Unsupervised Learning

Presented by:
Talha Mehmood
Zaupash Tariq

Table of Contents

cbt

Unsupervised Learning

- Background
- Definition
- More subjective
- Cannot assess results
- Customer segmentation

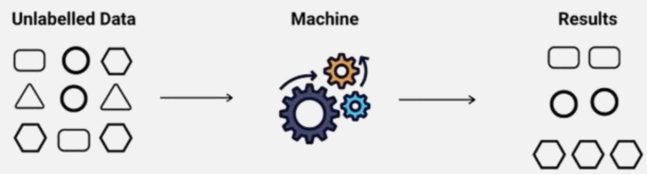


Figure 1: Unsupervised learning

Backgorund: In previous chapters we have discussed supervised learning techniques where output variable y is known in unsupervised learning output variable y is unknown here our goal is to discover subgroups among the variables or among the observations?

Definition: unsupervised learning algorithm analyzes and finds patterns in data without predefined outputs

More subjective: Unsupervised learning tends to be more subjective, and there is no simple goal for the analysis, such as prediction of a response Unsupervised learning is often performed as part of an exploratory data analysis

Cannot assess results: Furthermore, it can be hard to assess the results obtained from unsupervised learning methods, since there is no universally accepted mechanism for validating results as true answer is unknown

Customer preferences example : unsuperivsed learning can be used for segmenting customers based upon their preferences by placing Customers who buy similar products in same subgroup

Here we will discuss two particular types of unsupervised learning: principal components analysis and clustering. First we discuss principal components analysis

Principal Components Analysis

- Definition
- Why use PCA:
  - Visualization
  - Noise reduction
  - Feature extraction

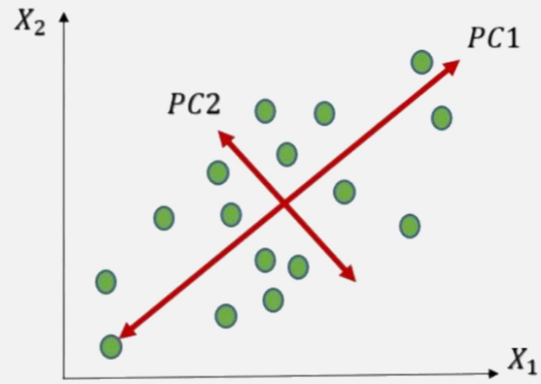
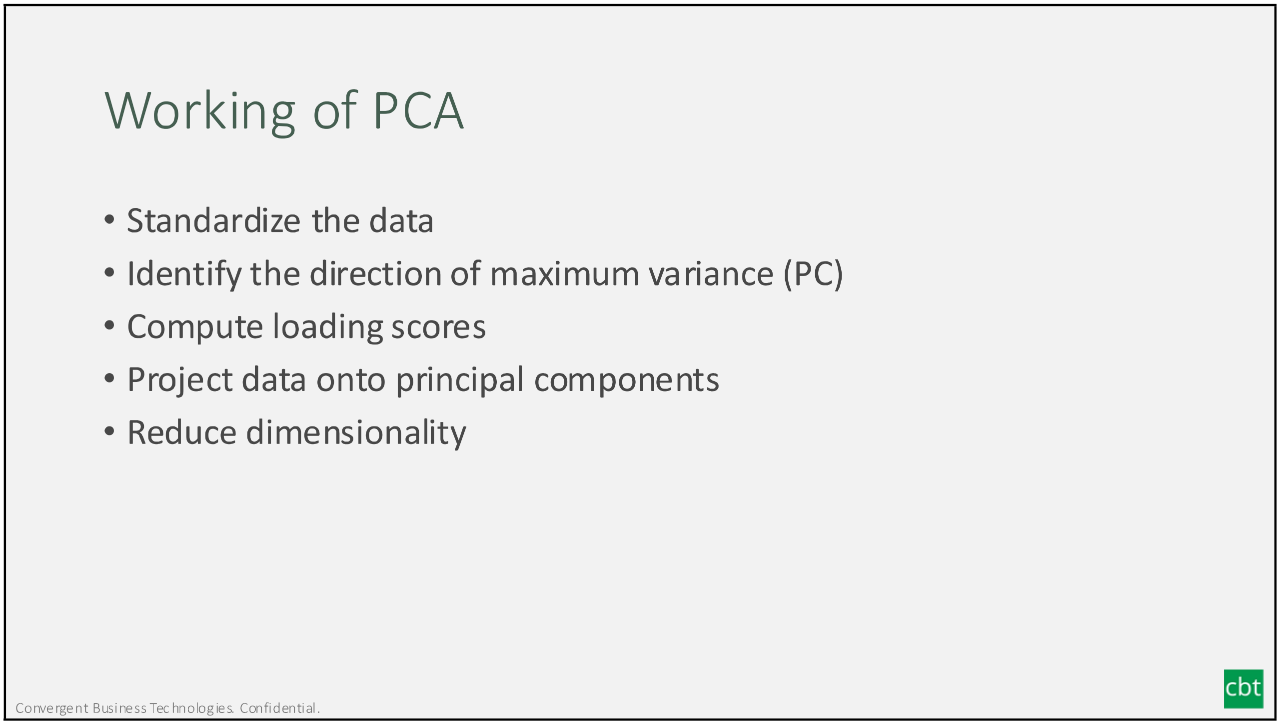


Figure 2: Principal component analysis

principal components analysis is a technique that finds smaller number of representative variables that collectively explain most of the variability in the original set

Why Use PCA?

