LING 506 - TOPICS IN COMPUTATIONAL LINGUISTICS

Introductory Machine Learning

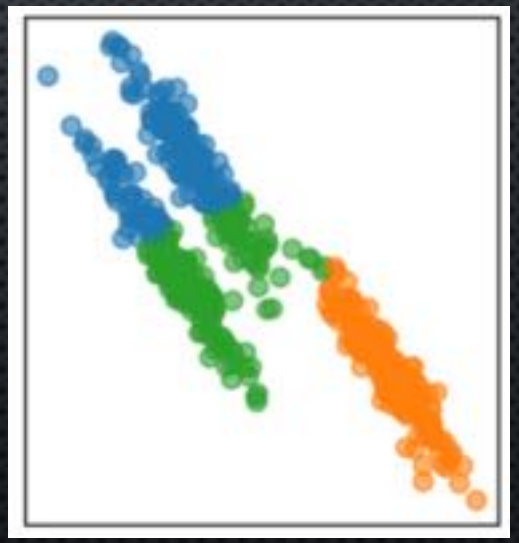
Yan Tang

Department of Linguistics, UIUC

Week 4

Last week…

- K-means clustering
  - Unsupervised ML
  - Groups data according to the distance between data and k centroids
  - Requires prior knowledge of k
  - One the fastest algorithms and scalable
  - Less well performance when clusters have different size, density or non-spherical distribution

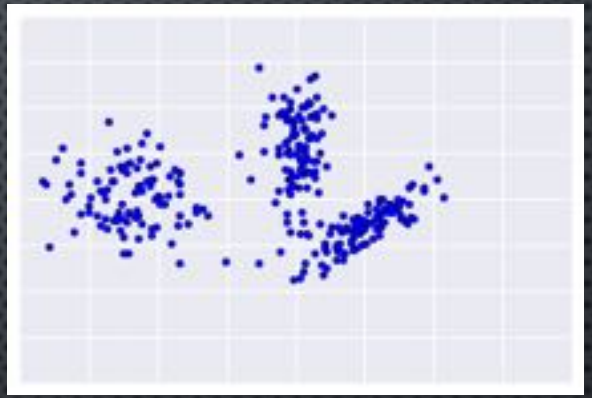


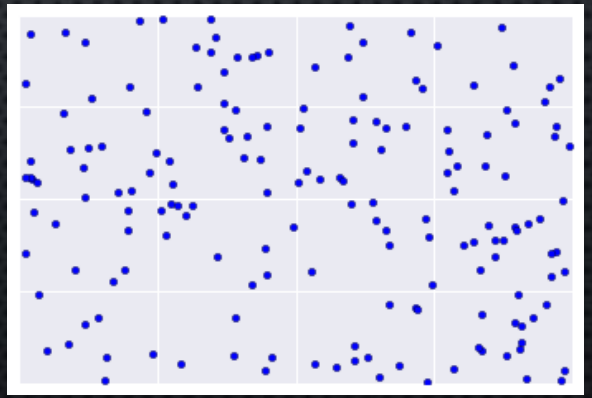
Last week…

- Clustering Gaussian Mixture Model (GMM)
  - Unsupervised ML
  - Groups data to one of the k Gaussians in which they have the highest probability
  - Requires prior knowledge of k
  - Allows soft clustering (by probability)
  - Extends k-means by considering covariance and centres of the latent Gaussians
  - Work well for data with ellipsoid distribution

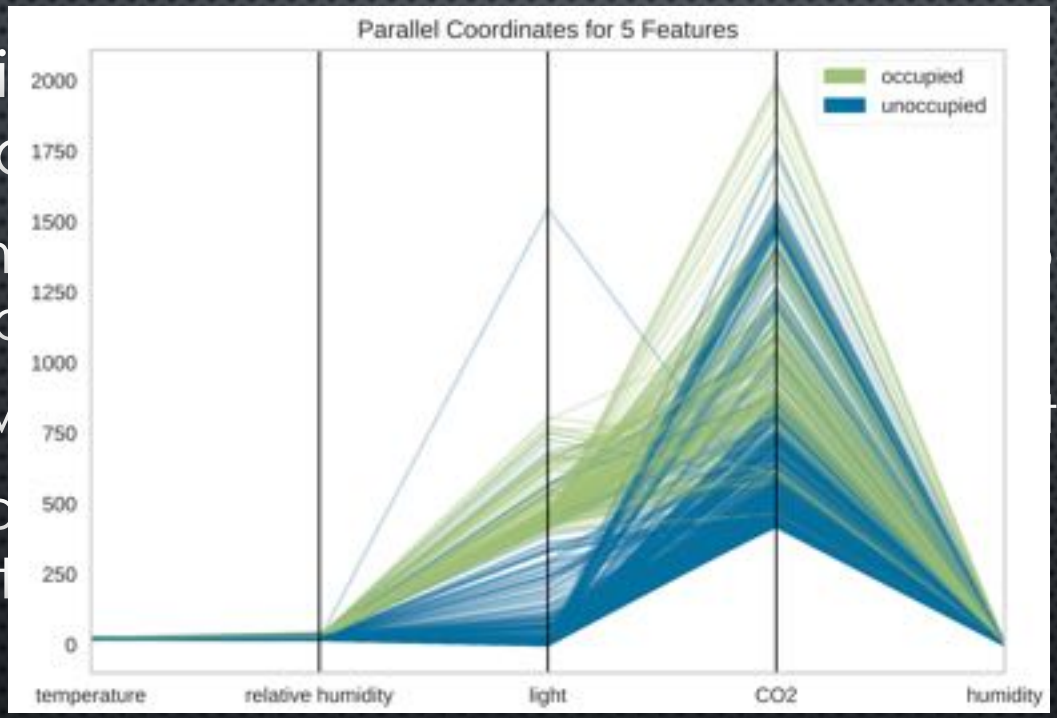
Cluster results interpretation and evaluation

- As an unsupervised ML problem, ground truth may not be available to validate results
- Gain insights in the relationships among variables/features and clusters
- Determine the potential optimal number of clusters for the data using different evaluation criteria





