Module 7: fastlink, Part I

Rebecca C. Steorts

Reading

- Binette and Steorts (2020)
- Edmorando et al. (2020)
- Fellegi and Sunter (1969)

Probabilistic Entity Resolution

While Fellegi and Sunter (1969) have provided a framework for probabilistic entity resolution, there are few implmentations that scale to large data sets.

Agenda

- We review fastlink, Edmorando et al. (2020)
- We illustrate a toy example on `RLdata10000`

fastlink

- Edmorando et al. (2020) developed fastlink a scalable implementation of the FS method.
- In addition, the authors incorporated auxiliary information such as population name frequency and migration rates.
- The authors used parallelization and hashing to merge millions of records in a near real-time on a laptop computer, and provided open-source software of their proposed methodology.

Agreement Patterns

- Assume two data sets (A and B) with overlapping variables in common (such as name, gender, address, etc.)
- Define an agreement value in field k for record pair (i, j) :

$$\gamma_k(i,j) = \begin{cases} \texttt{agree} \\ \\ \texttt{disagree} \end{cases}$$

Agreement Patterns

	First	Last	Age	Street
Data set $\mathcal{A}$				
1	James	Smith	35	**Devereux St.**
Data set $\mathcal{B}$				
7	James	Smit	43	**Dvereux St.**
	agree	agree	disagree	**agree**

Agreement Patterns

	First	Last	Age	Street
Data set $\mathcal{A}$				
1	`James`	`Smith`	`35`	**`Devereux St.`**
Data set $\mathcal{B}$				
7	`James`	`Smit`	`43`	**`Dvereux St.`**
	`agree`	`agree`	`disagree`	**`agree`**

Agreement pattern $\gamma(i,j) = \{\gamma_1(i,j), \gamma_2(i,j), \ldots, \gamma_K(i,j)\}$

Agreement Patterns

	First	Last	Age	Street
Data set $\mathcal{A}$				
1	James	Smith	35	**Devereux St.**
Data set $\mathcal{B}$				
7	James	Smit	43	**Dvereux St.**
	agree	agree	disagree	**agree**

Agreement pattern $\gamma(i, j) = \{\gamma_1(i, j), \gamma_2(i, j), \ldots, \gamma_K(i, j)\}$

One computational bottleneck is calculating these agreement patterns.

Agreement Patterns

- We **observe** the agreement patterns $\gamma(i, j)$
- We **do not observe** the matching status

$$C_{i,j} = \begin{cases} \texttt{non-match} \\ \\ \texttt{match} \end{cases}$$

fastLink Model

$$
\begin{aligned}
C(i,j) &\overset{\text{iid}}{\sim} \text{Bernoulli}(\mu) \\
\gamma(i,j) \mid C(i,j) = \text{non-match} &\overset{\text{iid}}{\sim} \mathcal{F}(\pi_{\text{NM}}) \\
\gamma(i,j) \mid C(i,j) = \text{match} &\overset{\text{iid}}{\sim} \mathcal{F}(\pi_{\text{M}}),
\end{aligned}
$$

where λ, π_{M}, π_{NM} are estimated via the EM algorithm

fastLink Model

More formally, we write

$$\begin{aligned} C(i,j) &\stackrel{\text{iid}}{\sim} \text{Bernoulli}(\mu) \\ \gamma(i,j) \mid C(i,j) &\stackrel{\text{iid}}{\sim} \text{Categorical}(\pi), \end{aligned}$$

`fastLink` Model

Independence assumptions:

1. Independence across pairs
2. Conditional Independence across linkage fields:

$$\gamma_k(i,j) \perp \gamma_{k'}(i,j) \mid C(i,j).$$

Log-likelihood

$$\log L(\lambda, \pi \mid \gamma(i,j)) \tag{1}$$

$$= \prod_{i=1}^{N_1} \prod_{i=1}^{N_2} \{\lambda \prod_{k=1}^{K} \prod_{\ell=1}^{L_k - 1} \pi_{Mk\ell}^{I(\gamma_k(i,j)=\ell)} + (1-\lambda) \prod_{k=1}^{K} \prod_{\ell=1}^{L_k - 1} \pi_{N_M k\ell}^{I(\gamma_k(i,j)=\ell)}\} \tag{2}$$

Exercises

Show the E and M steps.