

Machine Learning for Record Linkage at Statistics Canada

Presented at the HLG-MOS-ML meeting on Oct. 2020



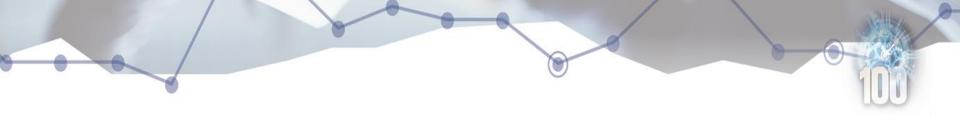
Delivering insight through data for a better Canada



Acknowledgements

- Gautier Gissler
- Tanvir Quadir
- Abdelnasser Saidi





Content

- Record linkage at Statistics Canada
- Probabilistic linkage with G-LINK
- Threshold problem
- An unsupervised solution
- Future work



- Many linkages of social or business data are routinely performed.
- In most cases, there are no training data.
- Various methods are used, which are deterministic, hierarchical or probabilistic.
- Probabilistic linkages are implemented with G-LINK.



- G-LINK is the agency system for probabilistic record linkage.
- A similarity score (the linkage weight) is assigned to each record pair.
- A pair is automatically linked if this score exceeds a threshold.



- Manually setting the G-LINK threshold is labor intensive.
- Question: How to automate this step when there are no training data?
- Answer: Consider an unsupervised solution.

An Unsupervised Solution

- Use two-means clustering to set the threshold.
- How does this work?
 - 1. Map each pair to features based on the similarity scores (the rules weights) for the different variables.
 - 2. Place each pair in one of two initial clusters.
 - 3. In an iteration, update the clusters centroids and place each pair in the cluster with the nearest centroid according to the Euclidian distance.
 - 4. Repeat until convergence.



An Unsupervised Solution (cont'd)

- Performance
 - Competitive relative to expectation-maximization based on a log-linear mixture under conditional independence (Fellegi and Sunter, 1969).
- Challenges
 - Choosing the initial clusters.
 - Imbalanced cluster sizes.



- The solution is available in G-LINK v3.4 and later versions.
 - See the macro %ml_classification().
- A related solution (Christen, 2007) combines the clustering procedure with a probit model.
 - 1. Use two-means to classify the pairs.
 - 2. Train the probit model on the classified pairs.



 Currently looking at the Python Record Linkage Toolkit (De Bruin, 2019).



THANK YOU!

abel.dasylva@canada.ca



- Christen, P. (2007). "A two-step Classification to Unsupervised Record Linkage", in Proceedings of the 6-th Australian Conference on Data Mining and Analytics, 70, 111-119.
- De Bruin, J. (2019). "Python Record Linkage Toolkit: A toolkit for record linkage and duplicate detection in Python". *Zenodo*.
 - https://doi.org./10.5281/zenodo.3559043
- Fellegi, I.P., and Sunter, A.B. (1969), "A theory of record linkage", Journal of the American Statistical Association, 64, 1183–1210.