Parallel Coordinates: visualise observations in clusters



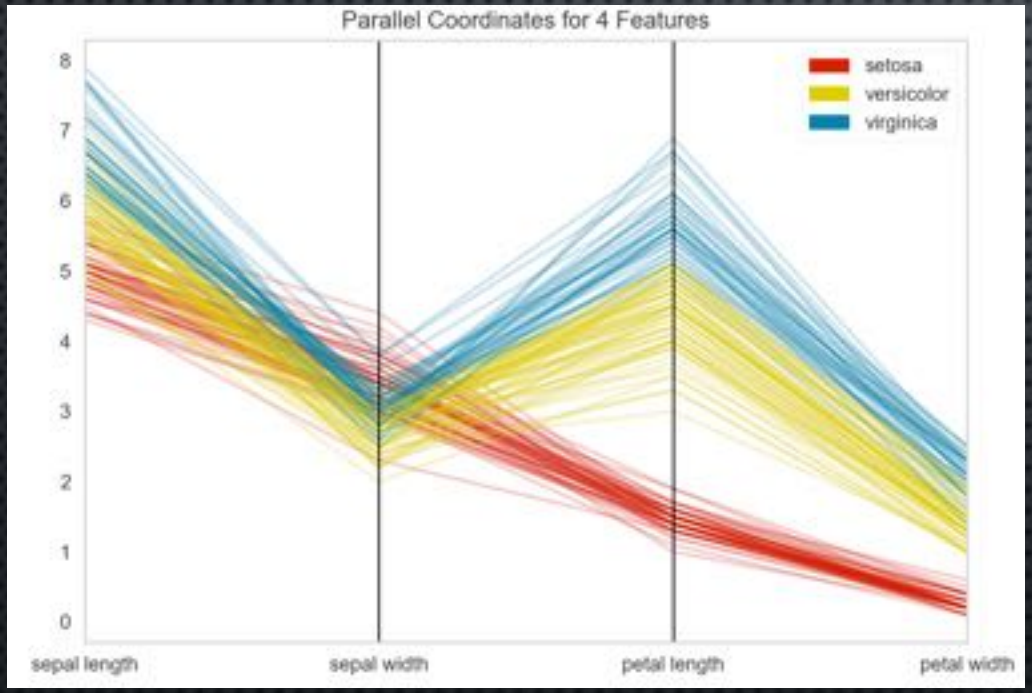
- A multi
  techni
  - Pern
    visu
  - Allov
  - Help
    clust
- Graph
  - x-axis: categorical; dimensions/variables
  - y-axis: values of each variable for individual observations

e

ta set.

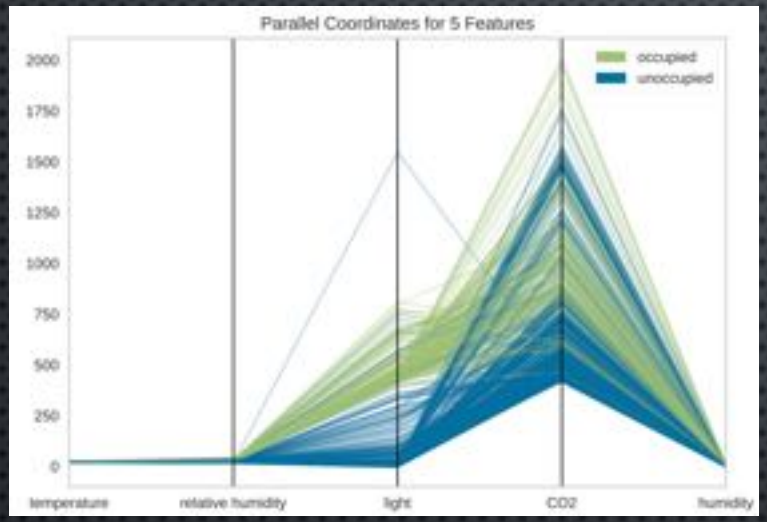
multiple

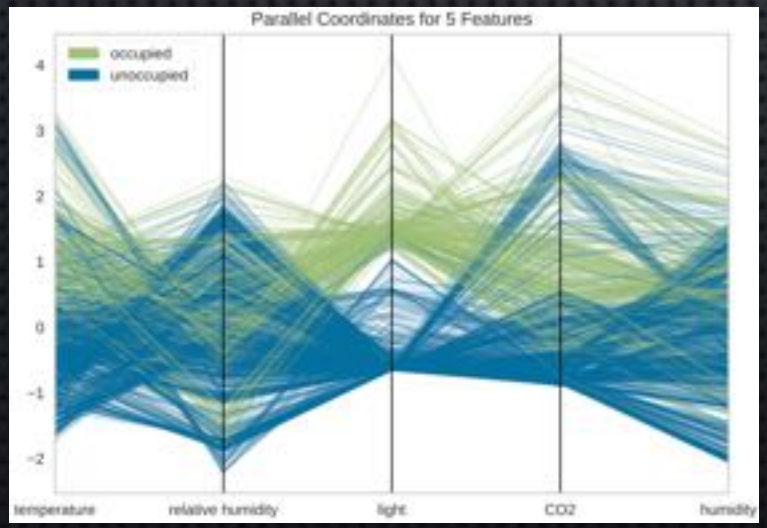
Parallel Coordinates: Iris data



- Distinctive range of occupation on "petal length" and "petal width"
- High correlations:
  - Sepal length vs petal length: 0.87
  - Sepal length vs petal length: 0.82
  - Patel length vs petal width: 0.96

Impact of data normalisation





- What are the issues of the current plot on the original data?
  - Variables have different unit, i.e. values may fall in different numeric range
- Data normalisation
  - To scale or normalise data into the same domain
  - To transform raw data to a more friendly representation to downstream algorithms
  - May improve model performance

