Machine Learning I: Fractal 2

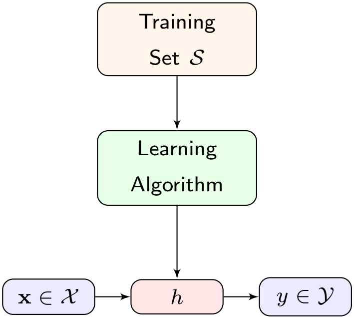
Rajendra Nagar

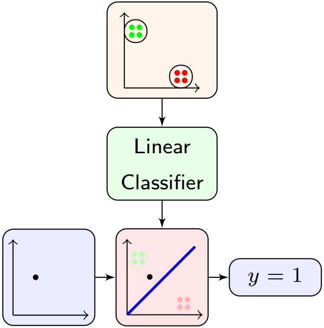
Assistant Professor
Department of Electircal Engineering
Indian Institute of Technology Jodhpur

These slides are prepared from the following book:
Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

Supervised Learning

Let $\mathcal{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$ be a training set. Here, $\mathbf{x}_i$ is the i^{th} training input, e.g. an image, and y_i is the corresponding label, e.g. "cat". Let $\mathcal{X}$ be the set of all inputs and let $\mathcal{Y}$ be the set of all possible output labels and let $h : \mathcal{X} \to \mathcal{Y}$ be a predictor. Then, our goal is to find h such that $h(\mathbf{x}_i)$ is equal to the true label of the input $\mathbf{x}_i$.

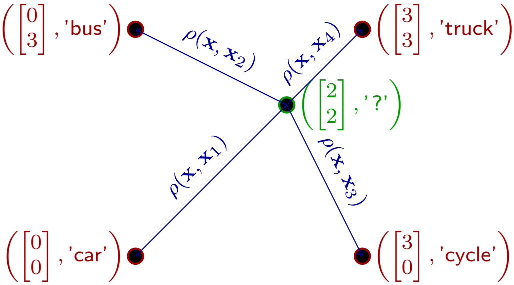




k-Nearest Neighbors Classifier

- Let $\mathcal{S} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$ be the training set.
- For each $\mathbf{x} \in \mathcal{X}$, let $\pi_1(\mathbf{x}), \ldots, \pi_m(\mathbf{x})$ be reordering of $\{1, 2, \ldots, m\}$ according to their distance to $\mathbf{x}$, $\rho(\mathbf{x}, \mathbf{x}')$.
- That is, for all $i < m$,

$$\rho(\mathbf{x}, \mathbf{x}_{\pi_i(\mathbf{x})}) \leq \rho(\mathbf{x}, \mathbf{x}_{\pi_{i+1}(\mathbf{x})})$$



Clustering

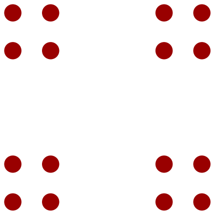
Input: A set of elements, $\mathcal{X}$, and a distance function to measure similarity.

Clustering

Input: A set of elements, $\mathcal{X}$, and a distance function to measure similarity.
Objective: A partition of the input domain $\mathcal{X}$ into groups $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ of similar elements such that $\cup_{i=1}^{k} \mathcal{C}_i = \mathcal{X}$, and $\mathcal{C}_i \cap \mathcal{C}_j = \phi, \forall i \neq j$.

Clustering

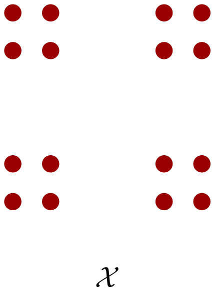
Input: A set of elements, $\mathcal{X}$, and a distance function to measure similarity.
Objective: A partition of the input domain $\mathcal{X}$ into groups $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ of similar elements such that $\cup_{i=1}^{k} \mathcal{C}_i = \mathcal{X}$, and $\mathcal{C}_i \cap \mathcal{C}_j = \phi, \forall i \neq j$.

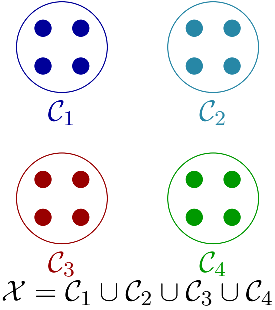


$\mathcal{X}$

Clustering

Input: A set of elements, $\mathcal{X}$, and a distance function to measure similarity.
Objective: A partition of the input domain $\mathcal{X}$ into groups $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$ of similar elements such that $\cup_{i=1}^{k} \mathcal{C}_i = \mathcal{X}$, and $\mathcal{C}_i \cap \mathcal{C}_j = \phi, \forall i \neq j$.





