k-means Clustering

Papers We Love
Bucharest Chapter
February 22, 2016

Hello!

I am Adrian Florea

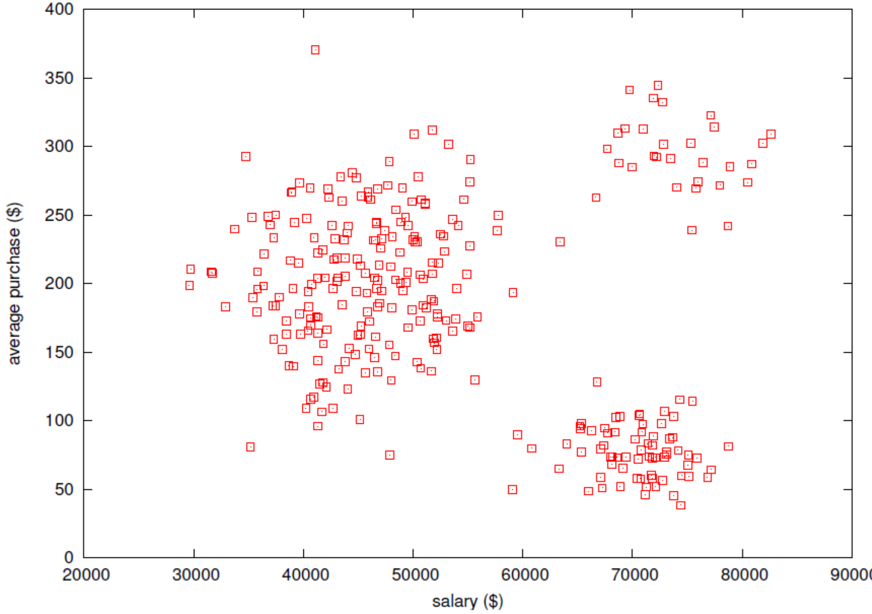
Architect Lead @ IBM Bucharest Software Lab

What is clustering?

- branch of Machine Learning (unsupervised learning i.e. find underlying patterns in the data)
- method for identifying data items that closely resemble one another, assembling them into clusters [ODS]