Data standardization/normalisation

- Common requirement for many ML algorithms
- Many types of standardisation/normalisation
  - Standardisation (mean-zero): subtract mean and scales to unit variance

$$X' = \frac{X - \text{mean}(X)}{\text{std}(X)}$$

  - MinMax: transformed data are in the range $[min', max']$

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)} \times (max' - min') + min'$$

  - MaxAbs: transformed data lies with the range [-1, 1]

$$X' = \frac{X}{\max(|X|)}$$

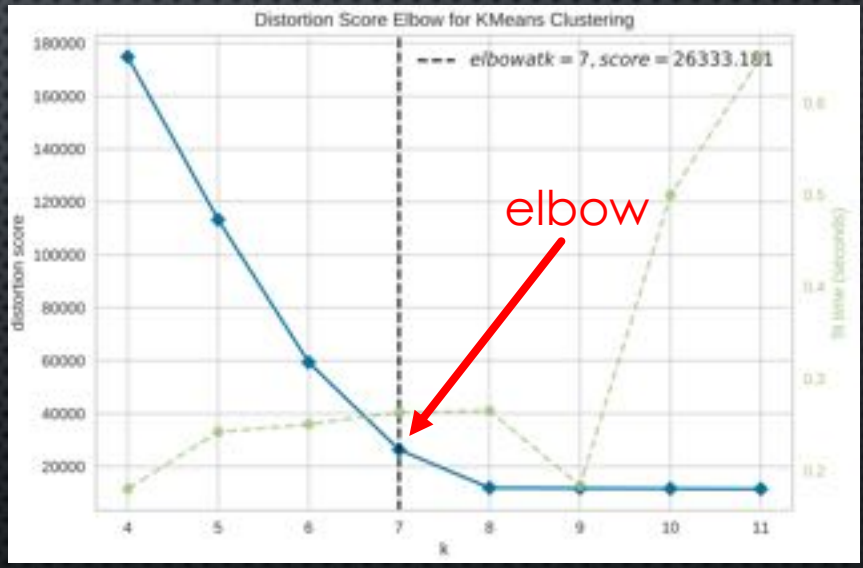
Data standardization does NOT always improve model performance!

Finding the optimal k: the elbow method

- A method to select the optimal number of clusters, k, by fitting the model with a range of values

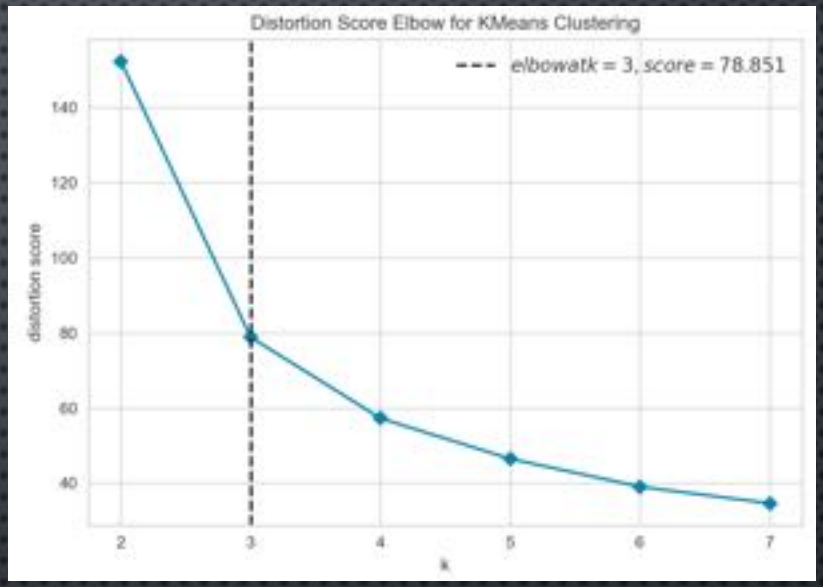
- The optimal k value is at the "elbow" (the point of inflection on the curve)

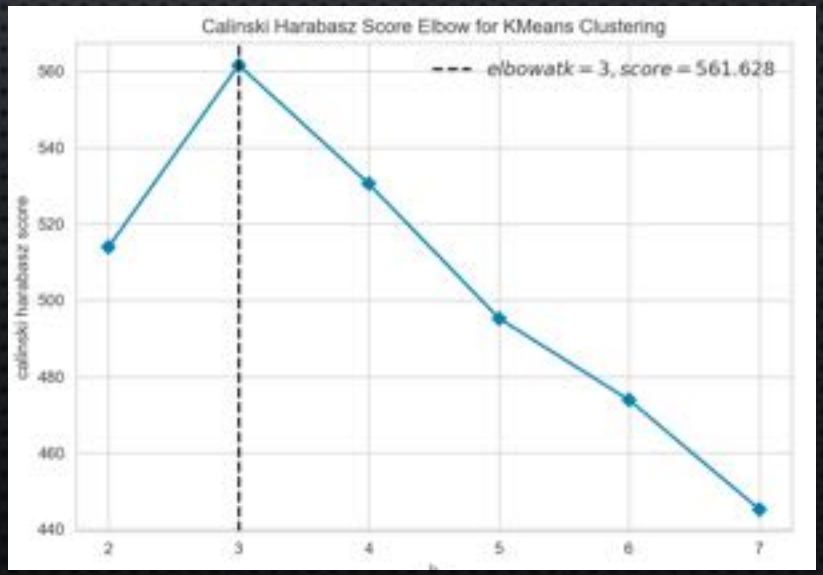
- Criteria: distortion, silhouette, calinski_harabasz



Finding the optimal k: the elbow method

- Optimal k for Iris data with different criteria.
- Still a rather coarse and empirical approach





Finding the optimal k: the Silhouette coefficient

- The density of clusters sought by the target model
  - Can be used to judge the quality of the clusters.

