K-Means Clustering

MATH 3220

Supplemental Presentation

by John Aleshunas

Outline

- Algorithm Definition
- Algorithm Fitness Function
- The Algorithm Implementation
- Issues
- Some Examples

Algorithm Definition

- The K-Means algorithm is an method to cluster objects based on their attributes into k partitions.
- It assumes that the *k* clusters exhibit Gaussian distributions.
- It assumes that the object attributes form a vector space.
- The objective it tries to achieve is to minimize total intra-cluster variance.

Source: Wikipedia

Algorithm Fitness Function

The K-Means algorithm attempts to minimize the squared error for all elements in all clusters.

The error equation is:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2$$

Where E is the sum of the square error for all elements in the data set; p is a given element; and m_i is the mean of cluster C_i

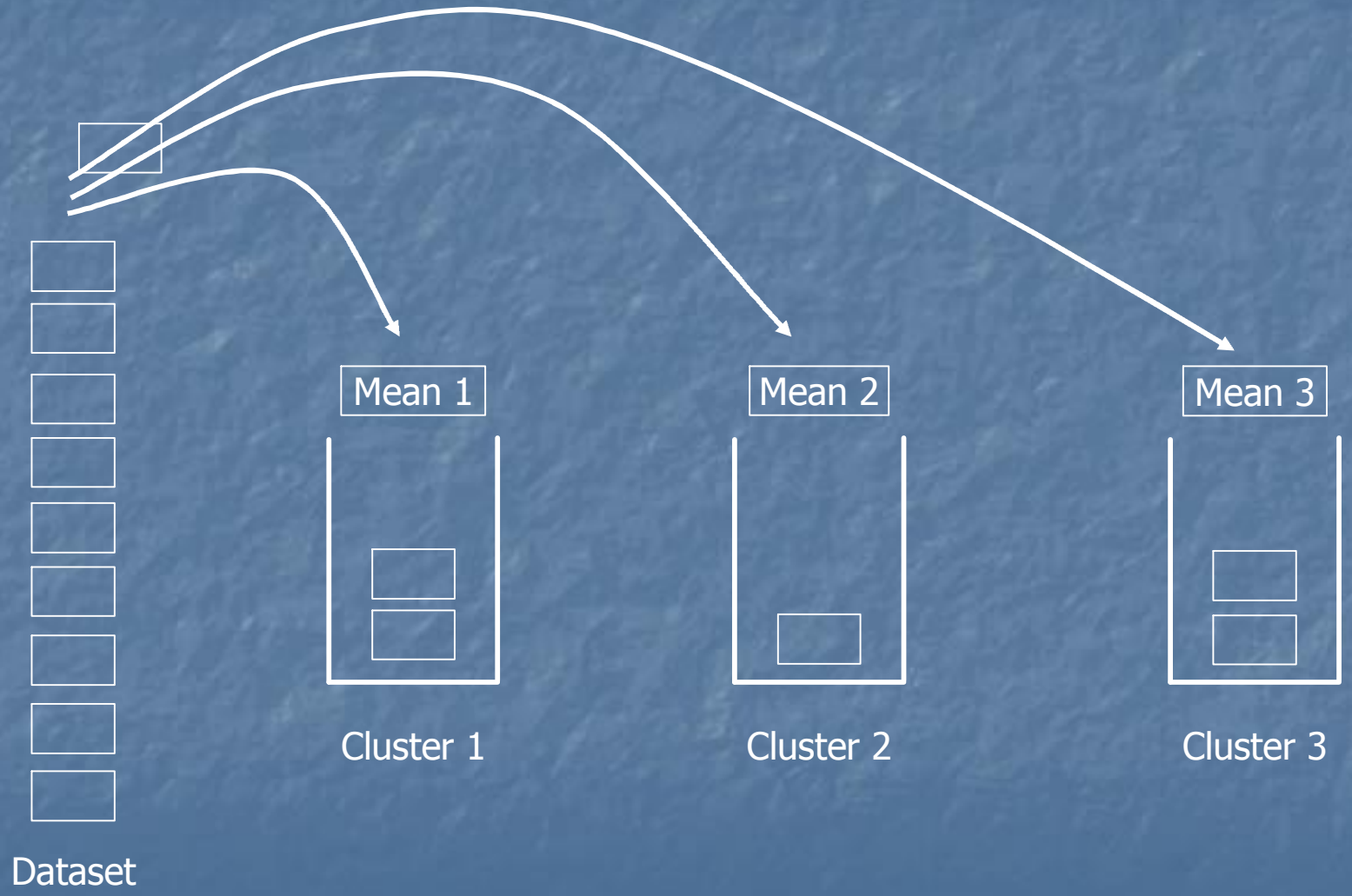
Source: Han

The Algorithm

- Input
  - k: the number of clusters
  - D: a dataset containing *n* elements
- Output: a set of *k* clusters
- Method
  - (1) arbitrarily choose *k* elements from *D* as the initial cluster mean values
  - (2) **repeat**
  - (3) assign each element to the cluster whose mean the element is *closest* to
  - (4) once all of the elements are assigned to clusters, calculate the *actual* cluster means
  - (5) **until** there is no change between the new and old cluster means