Clustering Algorithms

- Connectivity Based: Hierarchical Clustering

Clustering Algorithms

- Connectivity Based: Hierarchical Clustering
- Centroid models: k-Means

Clustering Algorithms

- Connectivity Based: Hierarchical Clustering
- Centroid models: k-Means
- Graph-based models: Spectral Clustering

Clustering Algorithms

- Connectivity Based: Hierarchical Clustering
- Centroid models: k-Means
- Graph-based models: Spectral Clustering
- Distribution models: Expectation Maximization

Clustering Algorithms

- Connectivity Based: Hierarchical Clustering
- Centroid models: k-Means
- Graph-based models: Spectral Clustering
- Distribution models: Expectation Maximization
- Density models: DBSCAN

Clustering Algorithms

- Connectivity Based: Hierarchical Clustering
- Centroid models: k-Means
- Graph-based models: Spectral Clustering
- Distribution models: Expectation Maximization
- Density models: DBSCAN
- Neural models: Self-organizing map

Linkage-Based Clustering Algorithm

- Start from the trivial clustering that has each data point as a single-point cluster.

Linkage-Based Clustering Algorithm

- Start from the trivial clustering that has each data point as a single-point cluster.
- Then, repeatedly, merge the "closest" clusters of the previous clustering.

Linkage-Based Clustering Algorithm

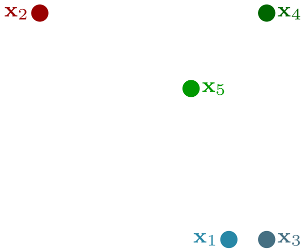
- Start from the trivial clustering that has each data point as a single-point cluster.
- Then, repeatedly, merge the “closest” clusters of the previous clustering.
- Consequently, the number of clusters decreases with each such round.

Linkage-Based Clustering Algorithm

- Start from the trivial clustering that has each data point as a single-point cluster.
- Then, repeatedly, merge the "closest" clusters of the previous clustering.
- Consequently, the number of clusters decreases with each such round.
- If kept going, this would eventually result in the trivial clustering in which all of the domain points share one large cluster.

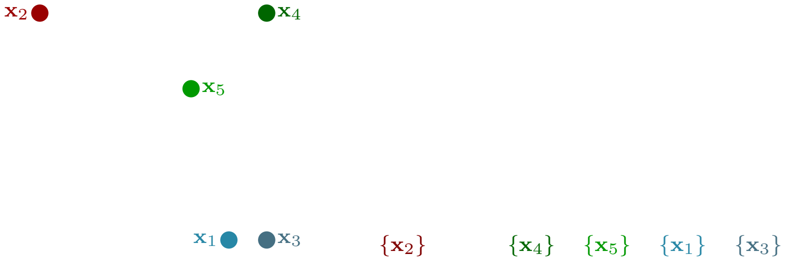
Linkage-Based Clustering Algorithm

- Start from the trivial clustering that has each data point as a single-point cluster.
- Then, repeatedly, merge the "closest" clusters of the previous clustering.
- Consequently, the number of clusters decreases with each such round.
- If kept going, this would eventually result in the trivial clustering in which all of the domain points share one large cluster.



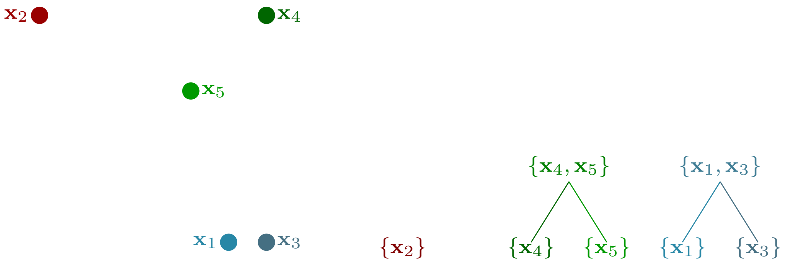
Linkage-Based Clustering Algorithm

- Start from the trivial clustering that has each data point as a single-point cluster.
- Then, repeatedly, merge the "closest" clusters of the previous clustering.
- Consequently, the number of clusters decreases with each such round.
- If kept going, this would eventually result in the trivial clustering in which all of the domain points share one large cluster.



