Unsupervised Learning

Introduction to Unsupervised Learning

1. In a supervised paradigm, we know the answers (or ***ground truth***)

Introduction to Unsupervised Learning

1. In a supervised paradigm, we know the answers (or ***ground truth***)

2. Real world data is rarely labeled, and often lacking in clear structure

Introduction to Unsupervised Learning

1. In a supervised paradigm, we know the answers (or ***ground truth***)
2. Real world data is rarely labeled, and often lacking in clear structure
3. But we can still do powerful analysis without this helping hand:

Introduction to Unsupervised Learning

1. In a supervised paradigm, we know the answers (or ***ground truth***)
2. Real world data is rarely labeled, and often lacking in clear structure
3. But we can still do powerful analysis without this helping hand:
    a. Classify fraudulent transactions
    b. Summarize complex text documents
    c. Develop novel encryption methods

Find groups in the data

- No labels nor response -> unsupervised

- Define groups based on similarity

Group customers, target ads

- A priori, you can’t really put labels on customers
- Group similar customers

You can then try to interpret the grouping, and send targeted ads to the groups

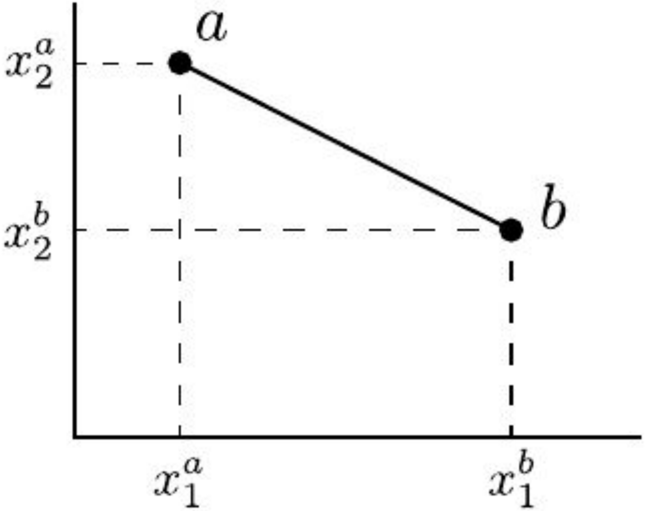
Note: *you may want to assign labels to groups a posteriori*

Defining similarity

After a pre-processing step, you have a data matrix with ***n*** rows (observations) and ***p*** columns (features). Each row is a “point”.

- How to define similarity between points?
- If the features are numerical, we can use euclidean distance
- What if some features are categorical?
    - Ignore
    - Embed into numerical

Euclidean distance



$$d(a,b)^2 = \sum_{i=1:2} (x_i^a - x_i^b)^2$$

Can be generalised from 2-D to n-D