- A normalised measure ($[-1, 1]$) of how close an observation is to others in the same cluster, compared to the observations in different clusters
  - $\rightarrow$ 1: well inside its own cluster and far from others
  - $\rightarrow$ 0: close to a cluster boundary
  - $\rightarrow$ -1: may have been assigned to the wrong cluster

Finding the optimal k: the Silhouette coefficient

- The Silhouette Coefficient of an observation, Si :

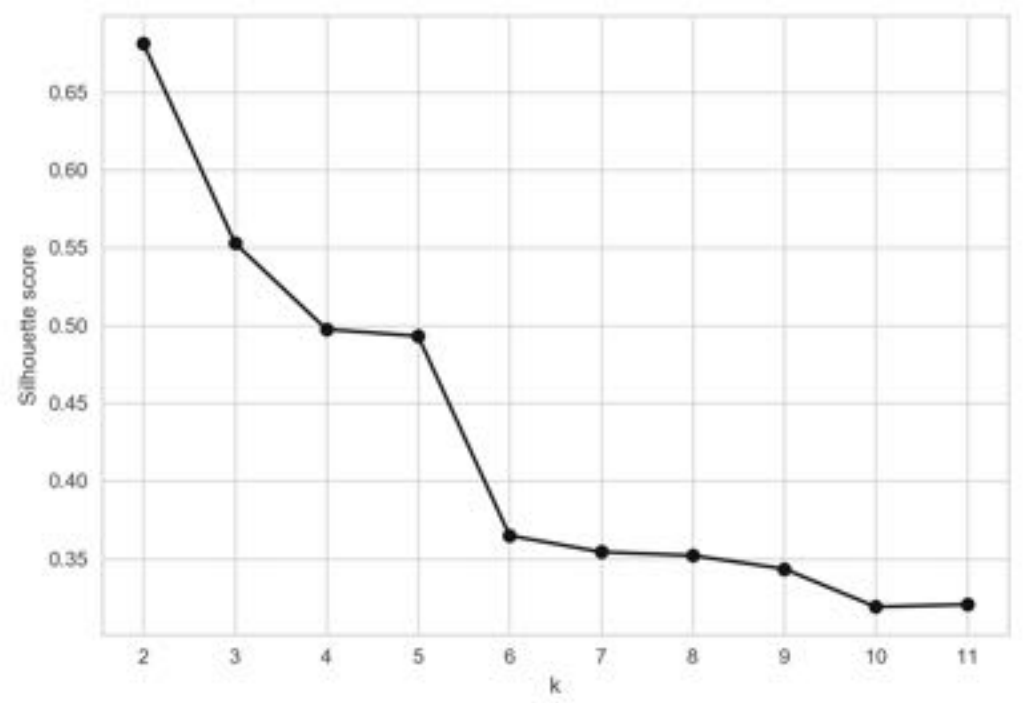
$$Si = \frac{b - a}{\max(a, b)}$$

a: mean distance between the current observation and other observations in the same cluster

b: minimum mean distance between the current observation to observations in a different cluster, minimized over clusters.

Finding the optimal k: the Silhouette score

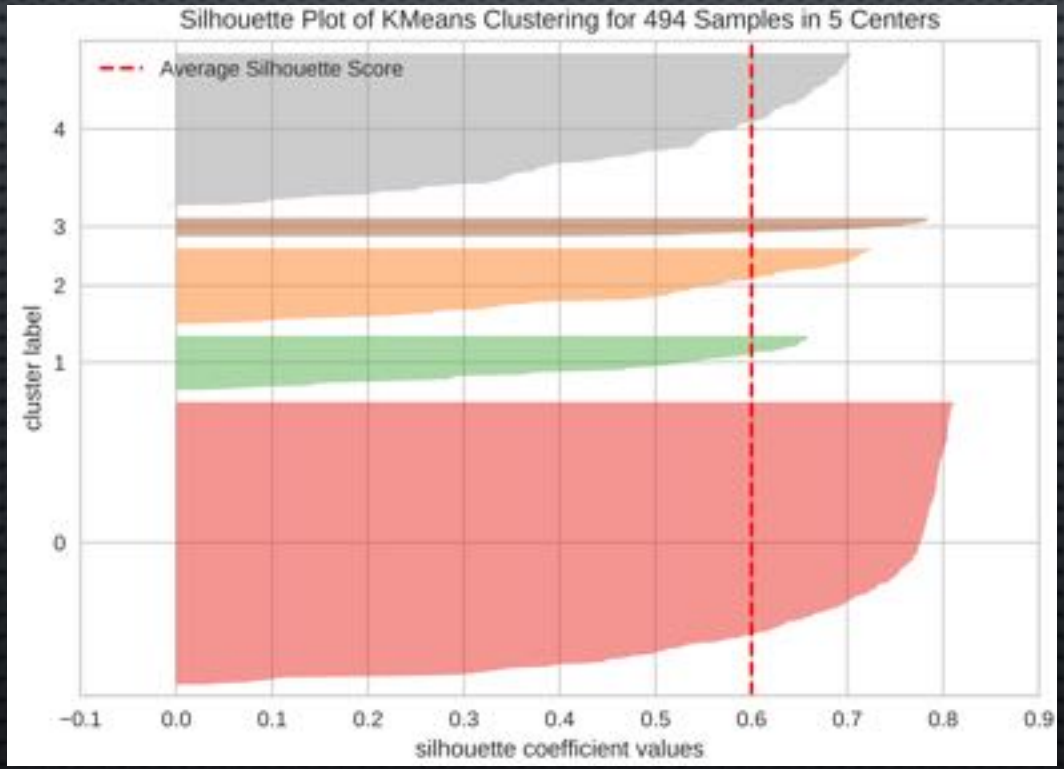
- The mean Silhouette coefficient across all the observations