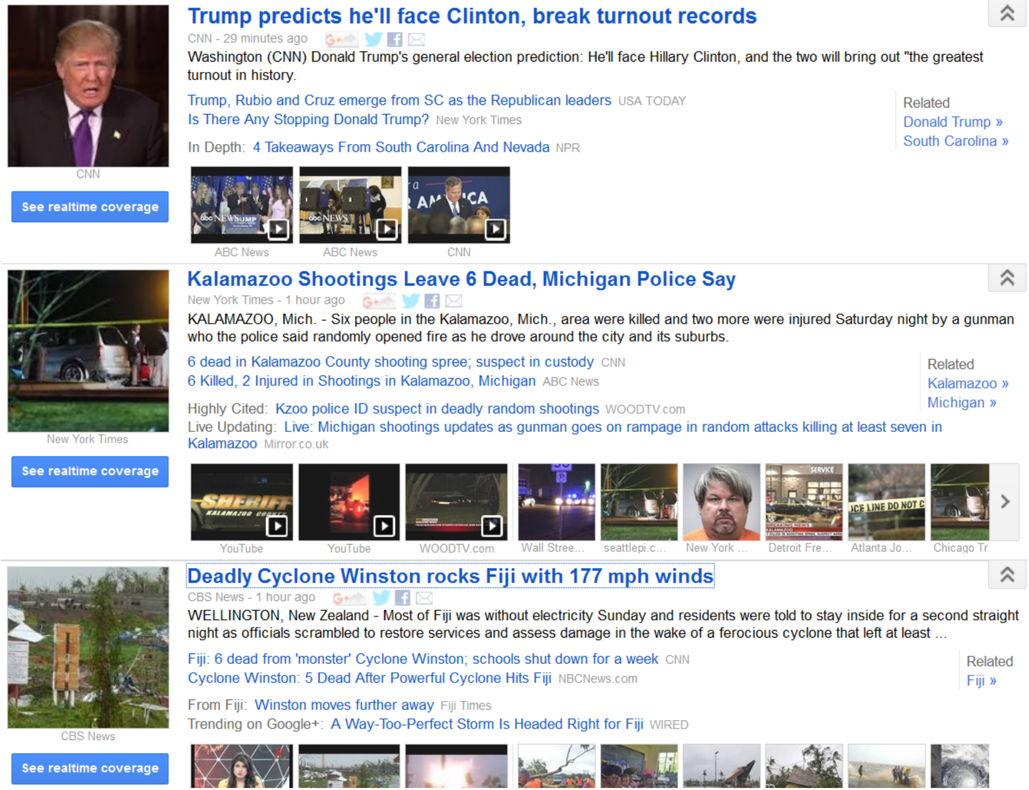
Clustering applications

- Customer segmentation



Clustering applications

- Google News



Definitions and notations

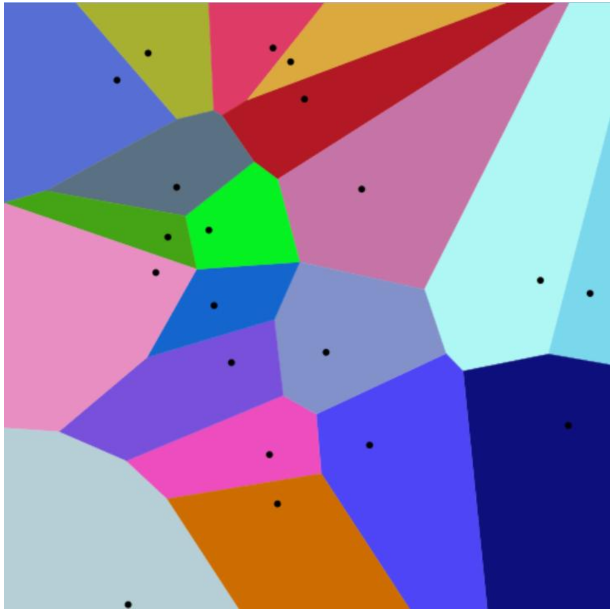
- data to be clustered (N data points) $D = \{\bar{x}_i \mid i = \overline{1, N};\ \bar{x}_i \in \boldsymbol{R}^d\}$
- iteratively refined clustering solution $C = \{\bar{c}_j \mid j = \overline{1, k};\ \bar{c}_j \in \boldsymbol{R}^d\}$
- cluster membership vector $m = \{m_i \mid i = \overline{1, N}; m_i = clusterID(x_i)\}$
- closeness cost function $KM(D, C) = \sum_{i=1}^{N} \min_{j=\overline{1,k}} d(\bar{x}_i, \bar{c}_j)$
- Minkowski distance

  p=1 (Manhattan), p=2 (Euclidian)

$$d(\bar{x}, \bar{c}) = \left(\sum_{i=1}^{d} |x_i - c_i|^p \right)^{1/p}$$

Voronoi diagrams

- Euclidian distance



- Manhattan distance

K-means algorithm

- **input**: dataset D, number of clusters k
- **output**: cluster solution C, cluster membership m
- initialize C by randomly choose k data points sets from D

