Module X: Bayesian Fellegi and Sunter

Rebecca C. Steorts

Reading

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

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

Give an overview of the framework

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 inefficient to utilized 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,

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

Back to our Example

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

Back to our Example

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

or

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

would correspond to this partition because

both $Z_1 = Z_3 = Z_4 = Z_5$ and Z_2 gets its own value.