Artificial Intelligence for Medical Data with Python

10 SAMPLE SLIDES

4th session – Clustering of Medical Data and Genetic Algorithms

UNIVERSITY OF THE AEGEAN

SCHOOL OF ENGINEERING

DEPARTMENT OF INFORMATION AND COMMUNICATION SYSTEMS ENGINEERING

Presenter: Panagiotis Symeonidis

Associate Professor

http://panagiotissymeonidis.com

psymeon@aegean.gr

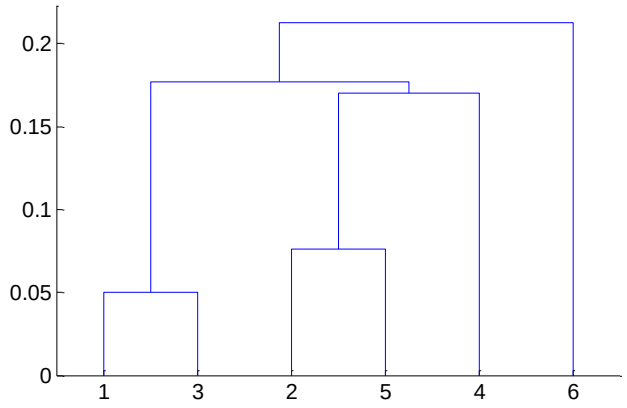
K-means Clustering

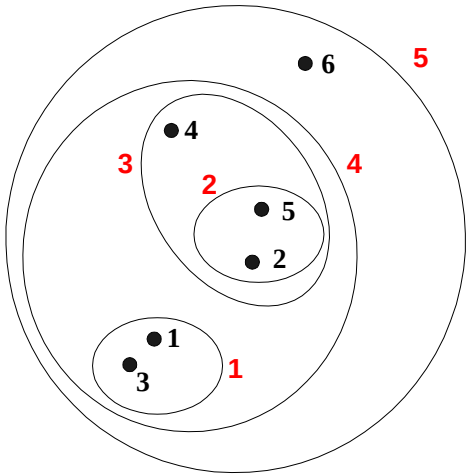
- Partitional clustering approach
- Number of clusters, K, must be specified
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- The basic algorithm is very simple

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits





DBSCAN algorithm

- General idea:
  - *Form clusters using core points, and assign border points to one of its neighboring clusters*

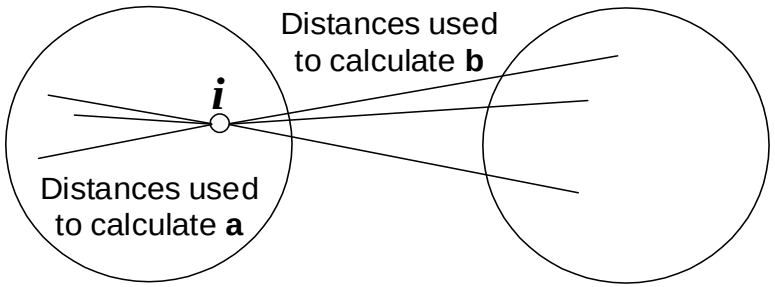
DBSCAN algorithm

1: Calculate distance matrix of points (Euclidean distance)
2: Based on Eps and MinPts, label all points as core, border, or noise points
3: Eliminate noise points
4: Put an edge between all core points within a distance *Eps* of each other
6: Make each group of connected core points into a separate cluster
7: Assign each border point to one of the clusters of its associated core points

Silhouette Coefficient

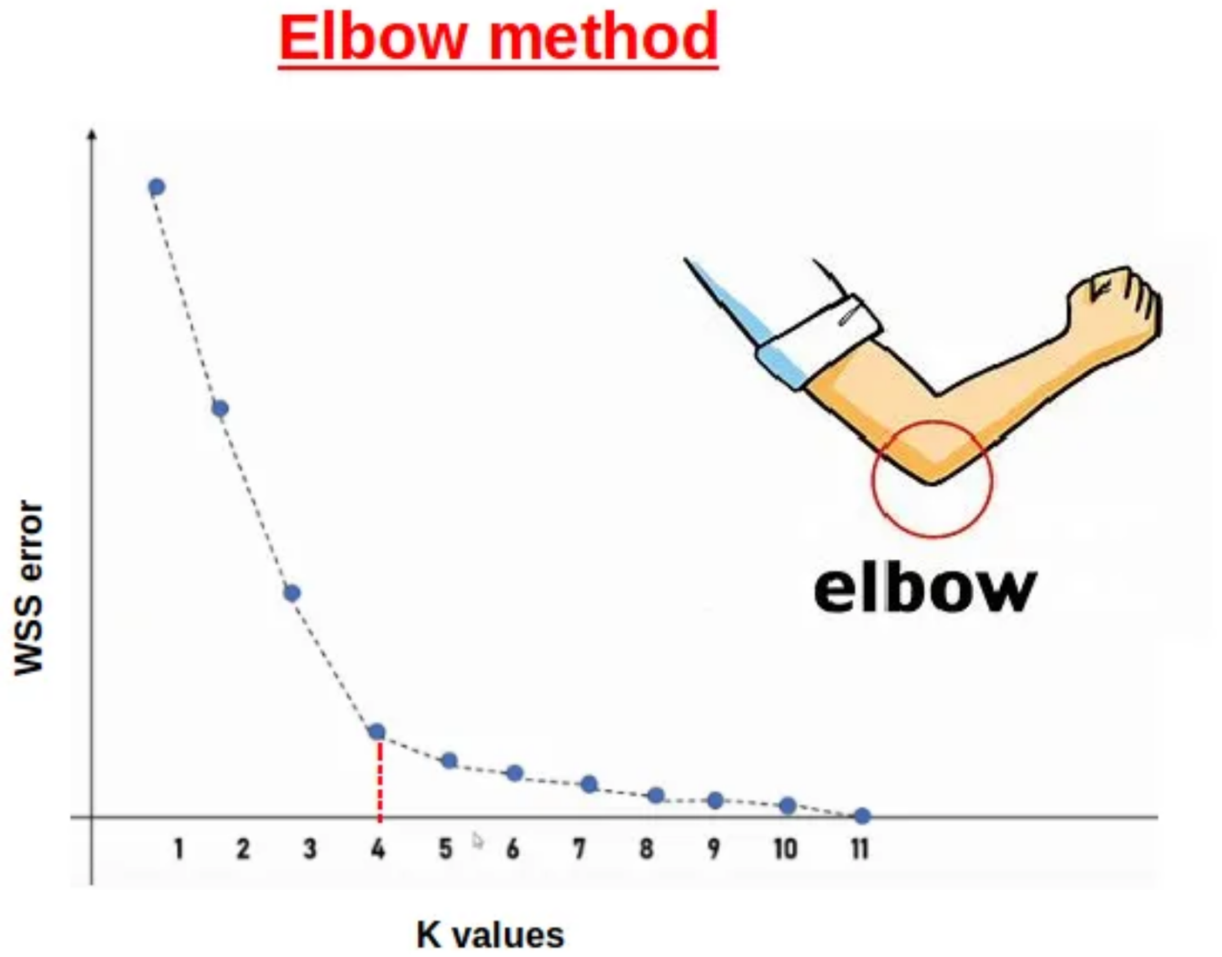
- Silhouette coefficient combines ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings
- For an individual point, ***i***
  - Calculate ***a*** = average distance of ***i*** to the points in its cluster
  - Calculate ***b*** = min (average distances of ***i*** to points in all other cluster)
  - The silhouette coefficient for a point is then given by