Source: Han

The Algorithm



Issues

- The algorithm can only be applied when the mean of a cluster is defined
- The numbers of clusters must be specified in advance
- This method is not suitable for clusters with non-convex shapes
- This method is sensitive to noise and outlier elements

Source: Han

An Example

- Iris dataset
- Use only the petal width attribute
- Specify 3 clusters
- Accuracy 95.33%

Cluster 1	Cluster 2	Cluster 3
46 Versicolor 3 Virginica Cluster mean 4.22857	4 Versicolor 47 Virginica Cluster mean 5.55686	50 Setosa Cluster mean 1.46275

Another Example (part 1)

- Iris dataset
- Use all attributes
- Specify 3 clusters
- Accuracy 66.0%

Cluster 1	Cluster 2	Cluster 3
47 Versicolor 49 Virginica Mean 6.30, 2.89, 4.96, 1.70	21 Setosa 1 Virginica Mean 4.59, 3.07, 1.44, 0.29	29 Setosa 3 Versicolor Mean 5.21, 3.53, 1.67, 0.35

Another Example (part 2)

- Iris dataset
- Use all attributes
- Specify 7 clusters
- Accuracy 90.67%

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
23 Virginica	1 Virginica	26 Setosa	12 Virginica	24 Versicolor 1 Virginica	26 Versicolor 13 Virginica	24 Setosa

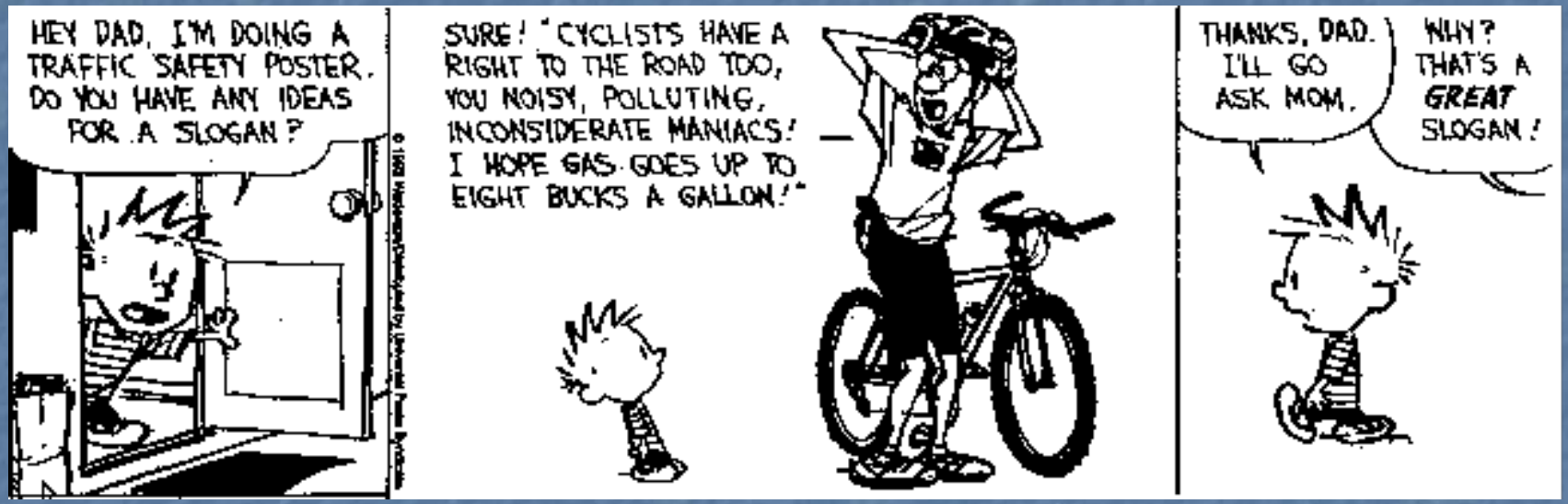
Review

- Algorithm Definition
- Algorithm Fitness Function
- The Algorithm Implementation
- Issues
- Some Examples

References

- Wikipedia, http://en.wikipedia.org/wiki/K-means
- Han, Jiawei, Data Mining: Concepts and Techniques, Elsevier Inc., 2006

Questions?


HEY DAD, I'M DOING A TRAFFIC SAFETY POSTER. DO YOU HAVE ANY IDEAS FOR A SLOGAN?
SURE! "CYCLISTS HAVE A RIGHT TO THE ROAD TOO, YOU NOISY, POLLUTING, INCONSIDERATE MANIACS! I HOPE GAS GOES UP TO EIGHT BUCKS A GALLON!"
THANKS, DAD. I'LL GO ASK MOM.
WHY? THAT'S A GREAT SLOGAN!