Applied Linear Algebra in Data Analysis

Case Study 01

Sivakumar Balasubramanian

Department of Bioengineering
Christian Medical College, Bagayam
Vellore 632002

What is this case study about?

- We will apply some of the concepts we have learned from the Vectors space, Matrcies, and Solutions to Linear Equations modules.

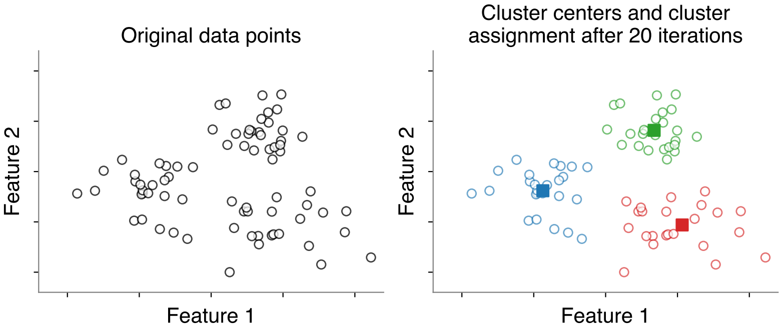
- We will be working with text data for this case study, in particular doctors' notes/reports from different specialities.

- We want to used these reports for two purposes:
  1. Cluster the set of reports into similar groups – to possibly identify which specialities these reports might be from.
  2. To use to reports to learn the relationship between different medical terms/concepts.
- We will make use of a dataset from kaggle for this case study: https://www.kaggle.com/datasets/gauravmodi/doctors-notes/data

Clustering of doctors' notes

Clustering: Grouping similar items together.

- There are various clustering algorithms: k-means, hierarchical clustering, Gaussian mixture models, etc.
- **k-means** is the simplest and most popular clustering algorithm.



Original data points

Cluster centers and cluster assignment after 20 iterations

k-means clustering

- The k-means algorithm is an iterative algorithm that divides a group of N samples (n-vectors) into k clusters.

- Clustering is done by minimizing the following cost,

$$J_{clust} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i \in C_j} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2$$

- Minimizing this cost for a given dataset is computational intensive, because the optimal choice for the means $\mathbf{m}_j$ and the cluster assignments C_j depend on each other.

- k-means takes a simpler approach: minimizing J_{clust} when either the means or the cluster assignments are fixed is easy.

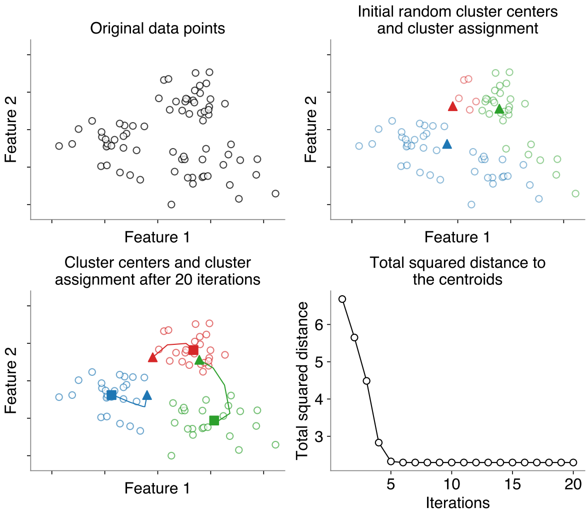
k-means clustering

- k-means solves the clustering problem by minimizing J_clust by alternatively fixing the means and the cluster assignments, while updating the other.

- k-means has two steps: we first randomly choose some cluster means.
  - **Cluster assignment update**: For a fixed set of cluster means, find cluster assignments that minimize J_{clust}.

  - **Cluster means update**: For a fixed cluster assignment, find the means $\mathbf{m}_i$ that miimize J_{clust}.
- Applying these two steps one after the other will lead to the algorithm converging towards a set of cluster means and assignments, because J_{clust} is guarnteed to reduce with each step.

k-means clustering



Clustering of doctors' notes

The details of this case study is in the `case_study_01.ipynb` file, which can be found in the `case_studies` folder.
The rest of the details are in the .ipynb file.

Case Study 01b: Co-occurence graph of medical terms

- Co-coccurence network or graph is a graph representing the relationship between different terms/concepts/keywords.
- The nodes of the graph are the terms and the edges represent a measure of how often two terms occur together in a text, sentence, etc.
- A co-occurence graph can be learned from a set of text blobs or documents containing the terms/keywords of interest.
- The details of this case study is in the `case_study_01b.ipynb` file, which can be found in the `case_studies` folder.
  The rest of the details are in the .ipynb file.