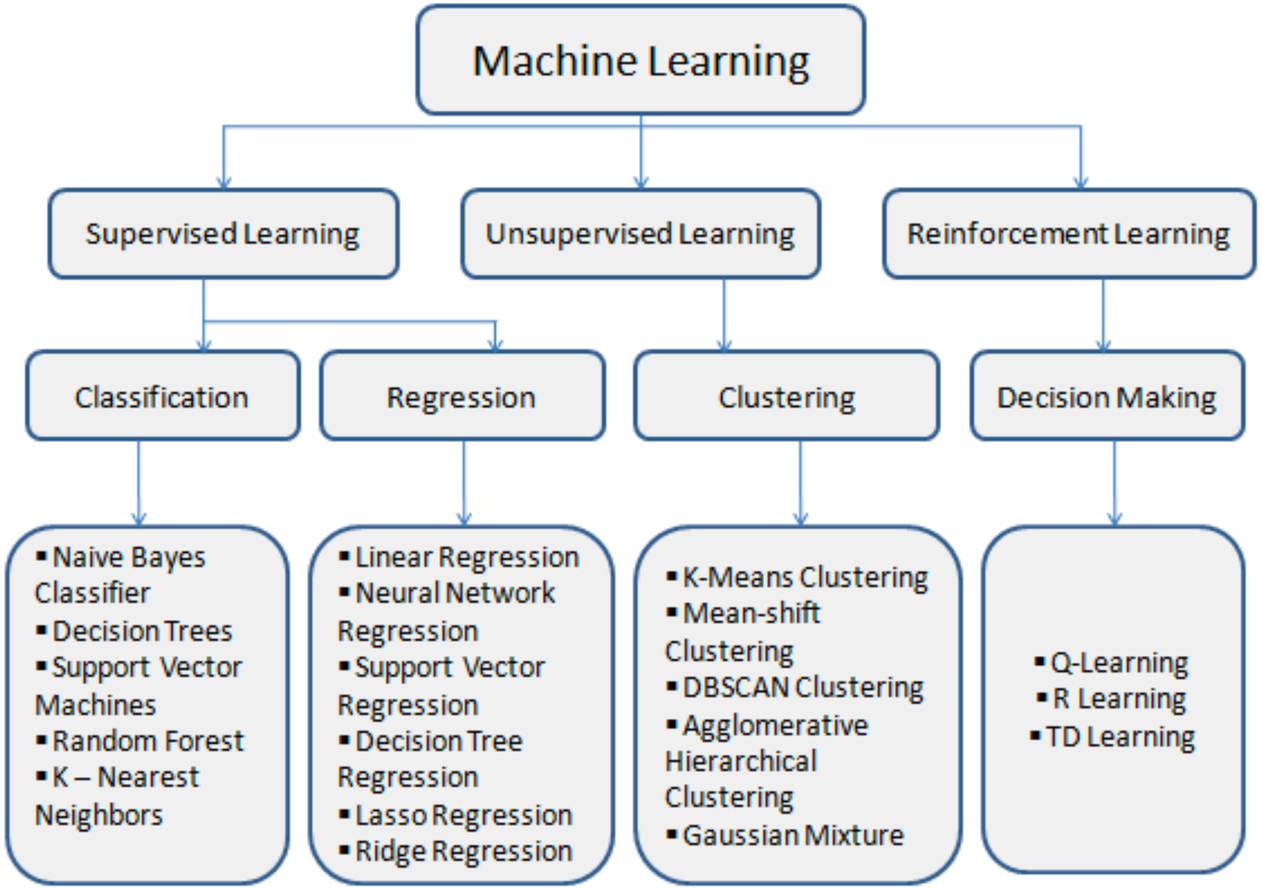
أكاديمية طويق
TUWAIQ ACADEMY

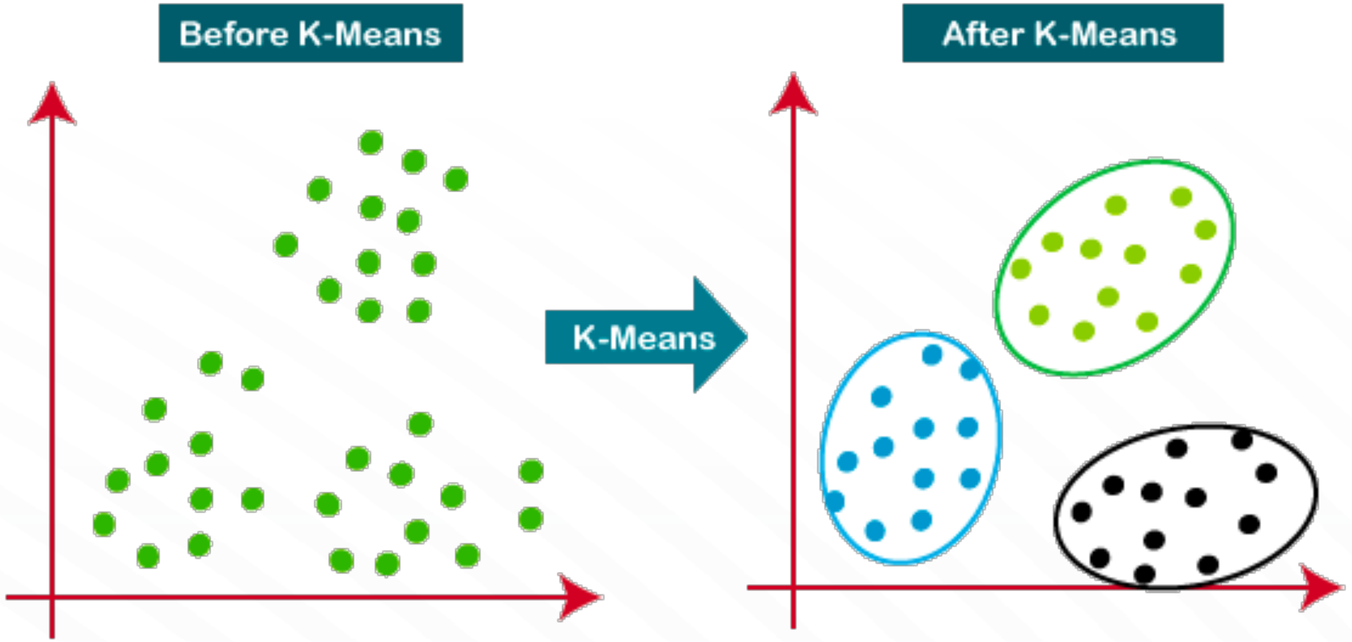
مقدمة في تعلم الآلة

خوارزميات تعلم الآلة (Machine Learning Algorithms)



خوارزمية (k-means clustering)

