Module 8: Bayesian Fellegi and Sunter

Rebecca C. Steorts

Reading

- Binette and Steorts (2020)
- Sadinle (2014)

Load Libraries

```
#install.packages("exchanger_0.2.0.tar.gz", repos = NULL, t
library(exchanger)
```

Duplicate detection

Duplicate detection is the task of finding sets of records that refer to the same entities within a data file.

Overview of Bayesian Fellegi and Sunter

- Sadinle (2014) proposes a Bayesian variant of Fellegi and Sunter (1969).
- In order to do this, he uses the same basic set up as the original authors, but puts priors on the m and u probabilities such that one has a posterior to approximate.
- This is appealing as one is able to incorporate prior knowledge into the record linkage process.
- We also, get an approximate posterior distribution, allowing use to know the posterior variance of the sample mean and full posterior distributions of evaluation metrics. Such uncertainty quanification is NOT possible from a probabilistic point of view.
- Finally, we can also do a fully unsupervised analyst, which Sadinle performs in his case study on the EI Salvadorian data set.

Notation

Assume there are a total of n records in a database.

Assume there is one database with r records labeled

$$\{1, 2, \ldots, r\}$$

where more than one record can refer to the same entity.

Assume that $n \leq r$.

Thus, we can view this problem as partitioning the database into n groups of matches/non-matches.

Representation of partitions

A partition of a set is a collection of nonempty and non-overlapping subsets whose union is the original set.

Sadinle (2014) refers such subsets **groups** or **cells.**

Example

Suppose the database has five records total $\{1, 2, 3, 4, 5\}$.

One potential partition can be represented by the following three groups:

$$\{1, 3\}, \{2\}, \{4, 5\}.$$

Each group represents an underlying entity.

In this example, records 1,3 are co-referent; records 4,5 are co-referent, and record 2 is a singleton record.

Co-reference matrix

A partition can also be represented by a matrix.

Consider the matrix Δ of dimension $r \times r$,

where

$$\Delta_{ij} = \begin{cases} 1, & \text{if records i,j are co-referent} \\ 0, & \text{otherwise.} \end{cases}$$

Δ is referred to as the co-reference matrix.

Δ is symmetric with only ones in the diagonal.

Labellings of the partition's groups

Unfortunately, it is not computationally efficient to utilize the co-reference matrix in practice.

An alternative is to use arbitrary labelings of the **partition's groups**.

Labellings of the partition's groups

Assume that r the maximum number of entities possibly represented in the database.

Define

$$Z_i = q, \quad i = 1, \ldots, r$$

if record i represents entity q, $\quad 1 \leq q \leq r$.

$$Z = (Z_1, Z_2, \ldots, Z_r)$$

contains all the records labels.

Thus,

$$\Delta_{ij} = I(Z_i = Z_j).$$

Example (Continued)

Recall our database has $\{1, 2, 3, 4, 5\}$ records and the partition can be represented by the three groups:

$$\{1, 3\}, \{2\}, \{4, 5\}.$$

What are some examples of vectors Z that would correspond to the partitions above?

Example (Continued)

The labelings

$$Z = (1, 2, 1, 3, 3)$$

or

$$Z = (4, 1, 4, 2, 2)$$

would correspond to this partition because both $Z_1 = Z_3$ and $Z_4 = Z_5$ and Z_2 gets its own value.

Connection to the rth Bell number

The number of ways in which a data file with r records can be partitioned is given by the rth Bell number, which grows rapidly with r.

See Rota (1964).

Comparison data

- Comparison data are obtained by comparing pairs of records, with the goal of finding evidence of whether two records refer to the same entity.
- Intuitively, two records referring to the same entity should be very similar.

Notation

Assume f features.

Compare features f of records (i, j) by computing some similarity measure $S_f(i, j)$.

The range of $S_f(i, j)$ is divided into $L_f + 1$ intervals

$$I_{f0}, I_{f1}, \ldots, I_{fL_f},$$

which represent levels of disagreement.

By convention, I_{f0} denotes the **highest level of agreement** and I_{fL_f} denotes the **lowest level of agreement**.

Notation

For records (i, j) define

$$\gamma_{ij}^{f} = \ell \quad \text{if} \quad S_f(i, j) \in I_{f\ell}.$$

- The larger the value of γ_{ij}^{f} the large the disagreement between records (i, j) with respect to feature f.
- These feature comparisons are collected into a vector for each record pair denoted by

  $$\boldsymbol{\gamma}_{ij} = (\gamma_{ij}^{1}, \gamma_{ij}^{2}, \ldots \gamma_{ij}^{F}),$$

  which denotes the comparison vector for records (i, j), where F is the number of features being compared.