Silhouette score as a function of k

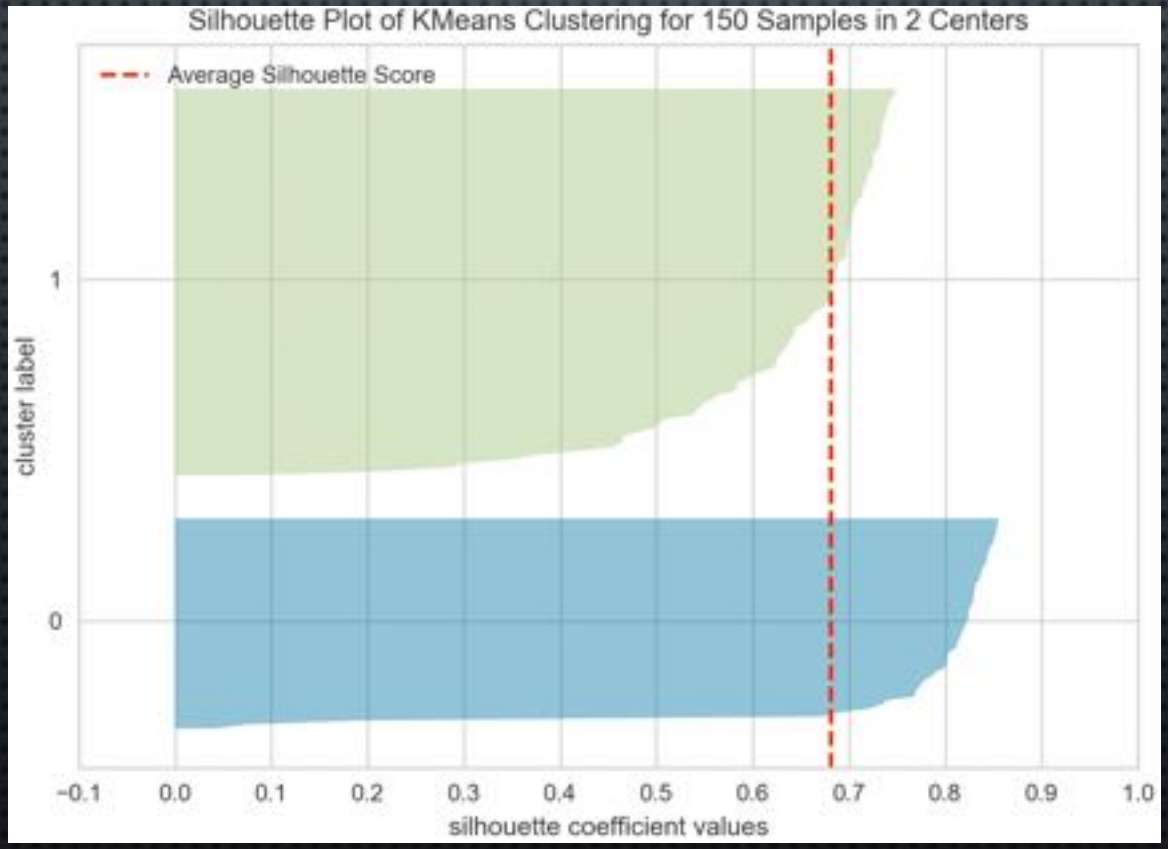
Finding the optimal k: the Silhouette diagram

- An in-depth look into Silhouette scores for a give k:



Finding the optimal k: the Silhouette diagram

- Iris data: $k = 2$

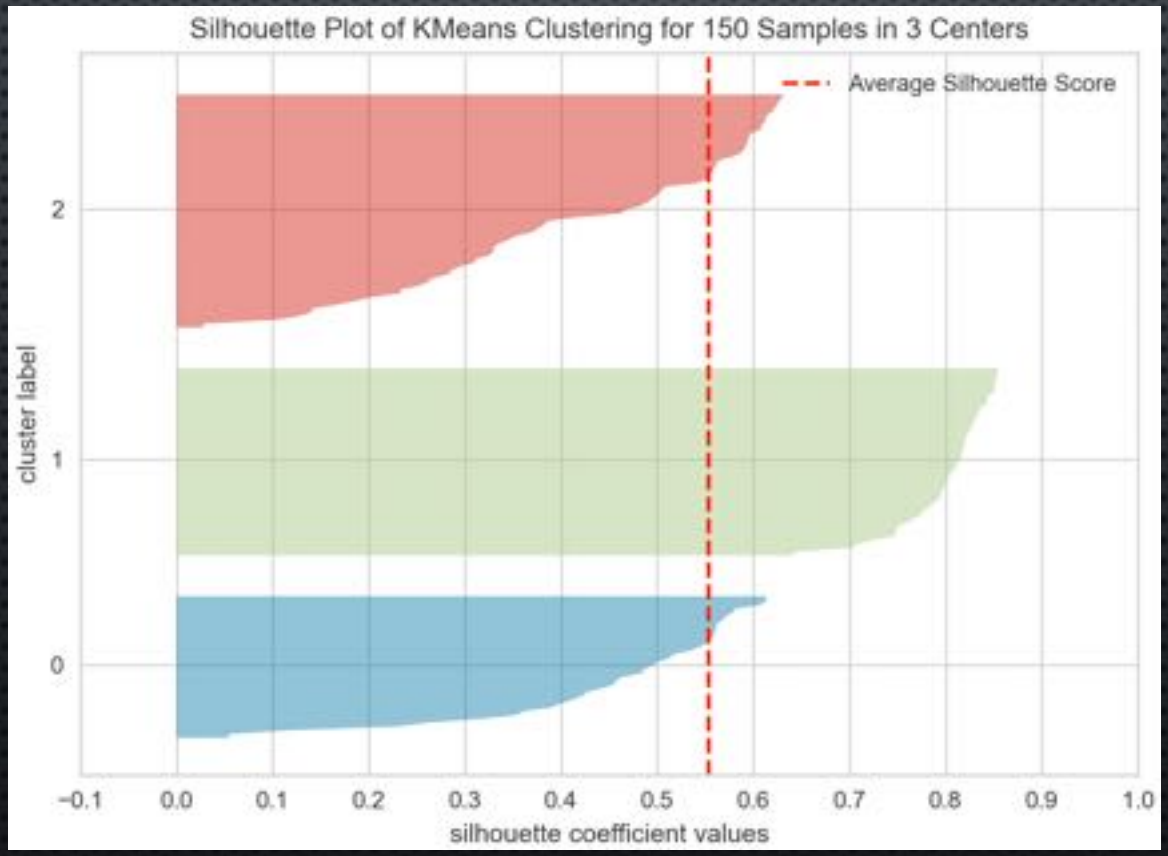


setosa

Finding the optimal k: the Silhouette diagram

- Iris data: $k = 3$

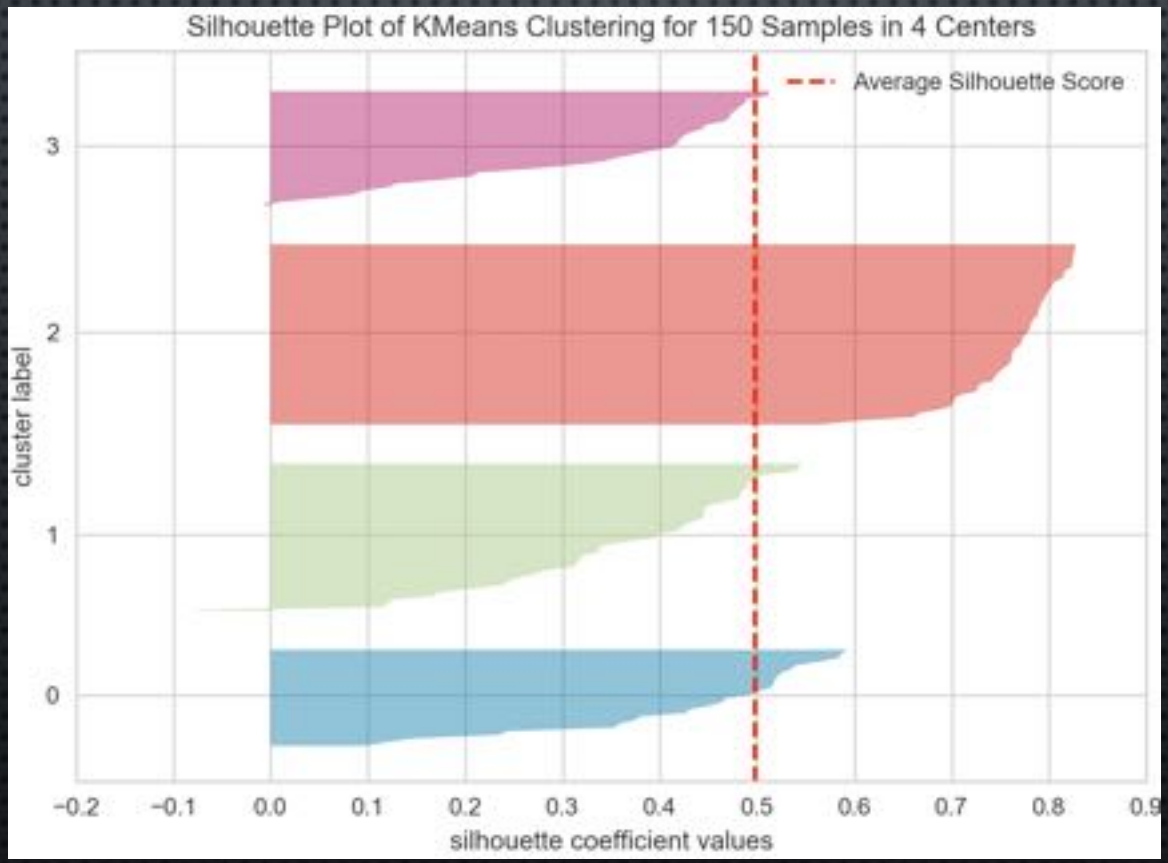
setosa



Finding the optimal k: the Silhouette diagram

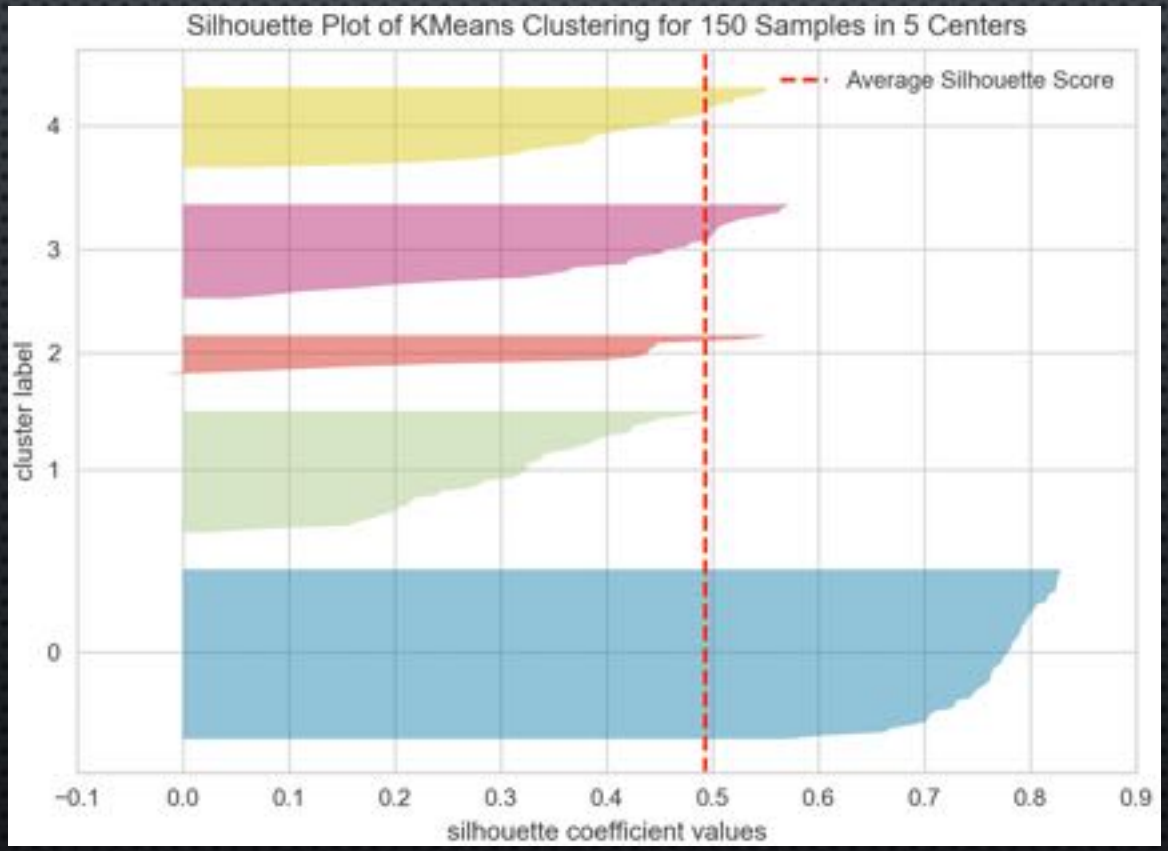
- Iris data: $k = 4$

setosa



Finding the optimal k: the Silhouette diagram

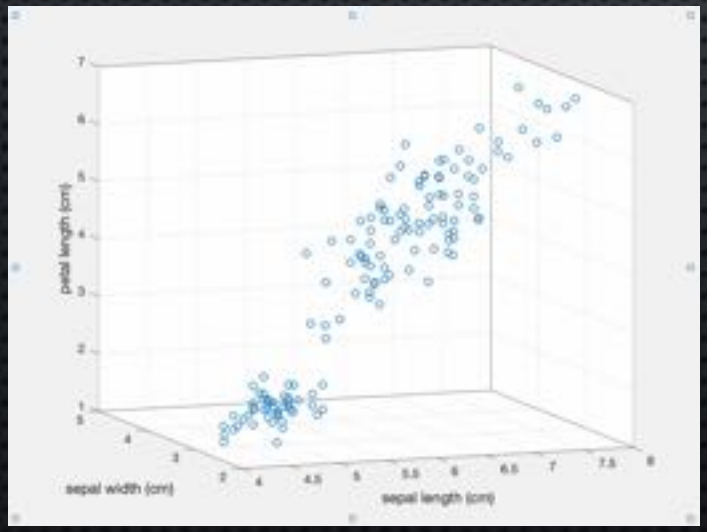
- Iris data: k=5

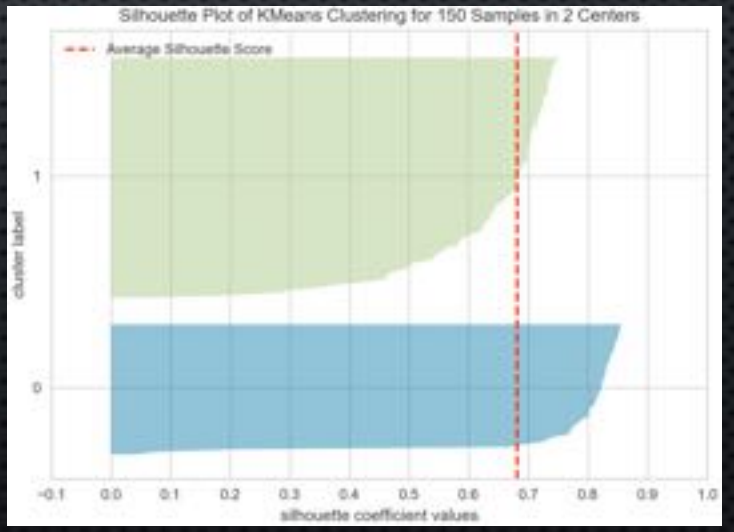


setosa

Hierarchical clustering