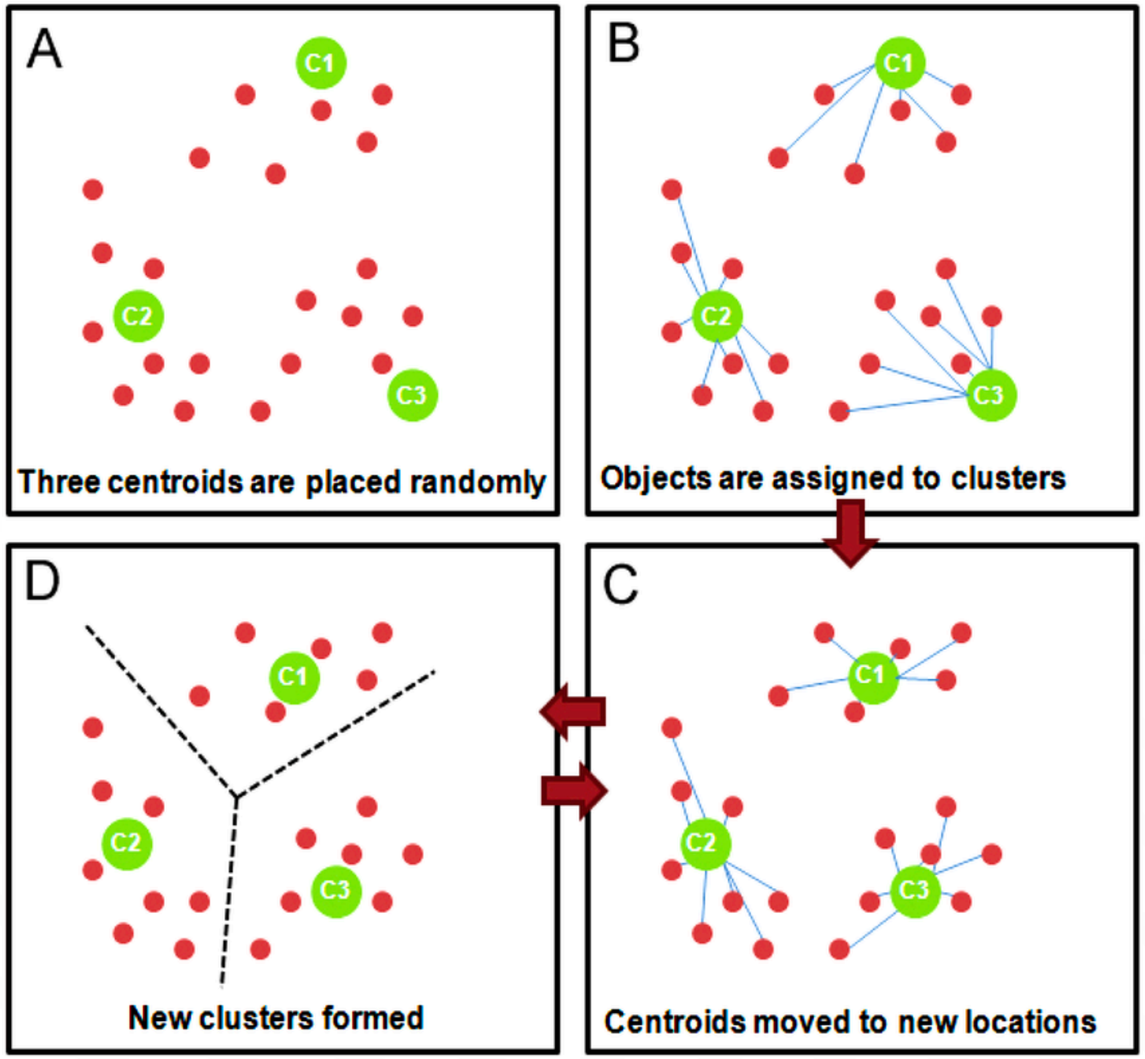
أحد أشكال الخوارزميات المستخدمة في Unsupervised Learning، وتهدف هذه الخوارزمية إلى تقسيم البيانات وتجميعها على شكل مجموعات تسمى Clusters بناء على التشابه بين هذه البيانات.