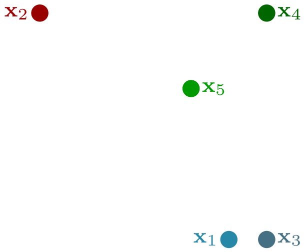
Linkage-Based Clustering Algorithm

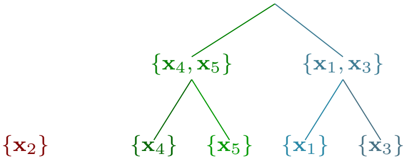
- Start from the trivial clustering that has each data point as a single-point cluster.
- Then, repeatedly, merge the "closest" clusters of the previous clustering.
- Consequently, the number of clusters decreases with each such round.
- If kept going, this would eventually result in the trivial clustering in which all of the domain points share one large cluster.



Linkage-Based Clustering Algorithm

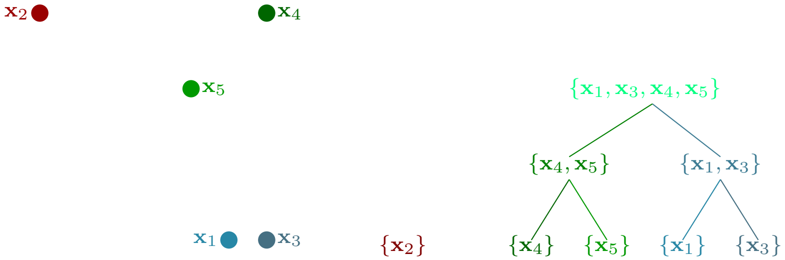
- Start from the trivial clustering that has each data point as a single-point cluster.
- Then, repeatedly, merge the “closest” clusters of the previous clustering.
- Consequently, the number of clusters decreases with each such round.
- If kept going, this would eventually result in the trivial clustering in which all of the domain points share one large cluster.





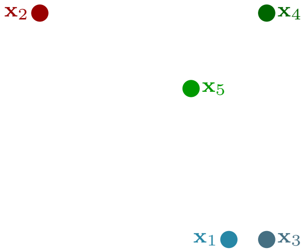
Linkage-Based Clustering Algorithm

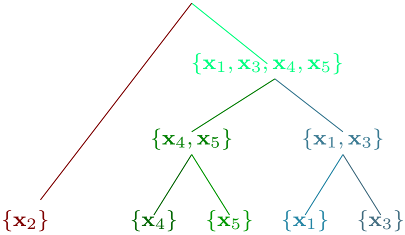
- Start from the trivial clustering that has each data point as a single-point cluster.
- Then, repeatedly, merge the "closest" clusters of the previous clustering.
- Consequently, the number of clusters decreases with each such round.
- If kept going, this would eventually result in the trivial clustering in which all of the domain points share one large cluster.



Linkage-Based Clustering Algorithm

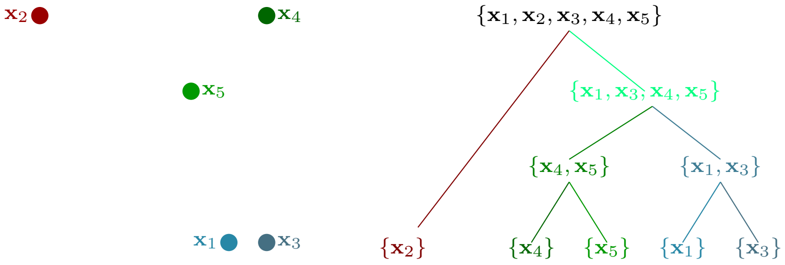
- Start from the trivial clustering that has each data point as a single-point cluster.
- Then, repeatedly, merge the "closest" clusters of the previous clustering.
- Consequently, the number of clusters decreases with each such round.
- If kept going, this would eventually result in the trivial clustering in which all of the domain points share one large cluster.





Linkage-Based Clustering Algorithm

- Start from the trivial clustering that has each data point as a single-point cluster.
- Then, repeatedly, merge the "closest" clusters of the previous clustering.
- Consequently, the number of clusters decreases with each such round.
- If kept going, this would eventually result in the trivial clustering in which all of the domain points share one large cluster.