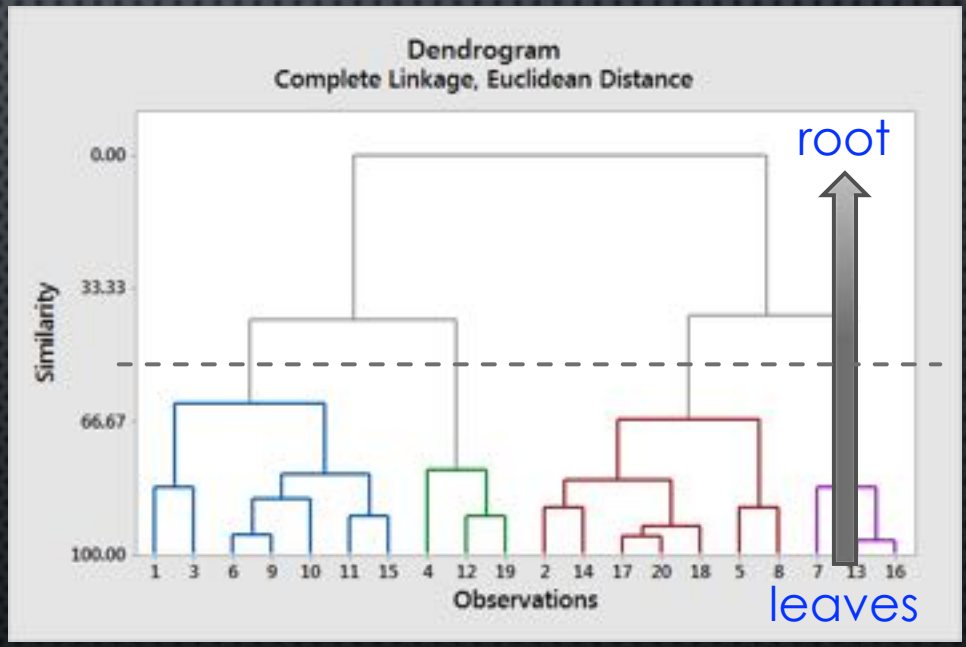
- Recall Iris data:
  - "setosa" tends to be an evident cluster
  - Less clear boundary between "versicolor" and "virginica"
  - Potential existence of sub-clusters





Hierarchical clustering

- Agglomerative clustering
  - Builds nested clusters by merging or splitting them successively.
  - This hierarchy of clusters is represented as a tree (visualised by dendrogram)
  - The root: the unique cluster that gathers all the samples; the leaves: the clusters with only one sample.