أمثلة لاستخدامات الخوارزمية:

- تصنيف المستندات المتشابهة
- تصنيف المستخدمين بناء على خصائصهم

خوارزمية (k-means clustering)



الخطوات المستخدمة لتطبيق الخوارزمية:

- اختيار عدد المجموعات K
- تصنيف البيانات بشكل عشوائي للمجموعات
- نقوم بحساب cluster centroid لكل مجموعة عن طريق حساب mean vector
- تعيين كل نقطة من البيانات إلى المجموعات التي تكون فيها هذه النقطة قريبة لقيمة centroid
- نقوم بتكرار الخطوة الثالثة و الرابعة حتى تتوقف cluster عن التغير

خوارزمية (k-means clustering)

الخطوات المستخدمة لتطبيق الخوارزمية:

Algorithm 1 k-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

خوارزمية (k-means clustering)

- تعتبر هذه الخوارزمية (Centroid-based Partitioning Method) وذلك لأنها تقوم باستخدام Center لتمثيل Cluster.
- يتم تمثيل Centroid عن طريق حساب mean للنقاط الموجودة في Cluster.

يتم حساب جودة (Quality of Cluster) عن طريق معادلة (Euclidean Distance):

- عن طريق حساب مجموع squared error لجميع النقاط في مجموعة البيانات.
- حيث **p** يمثل object بينما **ci** يمثل the centroid of Cluster.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c_i})^2$$

خوارزمية (k-means clustering)

A = {1, 2, 3, 6, 7, 8, 13, 15, 18}

- K=3 clusters
- Choose three cluster centroids (by random)
  - C_1={1}, M_1=1
  - C_2={2}, M_2=2
  - C_3={3}, M_3=3
- Step#1: Update three cluster centroids (by nearest cluster)
  - dist(M_3,6) < dist(M_2,6)
  - dist(M_3,6) < dist(M_1,6)
  - C_1={1}, M_1=1
  - C_2={2}, M_2=2
  - C_3={3, 6, 7, 8, 13, 15, 18}, M_3 = 70/7 = 10

خوارزمية (k-means clustering)