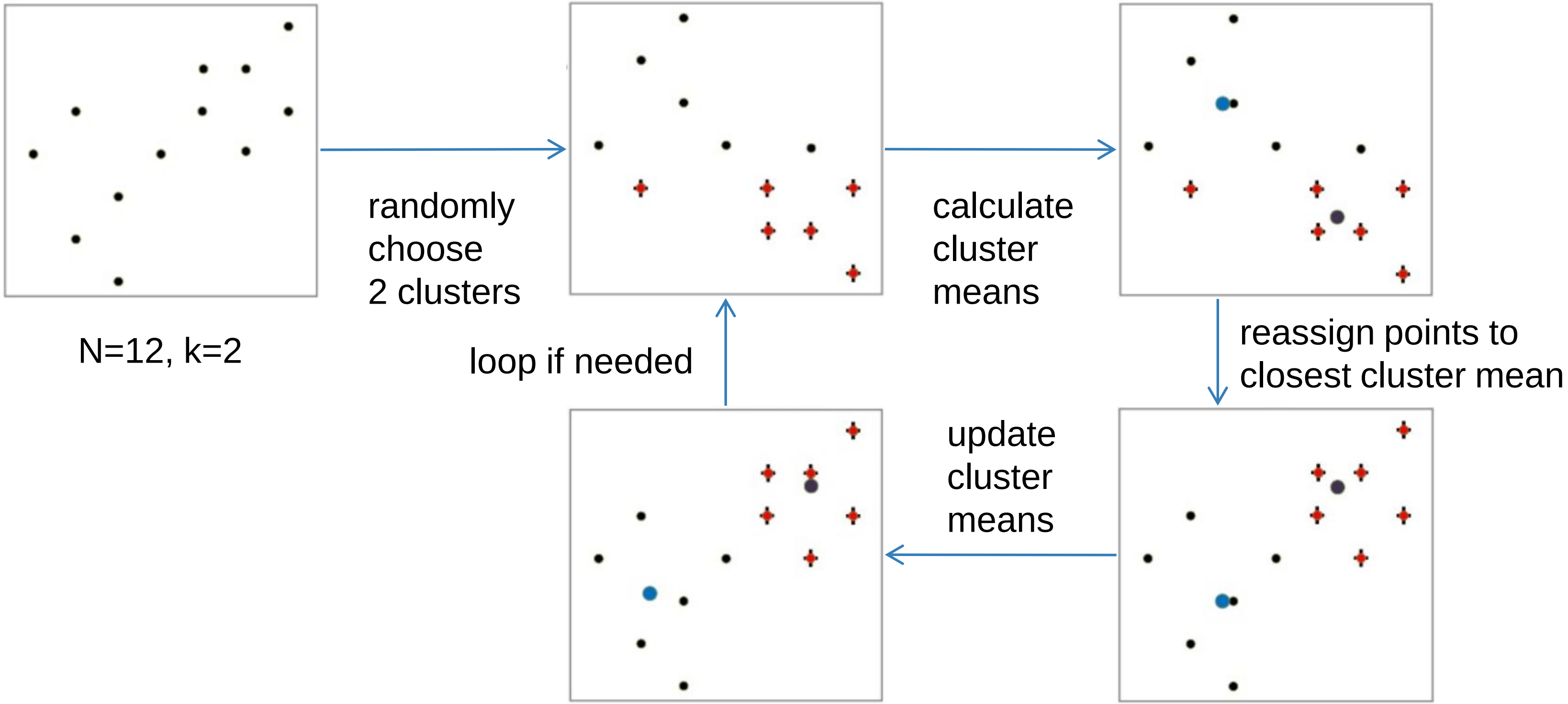
- **repeat**

  reasign points in D to closest cluster mean

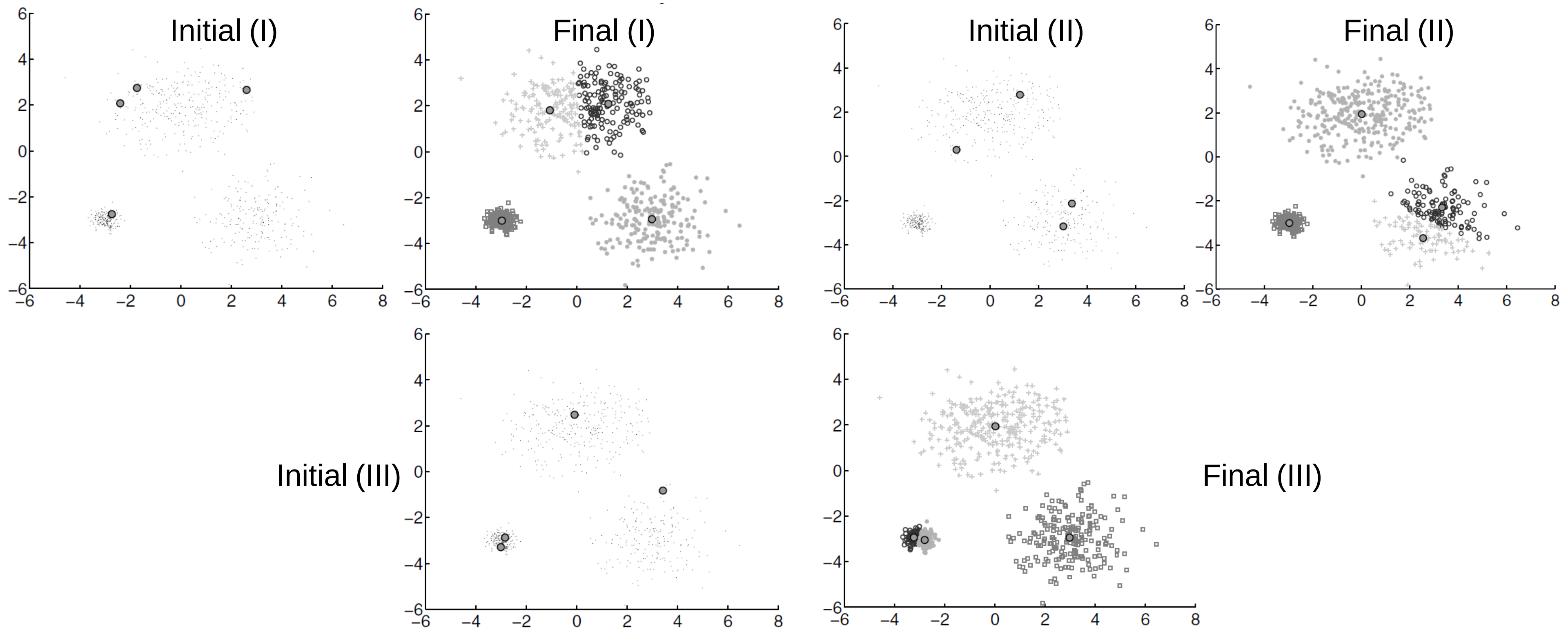
  update m

  update C such that c_j is mean of points in j^{th} cluster, $j = \overline{1, k}$
- **until** convergence of KM(D, C)

K-means iterations example



Poor initialization/Poor clustering



Pros

- simple
- common in practice
- easy to adapt
- relatively fast $O(N \cdot k \cdot d)$

Cons

- sensitive to initial partitions
- optimal k is difficult to choose
- restricted to data which has the notion of a center
- cannot handle well non-convex clusters
- does not identify outliers (because mean is not a “robust” statistic function)

Tips & tricks

- run the algorithm multiple times with different initial centroids and select the best result
- if possible, use the knowledge about the dataset in choosing k
- trying several k and choosing the one based on closeness cost function is naive
- use a more appropriate distance measure for the dataset
- use k-means together with another algorithm
- Exploit the triangular inequality to avoid to compare each data point with all centroids

Tips & tricks - continued

- recursively split the least compact cluster in 2 clusters by using 2–means
- remove outliers in preprocessing (anomaly detection)
- eliminate small clusters or merge close clusters at postprocessing
