review of clustering algorithms

DBSCAN, expectation-maximization, k-means

nico casale

NCSU

2017/11/29

outline

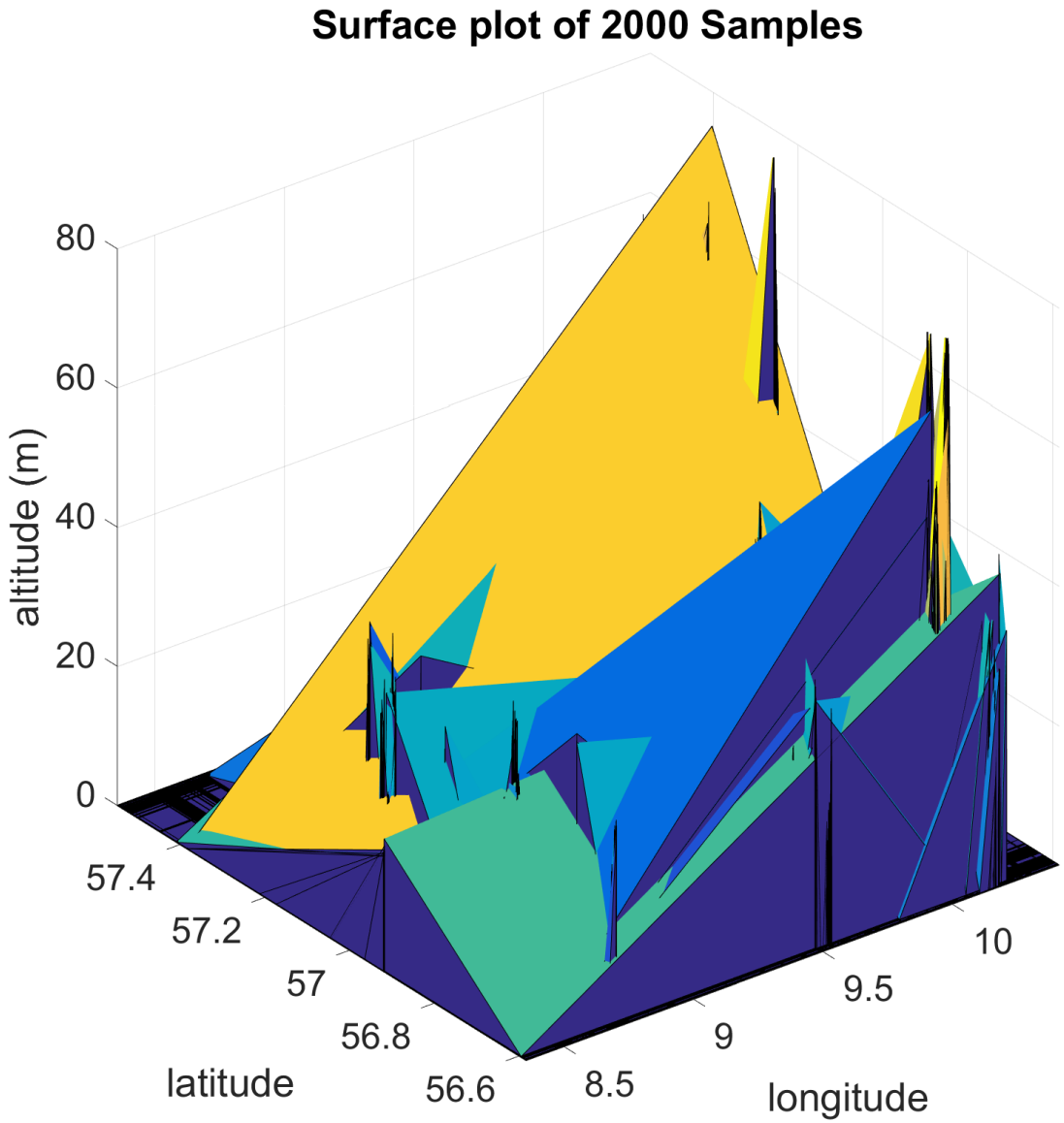
motivation

- labels for supervised machine learning
- identify patterns in data
- variety of clustering algorithms to choose from
- applications: gene sequencing, medical imaging, social networks, etc.

