



COMP3430 / COMP8430

Data wrangling

Lab 6: Evaluation for Record Linkage

Objectives of this lab

- Today's lab is the fourth in a series of five labs during which we will gradually build a complete record linkage system.
- We will be working with different evaluation measures and learn how they work and why they are important in the RL process.
- Completion of the evaluation module in the overall system.

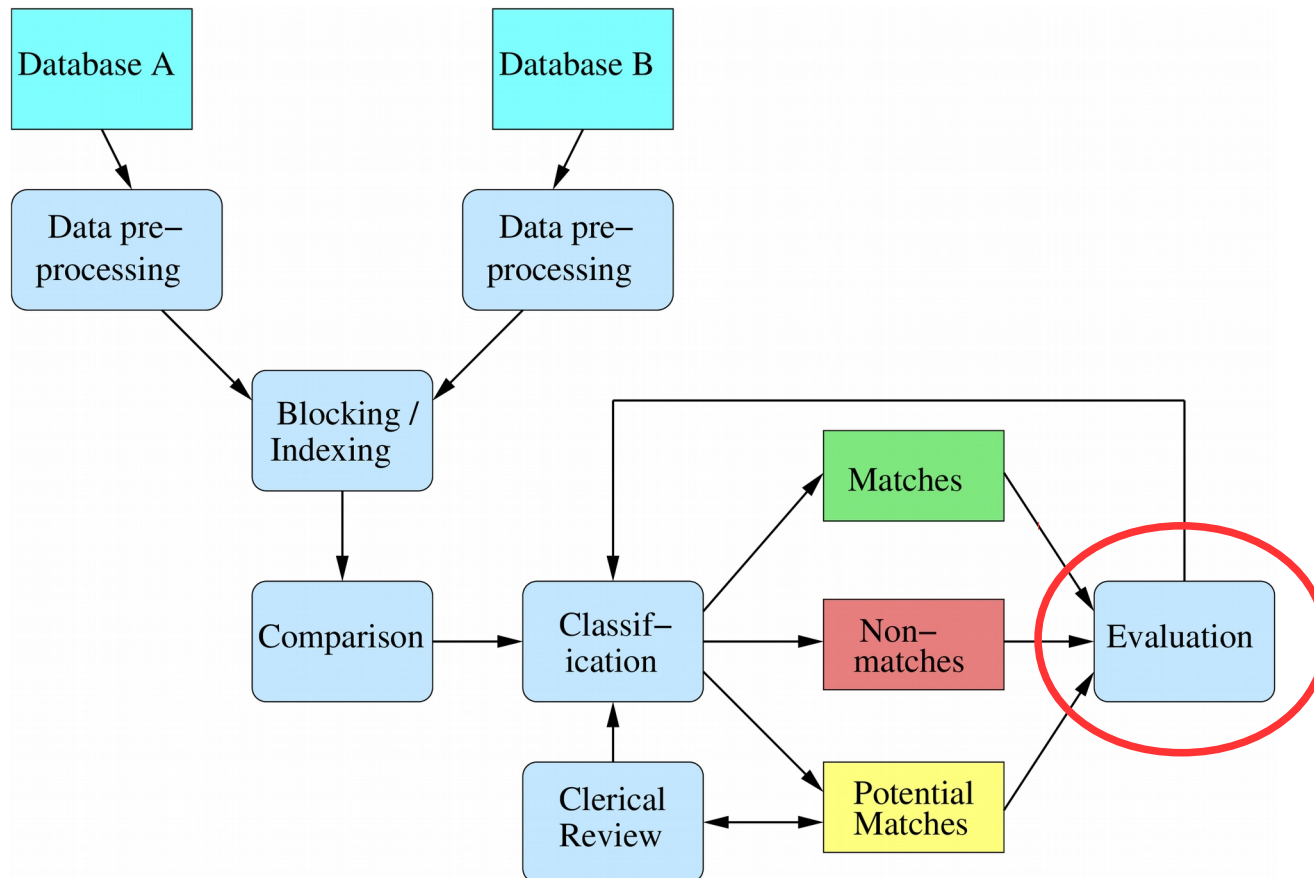
Outline of this lab

- Learn how different evaluation measures work
- Implement different evaluation measures
- Explore and experiment with different evaluation measures
- Summary

Preliminaries

- **Before you begin, aim to review lecture 19 if you have not already viewed them.**
- Go back over the work from lab 5 and remind yourself what we were doing and how the overall program is structured.
- You can download the classification module with sample solutions in week 8 and use it with your RL program if you find difficulties implementing the required classification techniques.

What is evaluation?



- This week we focus on the next step in the linkage process, evaluation.
- The aim of a evaluation metric is to measure the performance of a RL process and see how well it has linked the data sets.
- Why do you think we need different evaluation measures and are they equally important?

How to evaluate a linkage process

- Before we begin let us see how different evaluation metrics work. The evaluation measures are described in lecture 19.

	Predicted Matches	Predicted Non-matches
True Matches	1,000	400
True Non-matches	600	8,000

	Predicted Matches	Predicted Non-matches
True Matches	1,200	200
True Non-matches	800	7,800

- See if you can calculate the following measures from the above two confusion matrices:
 1. Accuracy
 2. Precision
 3. Recall

How to evaluate a linkage process

- Accuracy = $(TP + TN) / (TP + FP + FN + TN)$
= $1000 + 8000 / 10000$
= 0.9

- Precision = $TP / (TP + FP)$
= $1000 / (1000 + 600)$
= 0.625

- Recall = $TP / (TP + FN)$
= $1000 / (1000 + 400)$
= 0.7143

	Predicted Matches	Predicted Non-matches
True Matches	1,000 (TP)	400 (FN)
True Non-matches	600 (FP)	8,000 (TN)

How to evaluate a linkage process

- Accuracy = $(TP + TN) / (TP + FP + FN + TN)$
= $1200 + 7800 / 10000$
= 0.9

- Precision = $TP / (TP + FP)$
= $1200 / (1200 + 800)$
= 0.6

- Recall = $TP / (TP + FN)$
= $1200 / (1200 + 200)$
= 0.8571

	Predicted Matches	Predicted Non-matches
True Matches	1,200 (TP)	200 (FN)
True Non-matches	800 (FP)	7,800 (TN)

Implement different evaluation measures

- Now start looking at **evaluation.py** and explore how the evaluation functions work (inputs, return values, etc.).
- We have already provided two evaluation functions, **accuracy()** and **reduction_ratio()**.
- Run the RL program with different settings and see what the output of these two functions look like and how they perform.
- Now try to implement the other evaluation metrics as required in the lab tutorial document.

Questions to consider

- Are there any measures that are not useful, either because they are always extremely high, or low, or difficult to calculate, etc?
- What is the impact of the data quality on the linkage results? Does this vary depending on which functions you use for the blocking, comparison, and classification steps?
- What effect do the different blocking techniques have on the final record linkage results?
- **Extra task** – Run the RL program with different data sets provided.

Summary

- In this lab we implemented different evaluation measures and learnt how they can be used in the RL program to evaluate it performs.
- Make sure to complete any unfinished work in this module before you come to the next lab.
- In the next lab we will be conducting experiments with more data sets with different sizes and data quality.