A = {1, 2, 3, 6, 7, 8, 13, 15, 18}

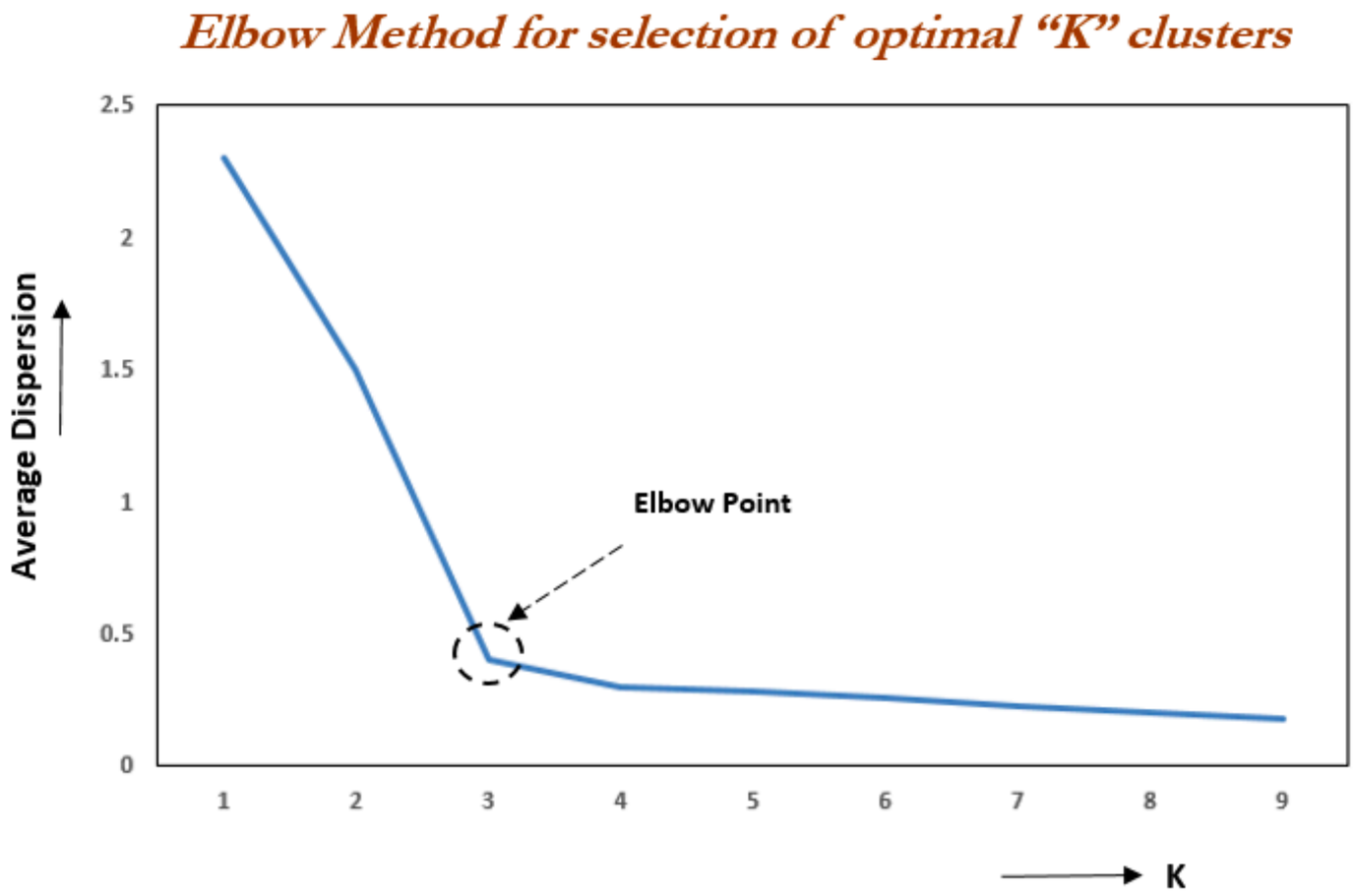
- **Step#2: Update three cluster centroids (by nearest cluster)**
  - dist(3, M2) < dist(3, M3) ➔ 3 goes to C2
  - dist(6, M2) = dist(6, M3) ➔ 6 goes to C2
  - C1={1}, M1=1
  - C2={2, 3, 6}, M2=3.7
  - C3={7, 8, 13, 15, 18}, M3= 12.2
- **Step#3: Update three cluster centroids (by nearest cluster)**
  - dist(2, M1) < dist(2, M2) ➔ 2 goes to C1
  - dist(7, M2) < dist(7, M3) ➔ 7 goes to C2
  - C1={1,2}, M1=1.5
  - C2={3,6,7}, M2=5.3
  - C3= {8,13,15,18}, M3=13.5

خوارزمية (k-means clustering)

A = {1, 2, 3, 6, 7, 8, 13, 15, 18}

- **Step#4: Update three cluster centroids (by nearest cluster)**
  - dist(3, M1) < dist(3, M2) ➔ 3 goes to C1
  - dist(8, M2) < dist(8, M3) ➔ 8 goes to C2
  - C1={1, 2, 3}, M1=2
  - C2={6, 7, 8}, M2=7
  - C3={13,15,18}, M3=15.3 → No change

خوارزمية (k-means clustering)



خوارزمية (k-means clustering)

مميزات خوارزمية k-means:

- فعالة (efficient) بشكل نسبي لأن (time complexity is O(nkt)) حيث يمثل (n=#objects, k=#clusters, t=#iterations)
- قابلية التطور بشكل نسبي (Relatively scalable) في معالجة مجموعات البيانات الكبيرة.

عيوب خوارزمية k-means:

- قابلة للتطبيق فقط على البيانات الرقمية أي عندما يمكننا حساب المتوسط (mean).
- تتطلب تحديد (number of clusters) بشكل مسبق.
- حساسة لوجود القيم الشاذة (outliers) أو (noisy data).

Resources

- Data Science: The Big Picture [https://app.pluralsight.com/library/courses/data-science-big-picture/table-of-contents].
- Introduction to Data Science [https://link.springer.com/book/10.1007/978-3-319-50017-1].
- Data Mining: Concepts and Techniques [https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques].