COMP3430 / COMP8430
Data wrangling

Lab 5: Classification for Record Linkage

Objectives of this lab

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

Outline of this lab

- Learn how different classification techniques work
- Implement different classification techniques
- Evaluate different classification techniques
- Summary

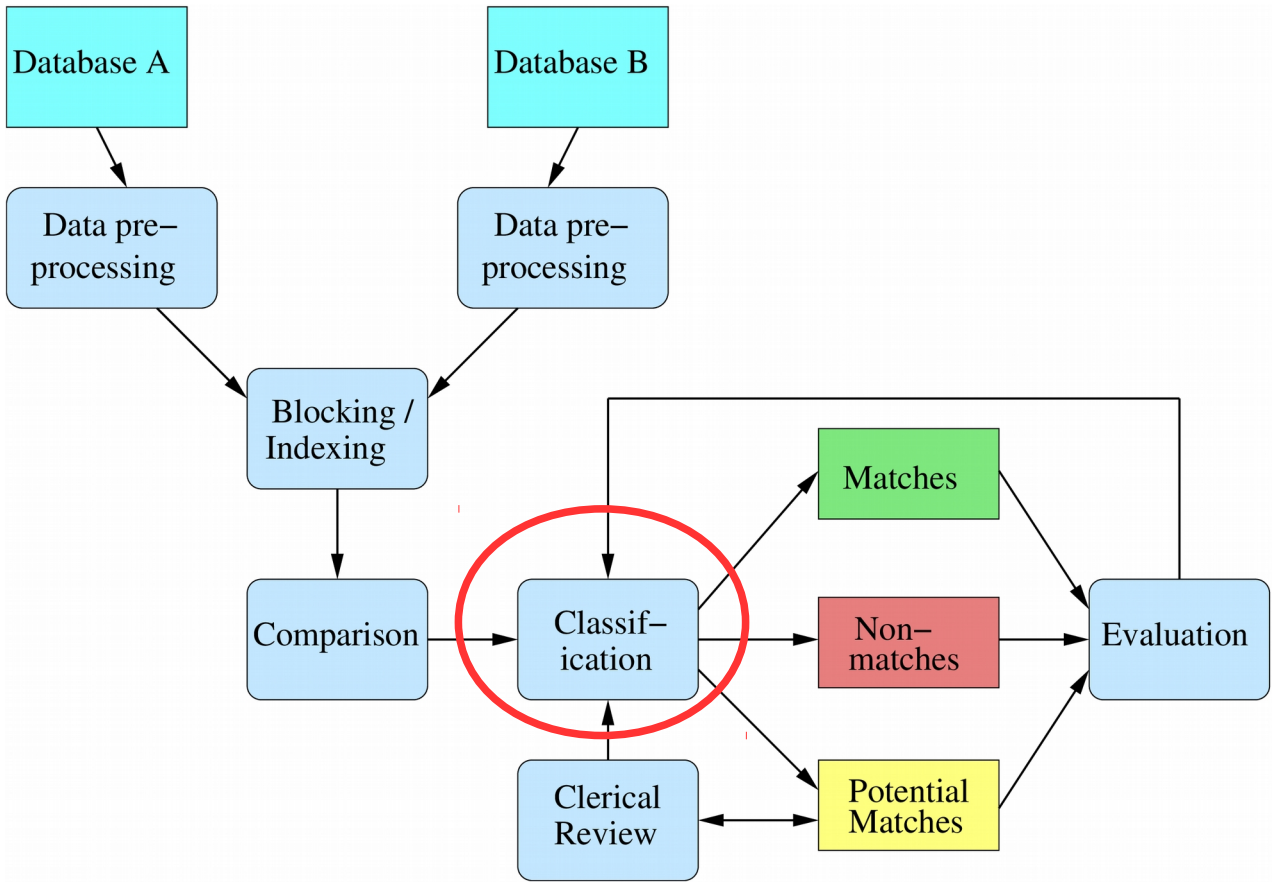
Preliminaries

- **Before you begin, aim to review lectures 17 and 18 if you have not already viewed them.**

- Go back over the work from lab 4 and remind yourself what we were doing and how the overall program is structured.

- You can download the comparison module with sample solutions in week 7 and use with your RL program if you find difficulties implementing the required comparison functions.

What is Classification?



- This week we focus on the next step in the linkage process, classification.
- The aim of a classification technique is to classify a given pair of records as either a match, a non-match, or a potential match.
- Why do you think we need different classification techniques other than exact classification?

How to classify record pairs

- Before we begin let us see how different classification techniques work. The classification techniques are outlined in lectures 17 and 18.

- Let us assume we have the following two vectors of similarities.
  - record pair (r1,r2) resulted in: [0.5, 0.9, 0.2, 0.8, 1.0, 0.0, 0.7]
  - record pair (r3,r4) resulted in: [0.8, 0.7, 1.0, 0.9, 0.6, 0.7, 0.9]

- See if you can compute the classification outcomes of above record pairs using:
  - Threshold based classification with thresholds 0.5 and 0.7
  - Minimum threshold based classification with thresholds 0.5 and 0.7

How to classify record pairs

Similarity vector for record pair (r1,r2): [0.5, 0.9, 0.2, 0.8, 1.0, 0.0, 0.7]

- **Threshold based classification with the threshold 0.5**

  Average similarity = (0.5 + 0.9 + 0.2 + 0.8 + 1.0 + 0.0 + 0.7) / 7 = 0.5857
  Since this average similarity > 0.5, the record pair (r1,r2) is classified as a **match**.

- **Minimum threshold based classification with the threshold 0.5**

  Chech if each similarity value in the similarity vector is at least 0.5.
  Since some similarties are not at least 0.5, the record pair (r1,r2) is classified as a **non-match**.

Implement different classification techniques

- Now start looking at **classification.py** and explore how the classification techniques work (inputs, return values, etc.).
- We have already provided a classification technique, **exactClassify().**

- Run the RL program using this classification technique and see what the output looks like and how it performs**.**

- Now try to implement the other classification functions as required in the lab tutorial document.

Questions to consider

- How many matches do you find with each classification technique for the same threshold value?
- Do different thresholds have an effect on the number of matches?
- How do different weights for the weighted average function influence the number of matches?
- Extra tasks – see if you can change the program to learn the weights for attributes based on their value distributions.

Summary

- In this lab we implemented different classification techniques and learnt how they can be used in the RL program.
- Make sure to complete any unfinished work in this module before you come to the next lab.
- In the next lab we will be looking at how different evaluation measures work and how they can be used in the RL program to evaluate performance.