Linkage-Based Clustering Algorithm

- The linkage-based clustering algorithms are agglomerative in the sense that they start from data that is completely fragmented and keep building larger and larger clusters as they proceed.

Linkage-Based Clustering Algorithm

- The linkage-based clustering algorithms are agglomerative in the sense that they start from data that is completely fragmented and keep building larger and larger clusters as they proceed.
- Without employing a stopping rule, the outcome of such an algorithm can be described by a clustering *dendrogram*: that is, a tree of domain subsets, having the singleton sets in its leaves, and the full domain as its root.

Linkage-Based Clustering Algorithm

- The linkage-based clustering algorithms are agglomerative in the sense that they start from data that is completely fragmented and keep building larger and larger clusters as they proceed.
- Without employing a stopping rule, the outcome of such an algorithm can be described by a clustering *dendrogram*: that is, a tree of domain subsets, having the singleton sets in its leaves, and the full domain as its root.
- Two parameters, then, need to be determined to define such an algorithm clearly.

Linkage-Based Clustering Algorithm

- The linkage-based clustering algorithms are agglomerative in the sense that they start from data that is completely fragmented and keep building larger and larger clusters as they proceed.
- Without employing a stopping rule, the outcome of such an algorithm can be described by a clustering *dendrogram*: that is, a tree of domain subsets, having the singleton sets in its leaves, and the full domain as its root.
- Two parameters, then, need to be determined to define such an algorithm clearly.
- First, we have to decide how to measure (or define) the distance between clusters, and, second, we have to determine when to stop merging.

Linkage-Based Clustering Algorithm

Single Linkage clustering

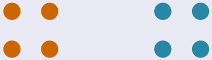
Distance between two clusters is the min. distance between their elements

Linkage-Based Clustering Algorithm

Single Linkage clustering

Distance between two clusters is the min. distance between their elements

$$D(\mathcal{A}, \mathcal{B}) = \min\{\rho(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in \mathcal{A}, \mathbf{y} \in \mathcal{B}\}$$

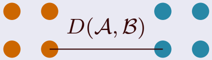


Linkage-Based Clustering Algorithm

Single Linkage clustering

Distance between two clusters is the min. distance between their elements

$$D(\mathcal{A}, \mathcal{B}) = \min\{\rho(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in \mathcal{A}, \mathbf{y} \in \mathcal{B}\}$$



Linkage-Based Clustering Algorithm

Max Linkage clustering

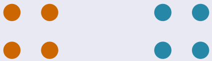
Distance between two clusters is the max. distance between their elements

Linkage-Based Clustering Algorithm

Max Linkage clustering

Distance between two clusters is the max. distance between their elements

$$D(\mathcal{A}, \mathcal{B}) = \max\{\rho(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in \mathcal{A}, \mathbf{y} \in \mathcal{B}\}$$

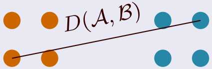


Linkage-Based Clustering Algorithm

Max Linkage clustering

Distance between two clusters is the max. distance between their elements

$$D(\mathcal{A}, \mathcal{B}) = \max\{\rho(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in \mathcal{A}, \mathbf{y} \in \mathcal{B}\}$$



Linkage-Based Clustering Algorithm

Average Linkage clustering

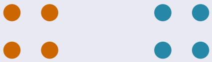
Distance between two clusters is defined to be the average distance between a point in one of the clusters and a point in the other

Linkage-Based Clustering Algorithm

Average Linkage clustering

Distance between two clusters is defined to be the average distance between a point in one of the clusters and a point in the other

$$D(\mathcal{A}, \mathcal{B}) = \frac{1}{|\mathcal{A}| \times |\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{A}} \sum_{\mathbf{y} \in \mathcal{B}} \rho(\mathbf{x}, \mathbf{y})$$

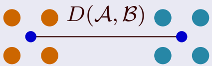


Linkage-Based Clustering Algorithm

Average Linkage clustering

Distance between two clusters is defined to be the average distance between a point in one of the clusters and a point in the other

$$D(\mathcal{A}, \mathcal{B}) = \frac{1}{|\mathcal{A}| \times |\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{A}} \sum_{\mathbf{y} \in \mathcal{B}} \rho(\mathbf{x}, \mathbf{y})$$



Linkage-Based Clustering Algorithm

- Fixed number of clusters – fix some parameter, k, and stop merging clusters as soon as the number of clusters is k.

