(\gamma_i(a,b))\big).$$

With some additional hand waving, you can convince yourself that if γ_i is the simple agreement indicator for a certain binary attribute ξ, such as the presence of particular surname or a given month of birth:

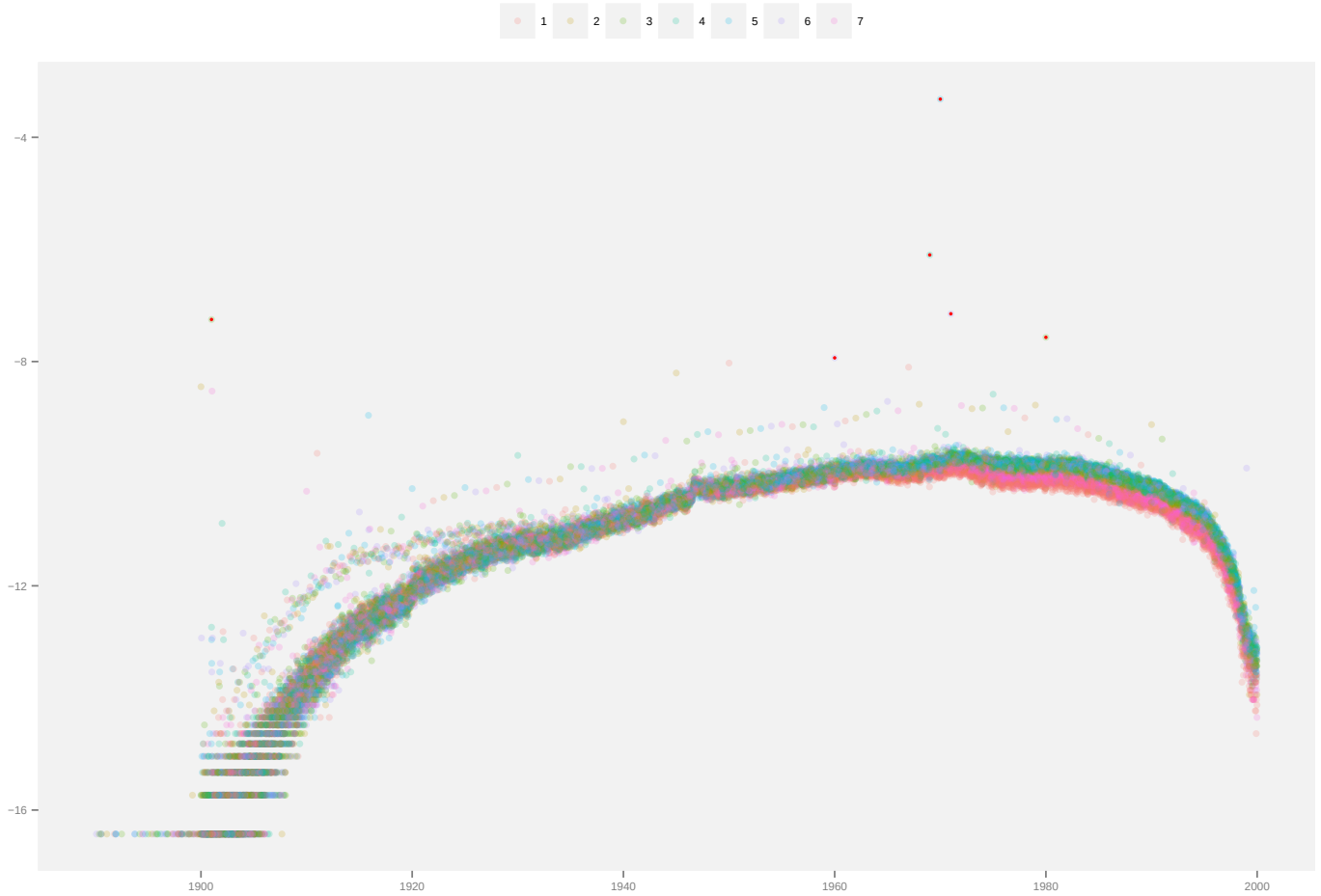
$$\gamma_i(a,b) = \begin{cases} 1, & \text{if } \xi(a) = 1 \text{ and } \xi(b) = 1, \\ 0, & \text{otherwise} \end{cases}$$

it might be reasonable to expect that when $\gamma_i\big((a,b)\big) = 1$:

$$\log\big(m_i(1)\big) - \log\big(u_i(1)\big) \approx \log\big(\pi_\xi\big) - \log\big(\pi_\xi^2\big) = -\log\big(\pi_\xi\big),$$

where π_ξ is the proportion of the combined population where the attribute is present. This gives us an intuitive scheme to assign weights to different matches, where e.g. rare names would be considered more informative than common ones.

CRODS date of birth log frequencies by day of the week



Connection to modern methods

Observe that the likelihood ratio can be approximated *directly*, i.e. :

$$\frac{m(\gamma)}{u(\gamma)} \approx \frac{P\big((a,b) \in M \,|\, \gamma\big)}{P\big((a,b) \in U \,|\, \gamma\big)} \frac{|U|}{|M|},$$

where the two components are estimated directed via probabilistic classifiers. Obvious downside of this is the hard requirement for "training data".