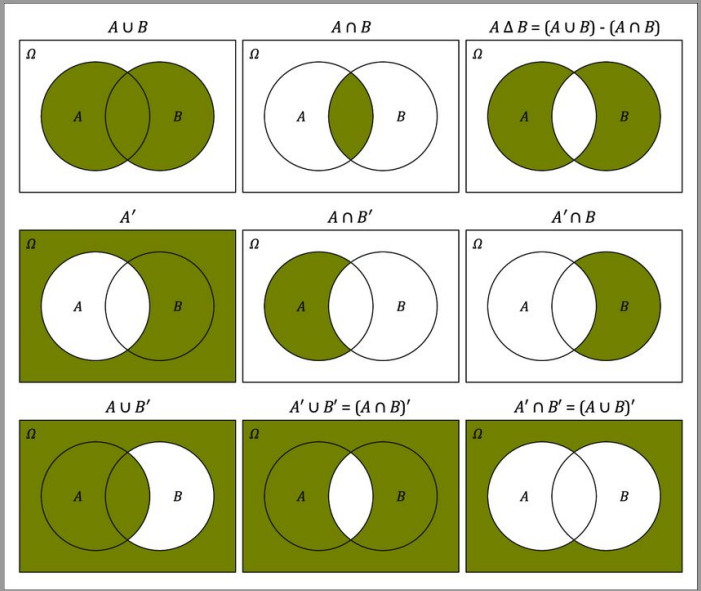
Optimisation problems



AI ITMO course 2024

Optimisation problems: standard techniques

Example 1.

Let $A=\{a_1,\ldots,a_n\}$ and $m<n$
Find a subset S of A such that:

(i) The cardinal of S is m (constraint)
(ii) The sum of elements in S is maximal (optimization criterion)

Remark. X = the set of all 2^n subsets of A
A brute force approach is of $O(2^n)$.

Optimisation problems: greedy techniques

Let's reformulate the optimization problem as follows:

Let $A=(a_1,\ldots,a_n)$ be a multiset (a set of not necessarily distinct elements).
Find $S=(s_1,\ldots,s_k)$such that S satisfies some constraints and optimizes a certain criterion.

Basic idea of greedy technique:

- S is constructed successively starting with the first element
- At each step a new element (that element which seems to be the best at that moment) is selected from A.
- Once a choice is made it is final (the greedy approach at each step takes the currently best element, without regard for future consequences; there are no back steps to make corrections)

Optimisation problems: greedy techniques

Pseudocode example

The general structure of a greedy algorithm:

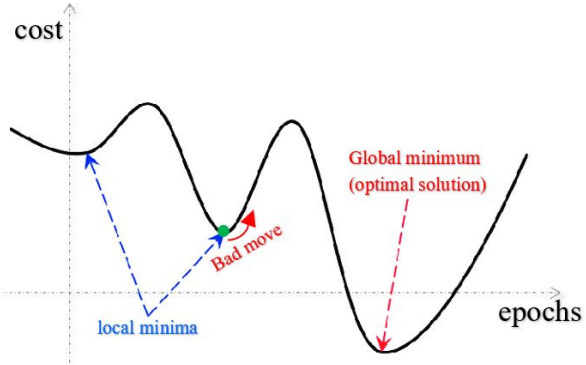
```
Greedy(A)
  S←Ø
  WHILE “S is not completed” AND “there exist unselected elements in A”
        DO
         “choose the best currently available element a from A”
         IF “by adding a to S the constraints are satisfied”
         THEN “add a to S”
  RETURN S
```

Optimisation problems: greedy techniques

Pseudocode example

The general structure of a greedy algorithm:

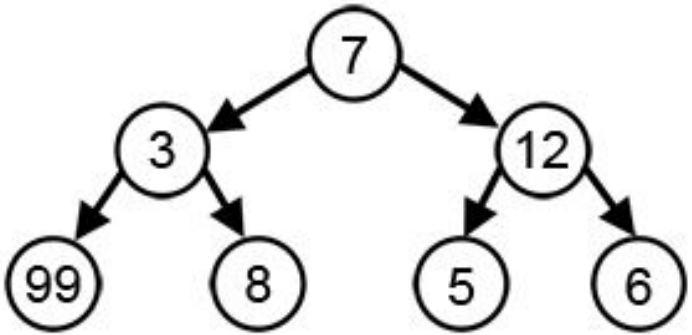
```
Greedy(A)
  S←Ø
  WHILE "S is not completed" AND "there exist unselected elements in A"
        DO
         "choose the best currently available element a from A"
         IF "by adding a to S the constraints are satisfied"
         THEN "add a to S"
  RETURN S
```



P.S. The selection criterion is frequently based on some heuristics (= technique based on experiential data and intuition rather than on theoretical analysis).

Optimisation problems: greedy techniques

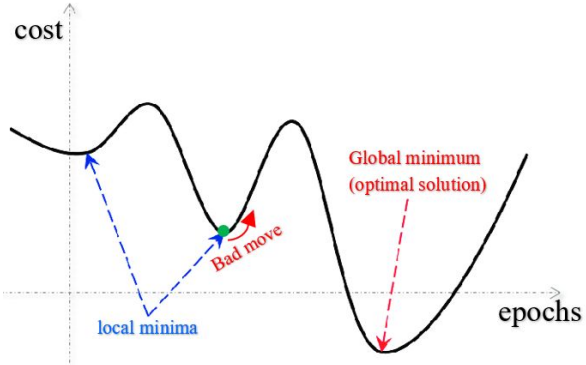
Are there algorithms, **where greedy algorithms won't lead to globally good solutions**? Yes, e.g. some problems do not have optimal solutions, in cases of weighted graphs algorithms, when the structure of the set on which the algorithm is running is not a matroid.



Optimisation problems: greedy techniques

Are there algorithms, where greedy algorithms will lead to globally good solutions?

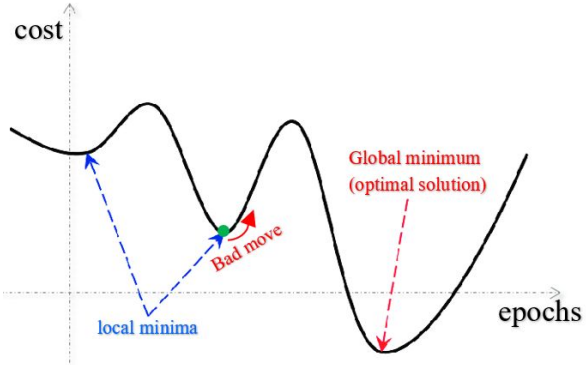
```
Subset(A[1..n],m)  //variant1
FOR i ← 1,m DO
  k ← i
  FOR j ← i+1,n DO
     IF A[k]<A[j] THEN k ← j  ENDIF
  ENDFOR
  IF k<>i THEN A[k]↔A[i] ENDIF
  S[i] ← A[i]
ENDFOR
RETURN S[1..m]
```



Optimisation problems: greedy techniques

Are there algorithms, where greedy algorithms will lead to globally good solutions?

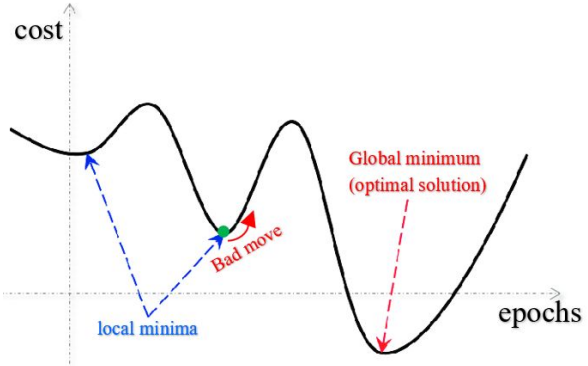
```
Subset(A[1..n],m)  //variant1
FOR i ← 1,m DO
  k ← i
  FOR j ← i+1,n DO
     IF A[k]<A[j] THEN k ← j  ENDIF
  ENDFOR
  IF k<>i THEN A[k]↔A[i] ENDIF
  S[i] ← A[i]
ENDFOR
RETURN S[1..m]
```



Materials on the optimisation problems

Are there algorithms, where greedy algorithms will lead to globally good solutions?
Lecture notes on the estimation of the complexity of the algorithm
https://myweb.uoi.gr/bekos/uploads/Lecture_Notes_Algorithms.pdf

Lecture notes and lecture on the Introduction to Approximation Algorithms" (PDF). Advanced Algorithms (2IL45) — Course Notes. TU Eindhoven. Archived (PDF) from the original on 2022-10-09