Linkage-Based Clustering Algorithm

- Fixed number of clusters – fix some parameter, k, and stop merging clusters as soon as the number of clusters is k.
- Distance upper bound fix some $r \in \mathbb{R}^+$. Stop merging as soon as all the between clusters distances are larger than r. We can also set r to be $\alpha \times \max\{\rho(\mathbf{x}, \mathbf{y}) : \mathbf{x}, \mathbf{y} \in \mathcal{X}\}$ for some $\alpha < 1$. In that case the stopping criterion is called "scaled distance upper bound."

Linkage-Based Clustering Algorithm

- Fixed number of clusters – fix some parameter, k, and stop merging clusters as soon as the number of clusters is k.
- Distance upper bound fix some $r \in \mathbb{R}^+$. Stop merging as soon as all the between clusters distances are larger than r. We can also set r to be $\alpha \times \max\{\rho(\mathbf{x}, \mathbf{y}) : \mathbf{x}, \mathbf{y} \in \mathcal{X}\}$ for some $\alpha < 1$. In that case the stopping criterion is called "scaled distance upper bound."

k-Means: Problem Formulation[1,2]

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set data-points, where $\mathbf{x}_i \in \mathbb{R}^d$. We want to partition $\mathcal{X}$ in groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ containing similar objects.

[1] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967.

[2] Lloyd, Stuart P. "Least squares quantization in PCM." IEEE Transactions on Information Theory, 1982.

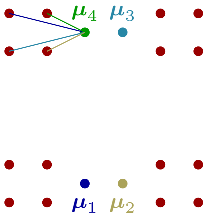
k-Means: Problem Formulation[1,2]

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set data-points, where $\mathbf{x}_i \in \mathbb{R}^d$. We want to partition $\mathcal{X}$ in groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ containing similar objects. Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ be their respective group representatives (centres), where $\boldsymbol{\mu}_i \in \mathbb{R}^d$.

[1] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967.

[2] Lloyd, Stuart P. "Least squares quantization in PCM." IEEE Transactions on Information Theory, 1982.

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set data-points, where $\mathbf{x}_i \in \mathbb{R}^d$. We want to partition $\mathcal{X}$ in groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ containing similar objects. Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ be their respective group representatives (centres), where $\boldsymbol{\mu}_i \in \mathbb{R}^d$.

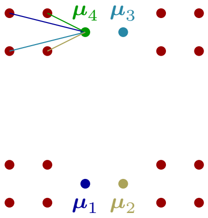


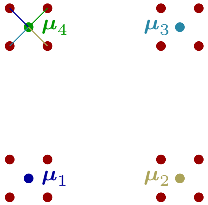
[1] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967.

[2] Lloyd, Stuart P. "Least squares quantization in PCM." IEEE Transactions on Information Theory, 1982.

k-Means: Problem Formulation[1,2]

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set data-points, where $\mathbf{x}_i \in \mathbb{R}^d$. We want to partition $\mathcal{X}$ in groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ containing similar objects. Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ be their respective group representatives (centres), where $\boldsymbol{\mu}_i \in \mathbb{R}^d$.





[1] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967.

[2] Lloyd, Stuart P. "Least squares quantization in PCM." IEEE Transactions on Information Theory, 1982.

k-Means: Problem Formulation[3,4]

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set data-points, where $\mathbf{x}_i \in \mathbb{R}^d$. We want to partition $\mathcal{X}$ in groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ containing similar objects.

[1] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967.

[2] Lloyd, Stuart P. "Least squares quantization in PCM." IEEE Transactions on Information Theory, 1982.

k-Means: Problem Formulation[3,4]

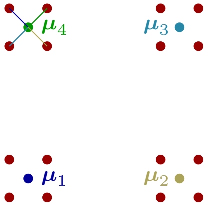
Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set data-points, where $\mathbf{x}_i \in \mathbb{R}^d$. We want to partition $\mathcal{X}$ in groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ containing similar objects.Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ be their respective group representatives (centres), where $\boldsymbol{\mu}_i \in \mathbb{R}^d$.

[1] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967.

[2] Lloyd, Stuart P. "Least squares quantization in PCM." IEEE Transactions on Information Theory, 1982.

k-Means: Problem Formulation[3,4]

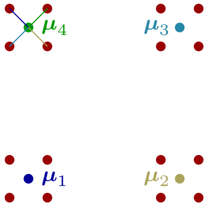
Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set data-points, where $\mathbf{x}_i \in \mathbb{R}^d$. We want to partition $\mathcal{X}$ in groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ containing similar objects.Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ be their respective group representatives (centres), where $\boldsymbol{\mu}_i \in \mathbb{R}^d$.



