Data Mining & Machine Learning

CS37300
Purdue University

Nov 8, 2023

Lagrange multipliers

- Constrained optimization:

$$\min_x f(x)\ subject\ to\ g(x) = 0$$

- Build the Lagrange equation

$$l(x, \lambda) = f(x) + \lambda g(x)$$

- Loose statement: Stationary points of $l(.)$ are optima for the original constrained problem

Lagrange multipliers -- example

- $f(x) = x$ s.t. $x^2 = 1$

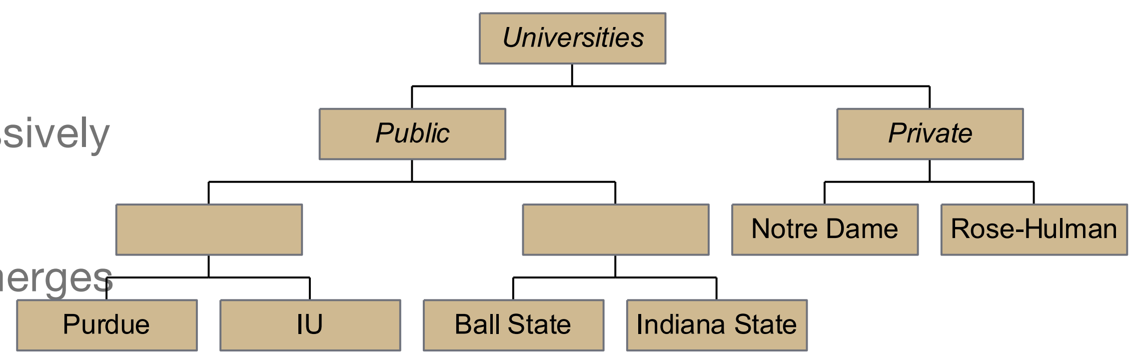
More generally:

- $f(x) = 1^\top x$ s.t. $||x||^2 = 1$
- $f(x) = \sum \log x_i$. s.t. $1^\top x = 1$

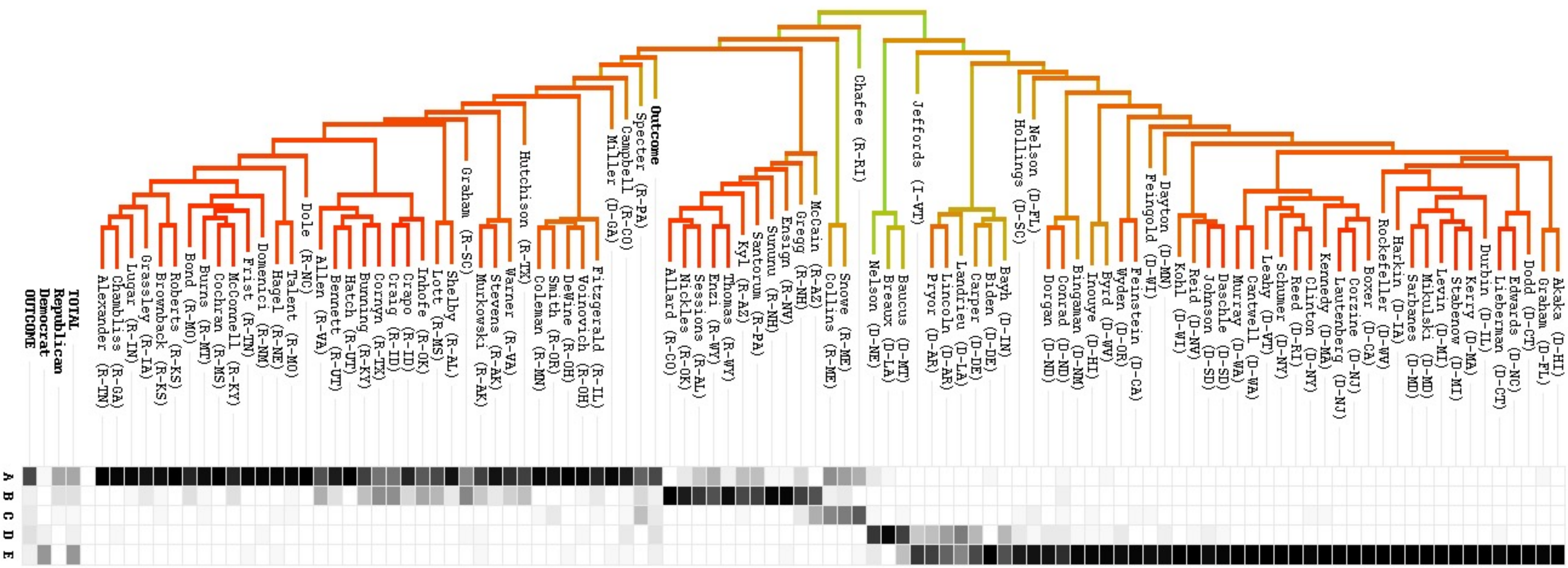
Hierarchical methods

- Construct a hierarchy of nested clusters rather than picking k beforehand
- Approaches:
    - Agglomerative: merge clusters successively
    - Divisive: divided clusters successively
- Dendrogram depicts sequences of merges or splits
    - Can use height to indicate distance