dataset

- source: UCI machine learning repository
- provided: Dr. Manohar Kaul of Aarhus University
- longitude & latitude (degrees), altitude (meters)
- ∼430k locations in North Jutland, Denmark [1]
- used first 2000 points– covers ∼4,500 sq. mi.

dataset

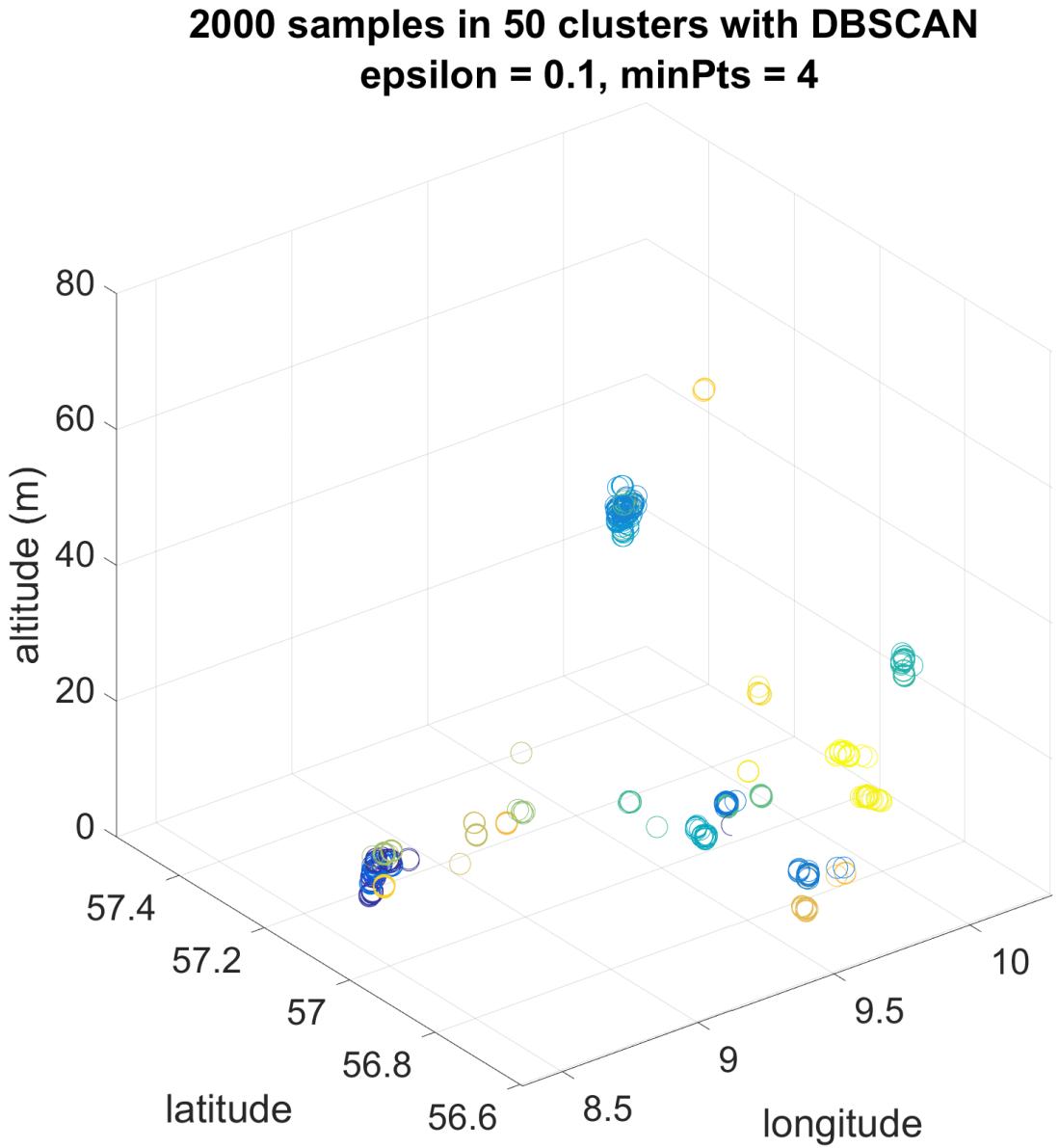

Surface plot of 2000 Samples
altitude (m)
80
60
40
20
0
57.4
57.2
57
56.8
56.6
latitude
8.5
9
9.5
10
longitude

progress

DBSCAN

- clusters dense regions of points
- requires three steps [2][3]
  1. find neighbors (using ϵ) of each point.
     assign points with $\geq$ *minPts* neighbors as *core points*.
  2. connect *core points* with mutual neighbors.
  3. assign *non-core* points to a cluster if ϵ-neighbor, otherwise ignore.