[1] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967.

[2] Lloyd, Stuart P. "Least squares quantization in PCM." IEEE Transactions on Information Theory, 1982.

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set data-points, where $\mathbf{x}_i \in \mathbb{R}^d$. We want to partition $\mathcal{X}$ in groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ containing similar objects.Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ be their respective group representatives (centres), where $\boldsymbol{\mu}_i \in \mathbb{R}^d$.



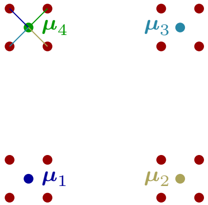
- $f(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|_2^2$

[1] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967.

[2] Lloyd, Stuart P. "Least squares quantization in PCM." IEEE Transactions on Information Theory, 1982.

k-Means: Problem Formulation[3,4]

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set data-points, where $\mathbf{x}_i \in \mathbb{R}^d$. We want to partition $\mathcal{X}$ in groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ containing similar objects.Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ be their respective group representatives (centres), where $\boldsymbol{\mu}_i \in \mathbb{R}^d$.



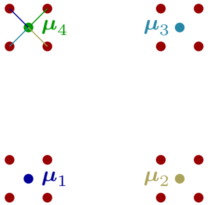
- $f(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|_2^2$
- $\underset{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k}{\arg\min} \; f(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)$.

[1] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967.

[2] Lloyd, Stuart P. "Least squares quantization in PCM." IEEE Transactions on Information Theory, 1982.

k-Means: Problem Formulation[3,4]

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set data-points, where $\mathbf{x}_i \in \mathbb{R}^d$. We want to partition $\mathcal{X}$ in groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ containing similar objects.Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ be their respective group representatives (centres), where $\boldsymbol{\mu}_i \in \mathbb{R}^d$.



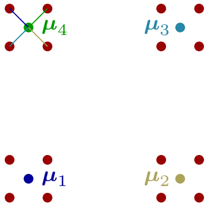
- $f(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|_2^2$
- $\underset{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k}{\arg\min}\ f(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)$.
- $\frac{\partial f}{\partial \boldsymbol{\mu}_i} = \mathbf{0} \Rightarrow \boldsymbol{\mu}_i = \frac{\sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x}}{|\mathcal{C}_i|}$

[1] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967.

[2] Lloyd, Stuart P. "Least squares quantization in PCM." IEEE Transactions on Information Theory, 1982.

k-Means: Problem Formulation[3,4]

Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be the set data-points, where $\mathbf{x}_i \in \mathbb{R}^d$. We want to partition $\mathcal{X}$ in groups $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$ containing similar objects.Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ be their respective group representatives (centres), where $\boldsymbol{\mu}_i \in \mathbb{R}^d$.



- $f(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k) = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|_2^2$
- $\underset{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k}{\arg\min} \; f(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)$.
- $\frac{\partial f}{\partial \boldsymbol{\mu}_i} = \mathbf{0} \Rightarrow \boldsymbol{\mu}_i = \frac{\sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x}}{|\mathcal{C}_i|}$
- $\boldsymbol{\mu}_i = \text{mean}(\mathcal{C}_i) \; \forall i \in \{1, 2, \ldots, k\}$.

[1] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967.

[2] Lloyd, Stuart P. "Least squares quantization in PCM." IEEE Transactions on Information Theory, 1982.

Algorithm 1 k-Means Algorithm

Input: $\mathcal{X} \subset \mathbb{R}^d$, Number of clusters k

Initialize: Randomly choose initial centroids $\boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_k^{(0)}$

while not converged **do**

  for $i \in [k]$ **do**

    $\mathcal{C}_i^{(t+1)} \leftarrow \left\{ \mathbf{x} : \|\mathbf{x} - \boldsymbol{\mu}_i^{(t)}\|^2 < \|\mathbf{x} - \boldsymbol{\mu}_j^{(t)}\|^2 \; \forall j \in [k] \backslash \{i\}, \mathbf{x} \in \mathcal{X} \right\}$

