

Evaluation Metrics for Entity Resolution

STA 325: Homework 2

General instructions for homeworks: Your code must be completely reproducible and must compile. No late homeworks will be accepted.

Reading Read the paper Binette and Steorts (2022) to get an overview of entity resolution. You'll want to refer to this during the course of the semester as it's meant to be a quick reference regarding the concepts that we will be covering. For more details, refer to the book by Christen (2012).

Advice: Start early on the homeworks and it is advised that you not wait until the day of as these homeworks are meant to be longer and treated as case studies.

Commenting code Code should be commented. See the Google style guide for questions regarding commenting or how to write code <https://google.github.io/styleguide/Rguide.xml>.

R Markdown Test

0. Open a new R Markdown file; set the output to HTML mode and "Knit". This should produce a web page with the knitting procedure executing your code blocks. You can edit this new file to produce your homework submission.

Total points on assignment: 5 (reproducibility) + 25 points for the assignment.

1. (4 points) What are the four main challenges of entity resolution?
2. (4 points, 1 point each) Suppose there are 10 records in a data set. a.) What are the total number of brute-force comparison needed to make all-to-all record comparisons? b.) Repeat this for 100 records, 1000 records, 10,000 records. c.) What do you observe about the number of comparisons that need to be made?
3. Consider the following record linkage data set with 1,000,000 total records that are matched between two databases. Assume that 500,000 are true matches. Assume a classifier (or method) finds 600,000 record pairs as matches, and of these 400,000 correspond as true matches. The number of $TP + FP + TN + FN = 50,000,000$.
 - a. (4 points) Given the information above, find the following information in the confusion matrix: TP, FP, TN, and FN.
 - b. (1 point) Calculate the accuracy. Comment on the reliability of this metric for this problem.
 - c. (1 point) Calculate the precision.
 - d. (1 point) Calculate the recall.
 - e. (1 point) Calculate the f-measure.
 - f. (1 point) Comment on the reliability of the precision, recall, and f-measure for this problem.