Hierarchical clustering: linkages

Type of Linkages:

- **Ward:** minimizes the variance of the clusters being merged
- **Maximum** or **complete:** minimizes the maximum distance between observations of pairs of clusters.
- **Average:** minimizes the average of the distances between all observations of pairs of clusters.
- **Single:** minimizes the distance between the closest observations of pairs of clusters.

- Each may produce different hierarchical structure

Hierarchical clustering: distance metric

Minkowski distance

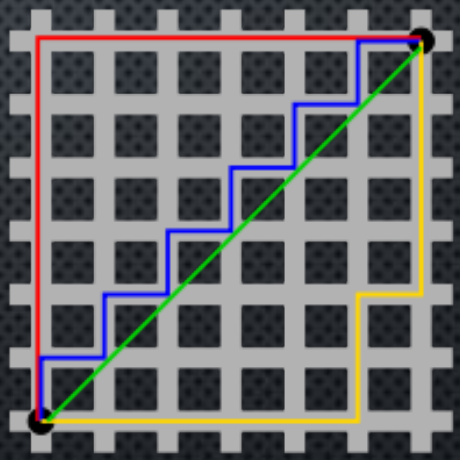
$$d(a,b) = \left(\sum_{i=1}^{n} |a_i - b_i|^p\right)^{\frac{1}{p}}$$

- Type of metrics
  - **Euclidean (L2)**

$$d(a,b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

  - **Manhattan (L1)**

$$d(a,b) = \sum_{i=1}^{n} |a_i - b_i|$$

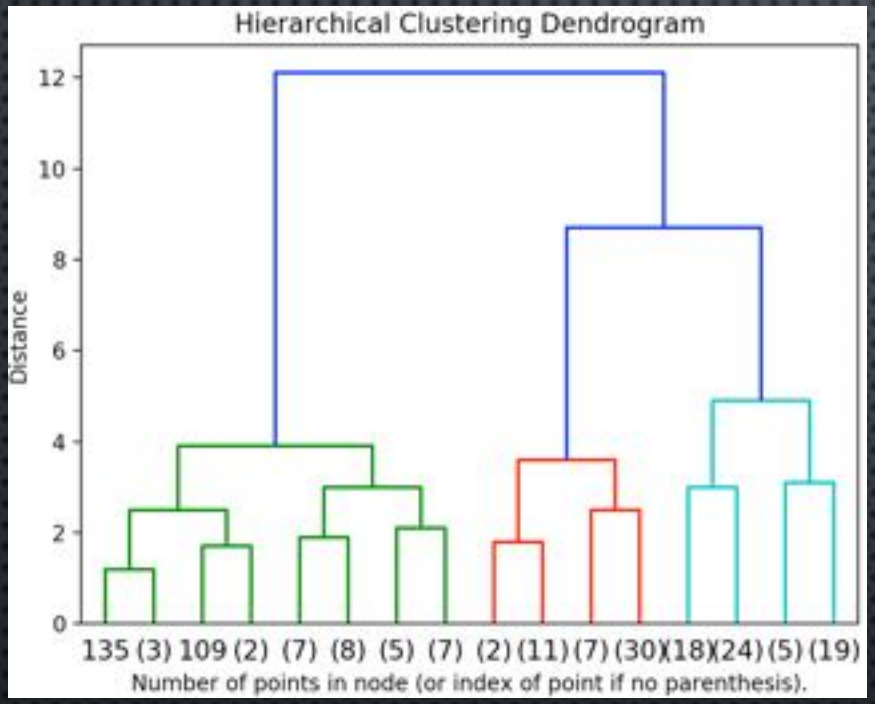


  - **Cosine**

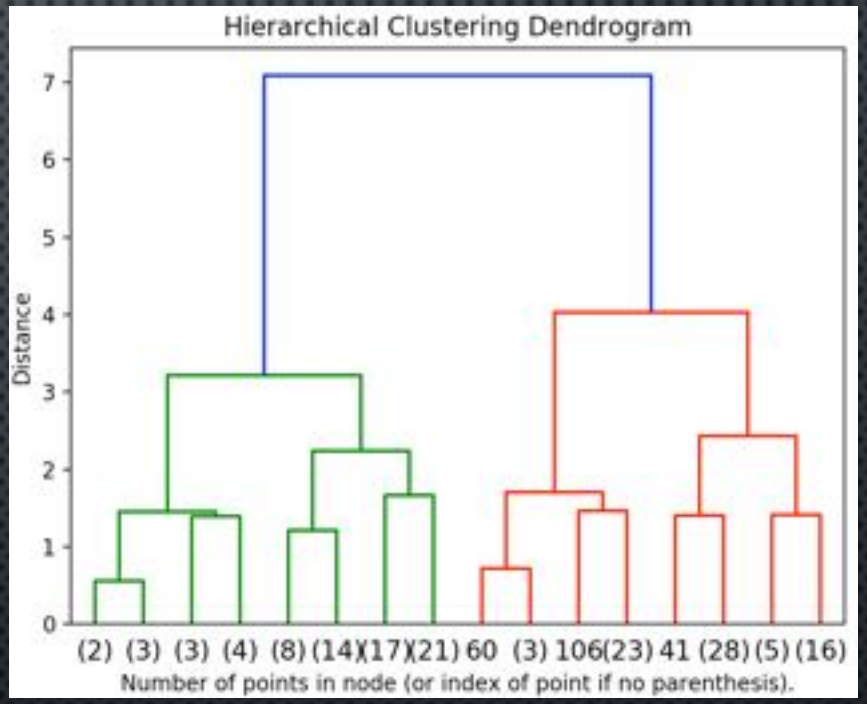
$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

- Some metrics can only be used by certain linkage

Hierarchical clustering: Iris data



d-metric : L1



d-metric: L2

- Linkage: “complete”
- Distance metric: “L1” and “L2”

Other clustering algorithms

- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)
- Spectral clustering
- DBSCAN
- Mean-shift
- …