$$s = (b - a) / \max(a, b)$$



  - Value can vary between -1 and 1
  - Typically ranges between 0 and 1.
  - The closer to 1 the better.

HOW TO DECIDE THE NUMBER OF CLUSTERS



Comparison: Elbow vs. Silhouette

- Pros and Cons of the Elbow Method:

  Pros: Simple, visual, widely applicable.

  Cons: Subjective, sometimes ambiguous, computationally intensive for large datasets.

  Pros and Cons of the Silhouette Coefficient:

  Pros: Quantitative, considers cohesion and separation, reduces subjectivity.

  Cons: Computationally intensive, may not always be clear-cut.

5 CLUSTERS of Diabetic Patients

- **Cluster 0: Current Smokers with Moderate Health Risks**
  - **Age**: Approximately 44 years
  - **Hypertension**: Low prevalence (7.95%)
  - **Heart Disease**: Very low prevalence (3.34%)
  - **BMI**: Around 28.4
  - **HbA1c Level**: Around 5.54
  - **Blood Glucose Level**: Around 139.5
  - **Diabetes**: Low prevalence (9.85%)
  - **Smoking History**: All individuals are current smokers
- **Cluster 1: Non-smokers with Low Health Risks**
  - **Age**: Approximately 42 years
  - **Hypertension**: Very low prevalence (0%)
  - **Heart Disease**: Very low prevalence (0%)
  - **BMI**: Around 27.8
  - **HbA1c Level**: Around 5.51
  - **Blood Glucose Level**: Around 137.2
  - **Diabetes**: Low prevalence (7.23%)
  - **Smoking History**: Majority never smoked (77.41%), with some having a history of smoking
- **Cluster 2: Younger Individuals with Minimal Health Issues**
  - **Age**: Approximately 32 years
  - **Hypertension**: Very low prevalence (0%)
  - **Heart Disease**: Very low prevalence (0%)
  - **BMI**: Around 25.1
  - **HbA1c Level**: Around 5.45
  - **Blood Glucose Level**: Around 134.6
  - **Diabetes**: Very low prevalence (3.04%)
  - **Smoking History**: Majority have no information on smoking history (100%)
- **Cluster 3: Former Smokers with Moderate to High Health Risks**
  - **Age**: Approximately 57 years
  - **Hypertension**: Higher prevalence (12.25%)
  - **Heart Disease**: Moderate prevalence (7.53%)
  - **BMI**: Around 29.6
  - **HbA1c Level**: Around 5.64
  - **Blood Glucose Level**: Around 142.9
  - **Diabetes**: Higher prevalence (16.32%)
  - **Smoking History**: All individuals are former smokers
- **Cluster 4: Older Individuals with High Health Risks**
  - **Age**: Approximately 64 years
  - **Hypertension**: Very high prevalence (73.51%)
  - **Heart Disease**: High prevalence (38.43%)
  - **BMI**: Around 30.3
  - **HbA1c Level**: Around 5.83
  - **Blood Glucose Level**: Around 150.2
  - **Diabetes**: Higher prevalence (28.29%)
  - **Smoking History**: Mixed smoking history with a significant proportion never smoked (52.09%) and some having a history of smoking
-

Genetic Algorithms

- The genetic algorithms exploit important mechanisms of the natural functions of organisms
  - selection,
  - crossover,
  - mutation

- Genetic algorithms follow a search process for the optimal solution, which is guided by a fitness function, that evaluates a large number of different possible solutions.

Architecture of a genetic algorithm