DBSCAN


2000 samples in 50 clusters with DBSCAN
epsilon = 0.1, minPts = 4
altitude (m)
80
60
40
20
0
57.4
57.2
57
56.8
56.6
latitude
8.5
9
9.5
10
longitude

progress

EM Algorithm

- assuming mixture of Gaussians, maximum-likelihood estimate (MLE) is the marginal likelihood

$$\mathcal{L}(\Theta; X) = p(X|\Theta) = \int p(X, Z|\Theta) dZ \tag{1}$$

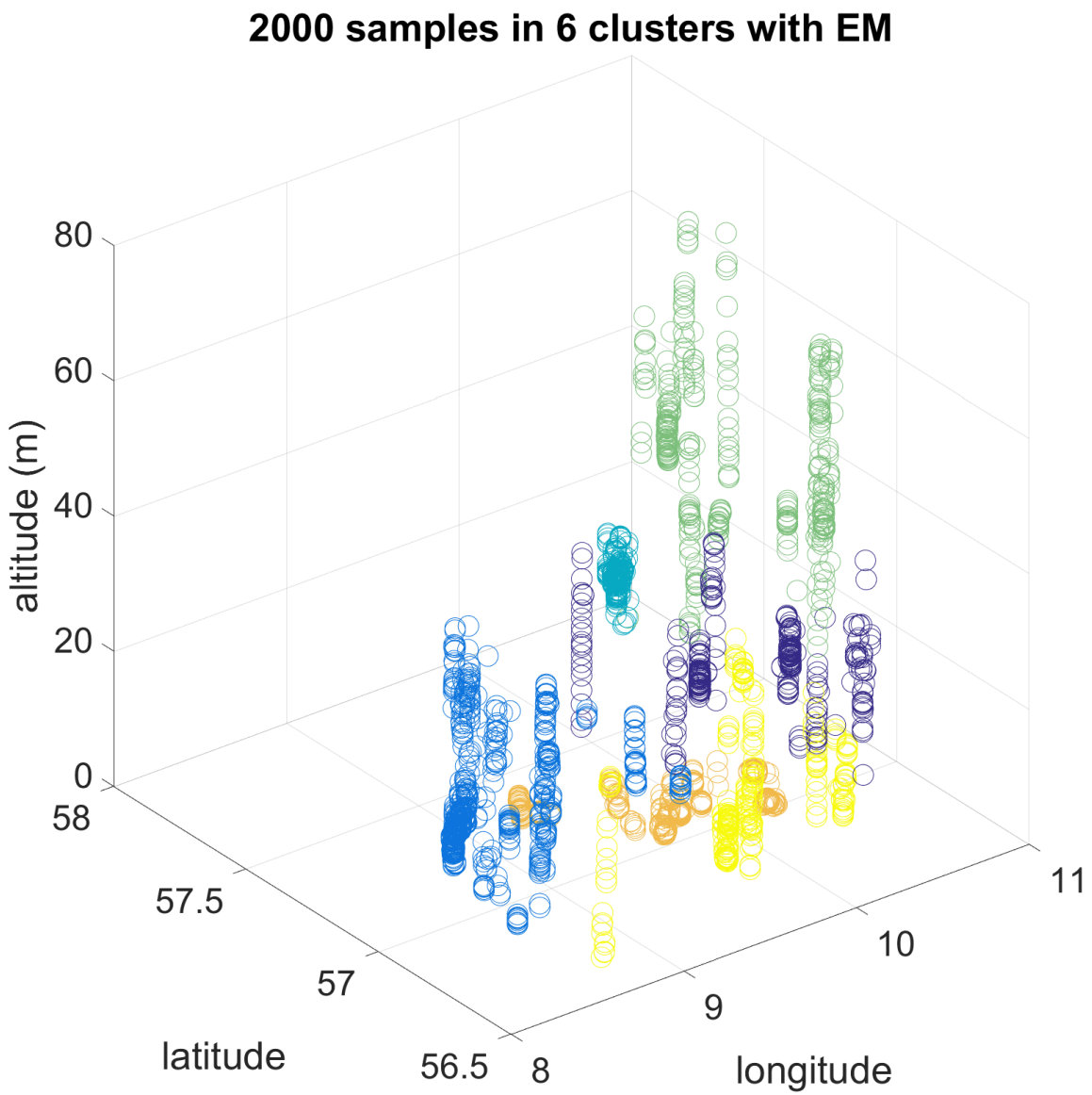
- approximates the MLE over iterations of two steps [4]
  1. expectation (E): find expected value of log-likelihood *w.r.t.* $P(Z|X, \Theta^{(t)})$

$$Q(\Theta|\Theta^{(t)}) = E_{Z|X,\Theta^{(t)}}[\log \mathcal{L}(\Theta; X, Z)] \tag{2}$$

     where t is the iteration, Θ are unknown parameters, $\Theta^{(t)}$ their estimate, X the observed data, and Z the unobserved data.
  2. maximization (M): find parameters that maximize

$$\Theta^{(t+1)} = \arg\max_{\Theta} Q(\Theta|\Theta^{(t)}) \tag{3}$$

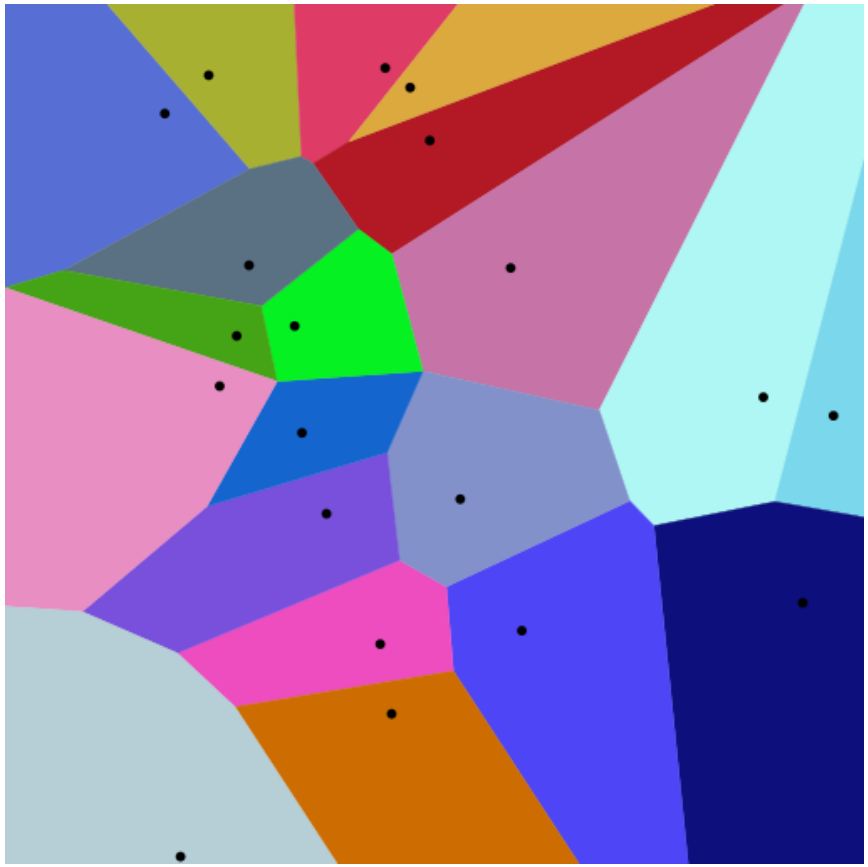
EM Algorithm



progress

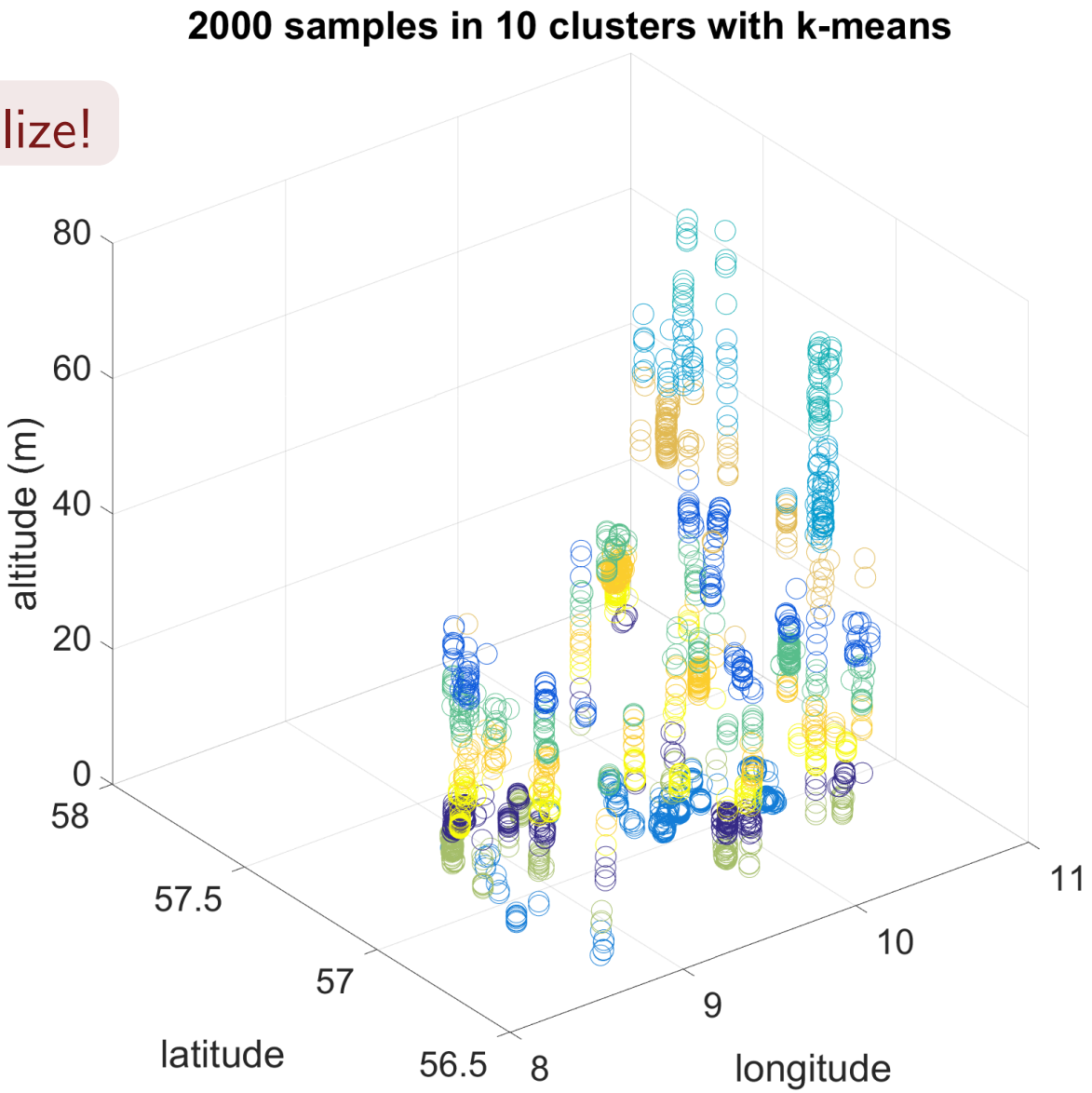
k-means

- partitions the input data into Voronoi cells
- equidistant from the k centroids/clusters [5]



k-means

need to normalize!



conclusion

- intuition and heuristics needed for each algorithm
- choice of algorithm depends on features of interest
- future work:
  - alternative distance metrics
  - normalize input data
  - parallelize algorithms

references


Manohar Kaul.
3d road network (north jutland, denmark) data set.
In *Picmdm (IEEE MDM)*, 2013.


Wikipedia.
Dbscan — wikipedia, the free encyclopedia, 2017.
[Online; accessed 28-November-2017].


Siddarth Agrawal.
Machine learning - dbscan, 2013.
[Online; accessed 28-November-2017].


Wikipedia.
Expectation maximization algorithm — wikipedia, the free encyclopedia, 2017.
[Online; accessed 29-November-2017].


Wikipedia.
K-means clustering — wikipedia, the free encyclopedia, 2017.
[Online; accessed 2-October-2017].

thanks!