UNIVERSITY OF TWENTE.

Introduction to Machine Learning and Clustering

Mahdi KHODADADZADEH
Assistant Professor
Faculty of Geo-Information Science and Earth Observation (ITC)
Department of Geo-information Processing (GIP)
m.khodadadzadeh@utwente.nl

May 2024

ITC FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

Why Kmeans is the most popular algorithm?

- From a theoretical standpoint, k-means is not a good clustering algorithm in terms of efficiency or quality (the running time, locally optimal solution, initialization, …)

- Why is it one of the top 10 algorithms in data mining? Why is it still popular even as datasets have grown in size?

➢ The advantage of k-means is its simplicity. In practice the speed and simplicity of k-means cannot be beat.

➢ Scaling k-means to massive data is relatively easy due to its simple iterative nature.

➢ Many works have focused on improving this algorithm.

Wu, Xindong, et al. "Top 10 algorithms in data mining." *Knowledge and information systems* 14.1 (2008): 1-37.
Bahmani, Bahman, et al. "Scalable k-means++." *arXiv preprint arXiv:1203.6402* (2012).

A better way to initialize

- Choosing the centers **one by one in a controlled fashion**.
- **k-means++** algorithm selects only the first center uniformly at random from the data.
- Each subsequent center is selected with a probability proportional to its contribution to the overall error given the previous selections.

Algorithm 1 k-`means++`(k) initialization.

1: $\mathcal{C} \leftarrow$ sample a point uniformly at random from X
2: **while** $|\mathcal{C}| < k$ **do**
3: Sample $x \in X$ with probability $\frac{d^2(x,\mathcal{C})}{\phi_X(\mathcal{C})}$
4: $\mathcal{C} \leftarrow \mathcal{C} \cup \{x\}$
5: **end while**

Arthur, David, and Sergei Vassilvitskii. *k-means++: The advantages of careful seeding*. Stanford, 2006.

Optimal k value

You may never find the right number of clusters but you can try to find an optimal one!

Run the algorithm for several consecutive number of clusters (k=1,2,3,…N).

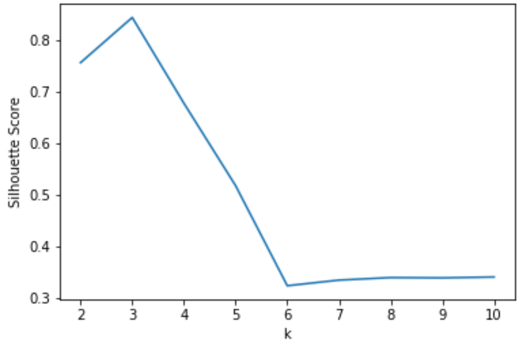
Compute the clustering performance for each number of clusters i.e., k.

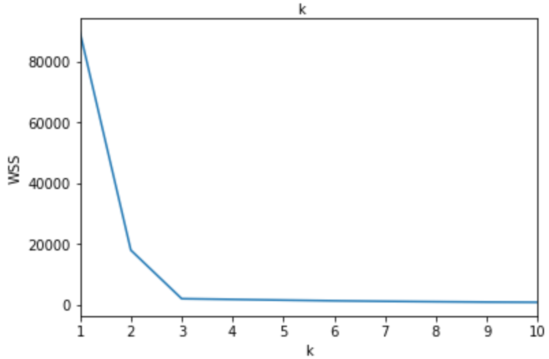
Determine the k such that it works well for your problem.

Silhouette Method → measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation) for different values of k. The Silhouette Score reaches its global maximum at the optimal k.

Elbow Method → calculates the Within-Cluster-Sum of Squared Errors (WSS) for different values of k and chooses the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.

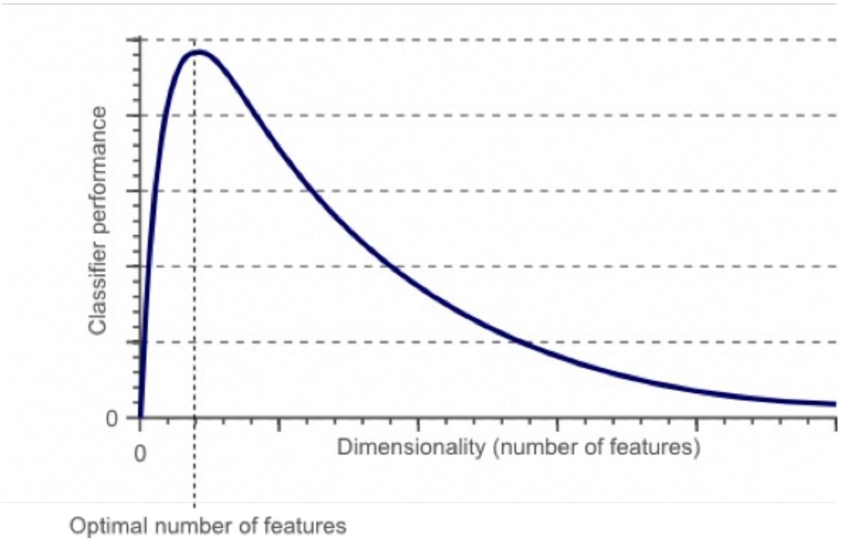
Chek out this link: https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb





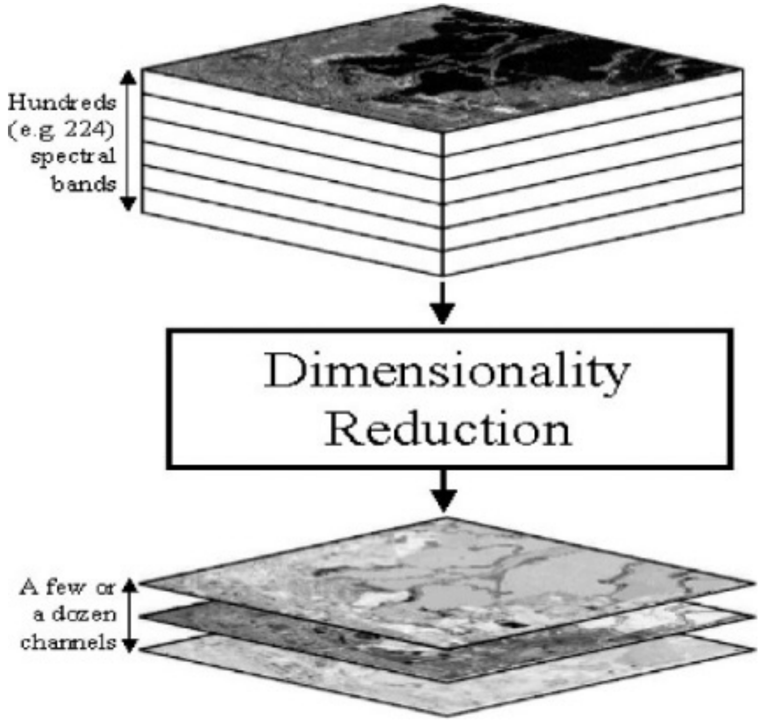
Feature Extraction

- Choosing discriminating and independent features is key to any machine learning algorithm
- In real applications usually many features are measured while only a very small percentage of them carry useful information towards our learning goal
- We usually need an algorithm that compress our feature vector and reduce its dimension
- Curse of dimensionality