    $\boldsymbol{\mu}_i^{(t+1)} \leftarrow \frac{1}{|\mathcal{C}_i^{(t+1)}|} \sum_{\mathbf{x} \in \mathcal{C}_i^{(t+1)}} \mathbf{x}$

    $t \leftarrow t + 1$

  end for

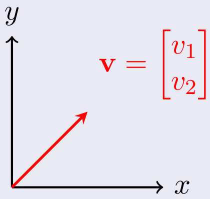
end while

Norm of a vector

Let $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$ be a vector in $\mathbb{R}^n$. Then, the norm of the vector $\mathbf{v}$ is defined as

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}.$$

$$\mathbf{v}^\top \mathbf{v} = v_1^2 + v_2^2 + \cdots + v_n^2 = \|\mathbf{v}\|^2.$$



Distance Between Two Vectors

$$\|\mathbf{u} - \mathbf{v}\|_2^2 = (\mathbf{u} - \mathbf{v})^\top (\mathbf{u} - \mathbf{v}) = \mathbf{u}^\top \mathbf{u} - 2\mathbf{u}^\top \mathbf{v} + \mathbf{v}^\top \mathbf{v}.$$

$$\frac{\partial f}{\partial \boldsymbol{\mu}_i} = \mathbf{0} \Rightarrow \frac{\partial}{\partial \boldsymbol{\mu}_i} \sum_{j=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_j} \|\mathbf{x} - \boldsymbol{\mu}_j\|_2^2 = \mathbf{0},\ \forall i \in [k]$$

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}_i} \sum_{j=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_j} \|\mathbf{x} - \boldsymbol{\mu}_j\|_2^2 &= \sum_{\mathbf{x} \in \mathcal{C}_i} \frac{\partial}{\partial \boldsymbol{\mu}_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|_2^2 \\
&= \sum_{\mathbf{x} \in \mathcal{C}_i} \frac{\partial}{\partial \boldsymbol{\mu}_i} \left((\mathbf{x} - \boldsymbol{\mu}_i)^\top (\mathbf{x} - \boldsymbol{\mu}_i) \right) \\
&= \sum_{\mathbf{x} \in \mathcal{C}_i} \frac{\partial}{\partial \boldsymbol{\mu}_i} \left((\mathbf{x}^\top - \boldsymbol{\mu}_i^\top)(\mathbf{x} - \boldsymbol{\mu}_i) \right) \\
&= \sum_{\mathbf{x} \in \mathcal{C}_i} \frac{\partial}{\partial \boldsymbol{\mu}_i} \left(\mathbf{x}^\top \mathbf{x} - \boldsymbol{\mu}_i^\top \mathbf{x} - \mathbf{x}^\top \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_i \right) \\
&= \sum_{\mathbf{x} \in \mathcal{C}_i} \frac{\partial}{\partial \boldsymbol{\mu}_i} \left(\mathbf{x}^\top \mathbf{x} - 2\boldsymbol{\mu}_i^\top \mathbf{x} + \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_i \right)
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial f}{\partial \boldsymbol{\mu}_i} \sum_{j=1}^{k} \sum_{\mathbf{x} \in \mathcal{C}_j} \|\mathbf{x} - \boldsymbol{\mu}_j\|_2^2 &= \sum_{\mathbf{x} \in \mathcal{C}_i} \frac{\partial f}{\partial \boldsymbol{\mu}_i} \left(\mathbf{x}^\top \mathbf{x} - 2\boldsymbol{\mu}_i^\top \mathbf{x} + \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_i \right) \\
&= \sum_{\mathbf{x} \in \mathcal{C}_i} (-2\mathbf{x} + 2\boldsymbol{\mu}_i) \\
\Rightarrow \sum_{\mathbf{x} \in \mathcal{C}_i} (-2\mathbf{x} + 2\boldsymbol{\mu}_i) &= \mathbf{0} \\
\Rightarrow \sum_{\mathbf{x} \in \mathcal{C}_i} \boldsymbol{\mu}_i &= \sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x} \\
\Rightarrow \boldsymbol{\mu}_i \sum_{\mathbf{x} \in \mathcal{C}_i} 1 &= \sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x} \\
\Rightarrow \boldsymbol{\mu}_i |\mathcal{C}_i| &= \sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x} \\
\Rightarrow \boldsymbol{\mu}_i &= \frac{1}{|\mathcal{C}_i|} \sum_{\mathbf{x} \in \mathcal{C}_i} \mathbf{x}
\end{aligned}
$$