- in case of empty clusters reinitialize their centroids or steal points from the largest cluster

Tools & frameworks

- Frontline Systems XLMiner (Excel Add-in)
- SciPy K-means Clustering and Vector Quantization Modules (scipy.cluster.vq.kmeans)
- stats package in R
- Azure Machine Learning

Bibliography

- H.-H. Bock, "Origins and extensions of the k-means algorithm in cluster analysis", JEHPS, Vol. 4, No. 2, December 2008
- D. Chappell, "Understanding Machine Learning", PluralSight.com, 4 February 2016
- J. Ghosh, A. Liu, "K-Means", pp. 21-35 in X. Wu, V. Kumar (Eds.), "The Top Ten Algorithms in Data Mining", Chapman & Hall/CRC, 2009
- G.J. Hamerly, "Learning structure and concepts in data through data clustering", PhD Thesis, University of California, San Diego, 2003
- J. Han, M. Kamber, J. Pei, "Chapter 10. Cluster Analysis: Basic Concepts and Methods" in "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2011
- S.P. Lloyd, "Least Squares Quantization in PCM", IEEE Transactions on Information Theory, Vol. 28, No. 2, 1982
- J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations", 5th Berkeley Symposium, 1967
- J. McCaffrey, "Machine Learning Using C# Succinctly", Syncfusion, 2014
- [ODS] G. Upton, I. Cook, "A Dictionary of Statistics", 3rd Ed., Oxford University Press, 2014
- ***, "k-means clustering", Wikipedia
- ***, "Voronoi diagram", Wikipedia

Thanks!

Any questions ?

You can find me at