Projects preparations

Examples of libraries for spatial analysis: osmnx (getting data, visualisation), scikitlearn (classification analysis), networkx (network analysis, EDA for spatial analysis methods), iduedu (connectivity patterns)

Identification of problems for network analysis from optimisation tasks:

Pipeline of steps in your notebook, find the template here:

1) Extraction of the raw data (from osm - osmnx or mapbox) which has many layers of information or use already existing data, data cleaning, data preprocessing - **format as geojson, csv**
2) Formalisation of the research question from the list here optimisation tasks:
   a) landuse specification (given some specific limitations of the territories),
   b) optimal transportation e.g. network connectivity tasks
3) Formalisation of the research questions related to network science/optimisation algorithms for 2):
   a) given your assigned labels of the tiles in your land, assign labels to the rest of tiles (classification task)
   b) optimal transportation e.g. network connectivity tasks networkx - **format of the data is graph G**

Projects preparations

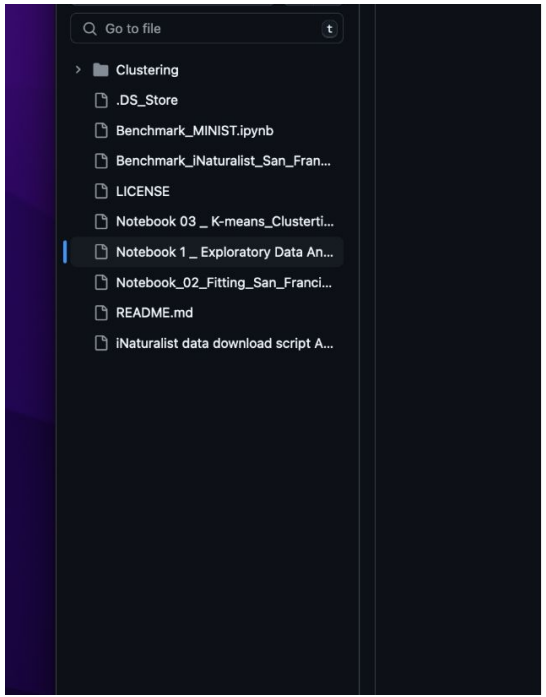
Example of the project network connectivity tasks**:**

1) Extraction of the raw data (from osm - osmnx or mapbox) which has many layers of information or use already existing data, data cleaning, data preprocessing - **format as geojson, csv**
2) Formalisation of the research question from the list here optimisation tasks:
   network connectivity tasks
3) Formalisation of the research questions related to network science/optimisation algorithms for 2):
   optimal transportation e.g. network connectivity tasks networkx - **format of the data is graph G**

   Studying traditional network centrality measures like:
   Identification of bottlenecks in graph G = nodes in a graph with high betweenness centrality
   Suggest some of the possible improvements, e.g connecting additional nodes to the network
4) Assess the complexity of your algorithm

Example of sustainability related projects

Github notebook from the data analysis courses and citizen science projects




```
Loading:  Francisco_Bay 2022
Loading:  Francisco_Bay 2023
```

Out[4]:

	id	observed_on_string	observed_on	time_observed_at	created_time_zone	created_at	
0	20069	1:15 pm.	2016-07-14	2016-07-14T13:15:00-07:00	America/ Los_Angeles	2011-06-03T14:51:45-07:00	2020-0
1	20070	1:00 pm.	2016-03-25	2016-03-25T13:00:00-07:00	America/ Los_Angeles	2011-06-03T14:53:13-07:00	2020-0
2	68373	6:30	2016-02-12	2016-02-12T06:30:00-08:00	America/ Los_Angeles	2012-04-20T20:36:48-07:00	2020-0
3	158736	2:19	2016-10-14	2016-10-14T14:19:00-07:00	America/ Los_Angeles	2012-12-06T20:23:52-08:00	2016-1
4	538018	2016-04-10 2:20:00 PM PDT	2016-04-10	2016-04-10T14:20:00-07:00	America/ Los_Angeles	2014-02-20T15:40:40-08:00	2016-0

5 rows × 38 columns

Dataset comprises of 4214727 observations and 38 characteristics.

In [10]:
```
dfall.shape
```

Out[10]:
```
(4214727, 38)
```