University Clustering



Clustering Represented with Dendrogram



Agglomerative

- For i = 1 to n:
  - Let $C_i = \{x_i\}$. $C = \{C_i\}$
- While target granularity is not reached:
  - Let C_i and C_j be the pair of clusters with min $D(C_i, C_j)$
  - $C_i = C_i \cup C_j$
  - Output C_i
  - Remove C_j from C

Divisive

- Let $C_0 = \{x_i\}$
- Divisive(C):
  - Output C as a cluster
  - If $|C|>1$
    - Divide C into C_1, C_2 with max $D(C_1,C_2)$
    - Divisive(C_1)
    - Divisive(C_2)

Distance measures between clusters

- Single-link/nearest neighbor:
  - $Dist(C_i, C_j) = \mathbf{min}\{\ d(x,y) \mid x \in C_i,\ y \in C_j\ \}$
  - Can produce long thin clusters
- Complete-link/furthest neighbor:
  - $Dist(C_i, C_j) = \mathbf{max}\{\ d(x,y) \mid x \in C_i,\ y \in C_j\ \}$
  - Particularly sensitive to outliers
- Average link:
  - $Dist(C_i, C_j) = \mathbf{avg}\{\ d(x,y) \mid x \in C_i,\ y \in C_j\ \}$

Agglomerative/Divisive: How to compute?

Agglomerative

- Exhaustive?
  - n^2 possibilities at first step
  - Shrinks as we go
  - But computing distance becomes more complex

Divisive

- Exhaustive?
  - Exponential possibilities at the start ($O(2^n)$)
- Heuristic solutions: Greedy
  - Choose a “high distance” point as start of new cluster
  - Move remaining points to what maximizes the distance

Hierarchical Summary

Agglomerative

- Knowledge representation?
  - Sequence of cluster merges
- Score function?
  - min/max/avg of distance/similarity
- Search?
  - Exhaustive possible

Divisive

- Knowledge representation?
  - Sequence of cluster divisions
- Score function?
  - min/max/avg of distance/similarity
- Search?
  - Greedy heuristic

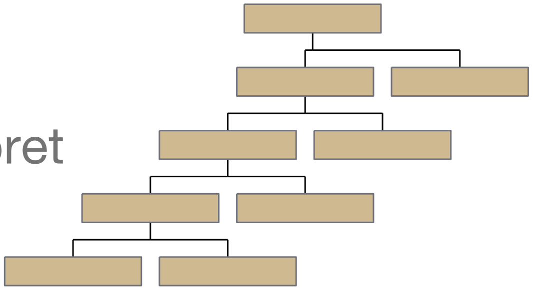
Hierarchical Summary

Advantages

- Can discover odd-shaped clusters
- No need to set number of clusters
- Natural, informative visualization

Disadvantages

- Sensitive to outliers
- Non-obvious choice of parameters
- Unclear when to terminate
- May give hard-to-interpret results



Pattern discovery

Pattern discovery

- Models describe entire dataset (or large part of it)
- Pattern characterize local aspects of data
- Pattern: predicate that returns “true” for the instances in the data where the pattern occurs and “false” otherwise
- Task: find descriptive associations between variables

Examples

- Supermarket transaction database
    - 10% of the customers buy wine and cheese
- Telecommunications alarms database
    - If alarms A and B occur within 30 seconds of each other then alarm C occurs within 60 seconds with p=0.5
- Web log dataset
    - If a person visits the CNN website, there is a 60% chance the person will visit the ABC News website in the same month

Pattern in tabular data

- Primitive pattern: subset of all possible observations over variables $X_1,\ldots,X_d$
  - If X_k is categorical then $X_k = c$ is a primitive pattern
  - If X_k is ordinal then $X_k \leq c$ is a primitive pattern
- Start from primitive patterns and combine using logical connectives such as AND and OR
  - age<40 AND income<100,000
  - chips=1 AND (beer=1 OR soda=1)

Pattern space

- Set of legal patterns; defined through set of primitive patterns and operators to combine primitives
    - Example: If variable $X_1,\ldots,X_d$ are all binary we can define the space of patterns to be all conjunctions of the form
      $(X_{i1}=1)$ AND $(X_{i2}=1)$ AND ... AND $(X_{ik}=1)$

- Typically there is a generalization/specialization relationship between patterns
    - Pattern α is **more general** than pattern β, if whenever β occurs, α occurs as well. This also means that pattern β is **more specific** than pattern α
    - Examples:
      age<40 AND income<100,000 is more ***specific*** *than age<40*
      chips=1 is more ***general*** *than chips=1 AND (beer=1 OR soda=1)*
    - This property is used during search

Pattern discovery task

- Find all “interesting” patterns in the data
- Approach: find all patterns that satisfy certain conditions
- Challenge: find the right balance between
    - Pattern complexity
    - Pattern accuracy
    - Computational complexity