Model for the Comparison Data

Assume that the comparison vector $\boldsymbol{\gamma}_{ij}$ is a realization of a random vector $\boldsymbol{\Gamma}_{ij}$.

The set of record pairs is composed of two types – co-referent and non co-referent record pairs.

Model for the Comparison Data

One expects the distribution of the comparison vectors $\mathbf{\Gamma}_{ij}$ of co-referent/non co-referent record pairs to be quite different.

For example, one expects to observe more agreements among co-referent pairs than among non0coreferent pairs.

Similarly, one expects many more disagreements among non-coreferent pairs than among co-referent pairs.

Model for the Comparison Data

This intuition can be formalized by assuming that the distribution of $\mathbf{\Gamma}_{ij}$ is the same for all record pairs that refer to the same entity and is the same for all record pairs that refer to different entities.

This is modeled as follows:

$$\mathbf{\Gamma}_{ij} \mid \Delta_{ij} = 1 \overset{iid}{\sim} G_1 \tag{1}$$

$$\mathbf{\Gamma}_{ij} \mid \Delta_{ij} = 0 \overset{iid}{\sim} G_0 \tag{2}$$

for all $i, j \in P$, where G_1, G_0 represent the models of the comparison vectors for pairs that are coreferent and noncoreferent, respectively.

Prior distribution on the coreference partition

The co-reference matrix reprents a partition of the entries of Δ.

Let D represent the set of possible co-reference partitions.

The author assumes a uniform prior that assigns equal probability to each partition in D.

This can be written as

$$\pi(\Delta) \propto I(\Delta \in D).$$

Prior distribution on the coreference partition

One simple way to obtain this prior for Z is to assign equal probability to each of the $r!/(r-n)!$ labelings of a partition with n cells/groups.

This leads to the prior

$$p(\boldsymbol{Z}) \propto \frac{(r - n(\boldsymbol{Z}))}{r!} I(\boldsymbol{Z} \in Z)$$

where $n(\boldsymbol{Z})$ measure the number of different labelings of $\boldsymbol{Z}$.

A model for independent comparison fields

We describe a simple parametrization for G_1 and G_0.

This model assumes that the comparison fields are independent for both co-referent and non co-referent records.

Assuming that Γ_{ij}^f takes $L_f + 1$ values corresponding to levels disagreement, one can model

$$\Gamma_{ij}^f \mid \boldsymbol{m}_f \sim \text{Multinomial}(\boldsymbol{m}_f),$$

where

$\boldsymbol{m}_f = (m_{f0}, m_{f1}, \ldots, m_{f,L_f-1}).$

A model for independent comparison fields

Similarly,

$$\Gamma_{ij}^{f} \mid \boldsymbol{u}_f \sim \text{Multinomial}(\boldsymbol{u}_f),$$

where

$$\boldsymbol{u}_f = (u_{f0}, u_{f1}, \ldots, u_{f,L_f-1}).$$

A model for independent comparison fields

Following the notation of Sadinle (2014),

$\Phi_0 = (u_1, \ldots, u_F)$ and $\Phi_1 = (m_1, \ldots, m_F)$ such that

$\Phi_f = (m_f, u_f)$.

Prior speciification for the model parameters

Sadinle (2014) specified that

$$m_{fl} \sim \text{TruncatedBeta}(\alpha^1_{f\ell}, \beta^1_{f\ell}, \lambda_{f\ell}, 1)$$

for $\ell = 2, \ldots, L_f - 1$.

In addition,

$$m_{f0} \sim \text{Beta}(\alpha^1_{f0}, \beta^1_{f0}).$$

Prior speciification for the model parameters

Sadinle (2014) specified that

$$u_{f\ell} \sim \texttt{Uniform}(0, 1)$$

for all features and disagreement.

The author stated that one might take

$$u_{f\ell} \sim \text{Beta}(\alpha^0_{f0}, \beta^0_{f0})$$

if prior information is available.

Gibbs sampling

In order to approximate the joint posterior of $\boldsymbol{Z}$ and Φ, one must use a Gibbs sampler, where details can be found in Sadinle's 2014 paper.

Review these details on your own.

Summarize

Provide summary.