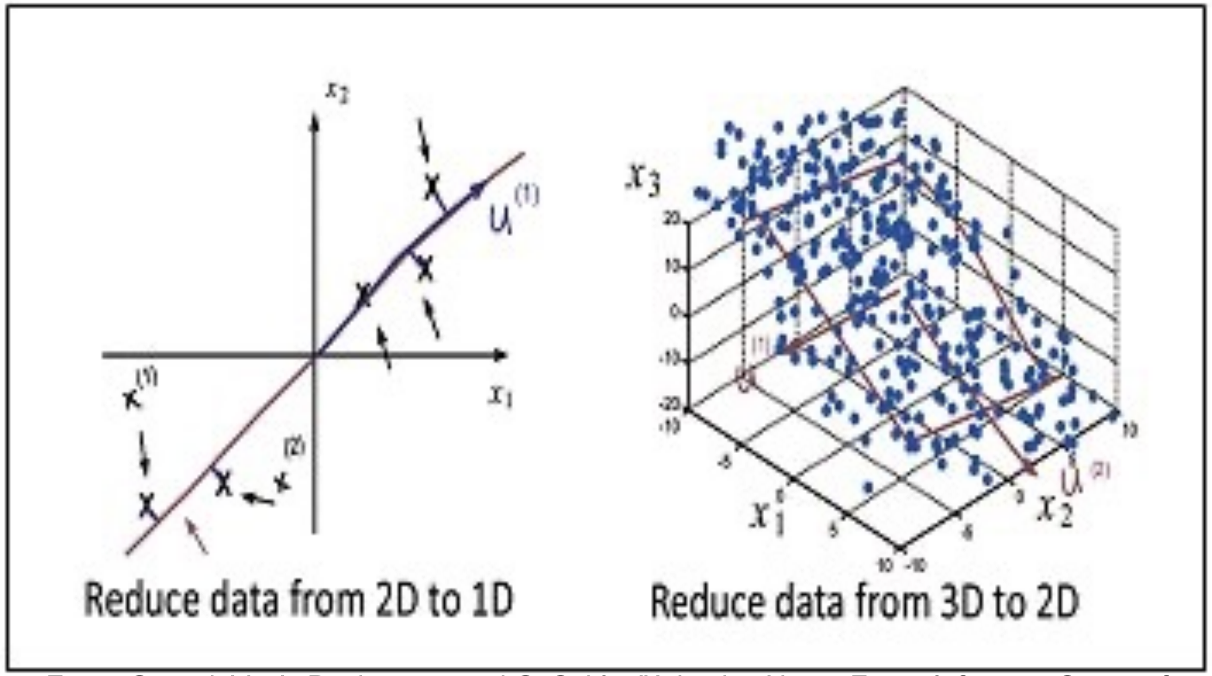
Principal component analysis

- Relatively simple and popular technique
- PCA converts a set of observations into a set of linearly uncorrelated variables, called principal components
- Represents data in a space that better describes the variation
- If a strong correlation between variables exists, the attempt to reduce the dimensionality is reasonable



Principal component analysis

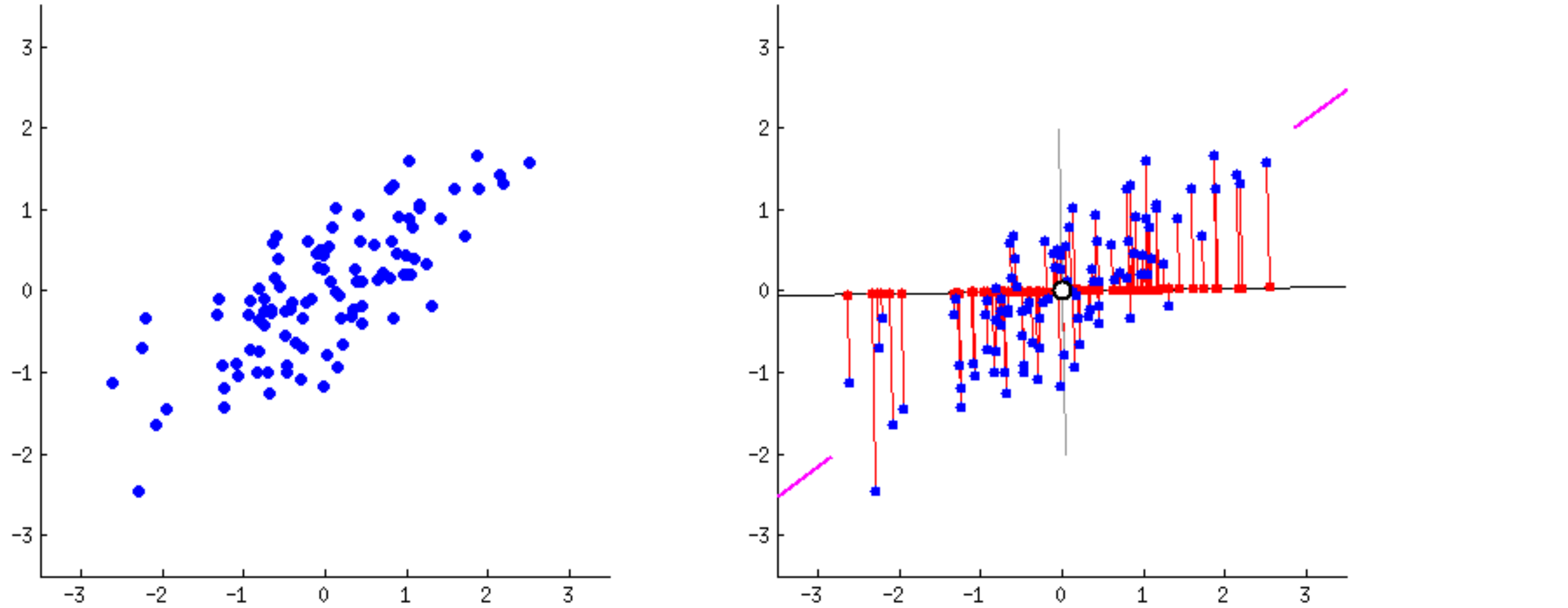
- Identifies directions of maximum variance (in high-dimensional data) and projects the data onto a smaller dimensional subspace while retaining most of the information.
- PCA projects the entire dataset onto a different feature (sub)space



From: Gosavi, V., A. Deshmane, and G. Sable. "Adaptive Neuro Fuzzy Inference System for Facial Recognition." *IOSR Journal of Electrical and Electronics Engineering* 14.3 (2019): 15-22.

Principal component analysis

- What the projections look like for different lines (red dots are projections of the blue dots)
- The reconstruction error are given by the length of the connecting red line



From: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Principal component analysis

- PCA is built on the concepts of eigenvector and eigenvalues
- Creates a projection matrix of the selected k eigenvectors.
- Transforms the original dataset X via the projection matrix and obtains a k-dimensional feature subspace Y

See this link for more details:

https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643