- **Visualization:** Reduces high-dimensional data to lower dimensions eg 2d and 3d for easier visualization.
- **Noise Reduction:** Removes noise and irrelevant information.
- **Feature Extraction:** Creates new, uncorrelated features that capture the most variance in the data.
- Now we discuss working of pca



1. Standardize the Data
First, PCA standardizes each variable so they all have a mean of zero and a standard deviation of one
2 Identify the direction of maximum variance (PC)
PCA then looks for the directions, or "axes," along which the data varies the most. These directions are called **principal components** (PCs)
3 Compute Loading Scores
Then it computes loading scores. The **loading scores** tell us how strongly each original variable relates to each principal component
4. Project the Data onto the Principal Components
PCA transforms the original data points by "projecting" them onto the principal components. This creates new data points (called **principal component scores**) in a reduced-dimensional space
5. Reduce Dimensionality
The final step is to decide how many principal components to keep. Often, just a few components capture most of the data's variability so we keep them
In last step we reduce dimensionality by selecting those components that capture most of the variance of data
Now I give you an example of pca

PCA Example

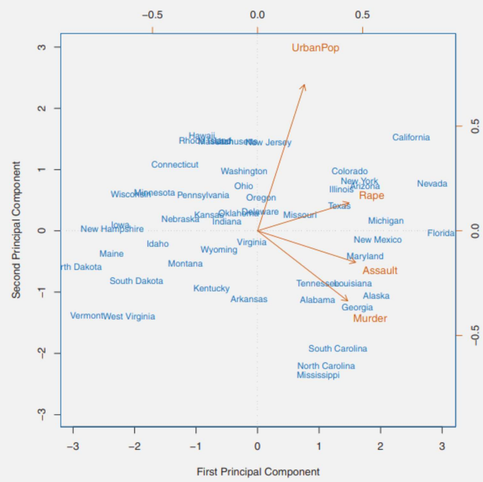


Figure 3: First two principal components for US Arrests data

cbt

The figure shows first two principal components for the USArrests data. The blue state names represent the scores for the first two principal components. The orange arrows indicate the first two principal component loading vectors
Here we can see loading of rape, assault and murder(which can be referred as crime rate) is more on first principal component and loading of urban population is more on second principal component
Here we can see florida has high crime rate and new jersey has high urban population

Now we discuss another interpretation of principal components

Another Interpretation of Principal Components

- Direction of maximum variance
- Surfaces closest to observations

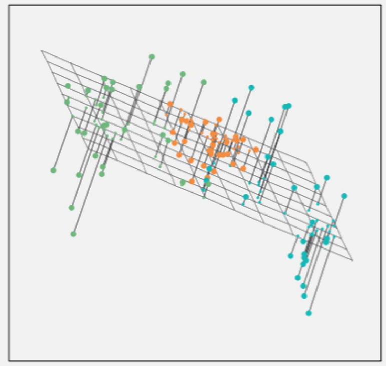


Figure 4: Principal component surface

cbt

previously, we described the principal component as the directions along which the data vary the most
However, an alternative interpretation for principal components can also be useful: principal components provide low-dimensional linear surfaces that are closest to the observations.
Now we discuss more detail of pca

More on PCA

- Scaling the variables
- Uniqueness of principle components
- Proportion of variance explained
- Deciding how many principal components to use
- Other uses of PCA

cbt

Uniqueness of principle components: Hence, if the sign is flipped on both the loading and score vectors, the final product of the two quantities is unchangedor result remains unchanged
Proportion of variance explained: we examine the proportion of variance explained by each principal component
Deciding how many principal components to use: We select principal components that explain most of the variance of data
Other uses of pca : **principal components** can be used in a variety of statistical methods, like **regression, classification, and clustering for reducing dimensionality and selecting important features only**

Clustering Methods

- Unsupervised technique
- Finds subgroups in data
- Groups by similarity
- Methods:
    - K-means
    - Hierarchical

cbt

Clustering is a versatile **unsupervised learning** technique that helps us find **hidden patterns** or **subgroups** in data by grouping similar observations. It's widely used in fields like **healthcare**, where it can reveal disease subtypes, or **marketing**, where it helps segment customers for targeted advertising. Unlike methods like **PCA** that reduce dimensionality, clustering specifically seeks **homogeneous groups**.
With **K-means clustering**, we can pre-define the number of clusters for clear segmentation. **Hierarchical clustering**, on the other hand, offers flexibility, creating a **dendrogram** that reveals clustering at multiple levels.

K-Means Clustering

- Partitions data into K clusters
- Non-overlapping clusters
- Based on similarity

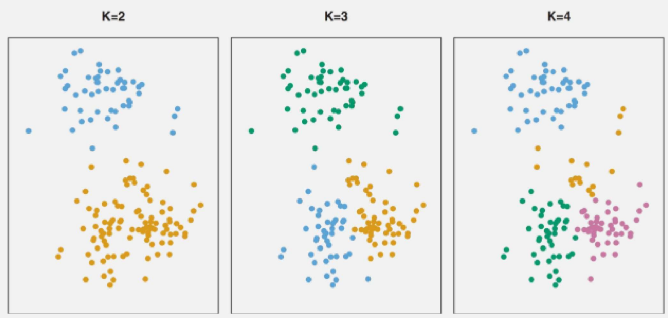


Figure 5: K-means clustering with different values of K

K-means clustering groups data into **K unique clusters** by assigning each point to the closest **cluster center** or centroid. It's a simple, effective way to separate data points based on **similarity**, resulting in non-overlapping, distinct groups

Algorithm Steps

- Randomly assign initial clusters
- Iterate
  - Compute centroid
  - Reassign observations
- Repeat until stable

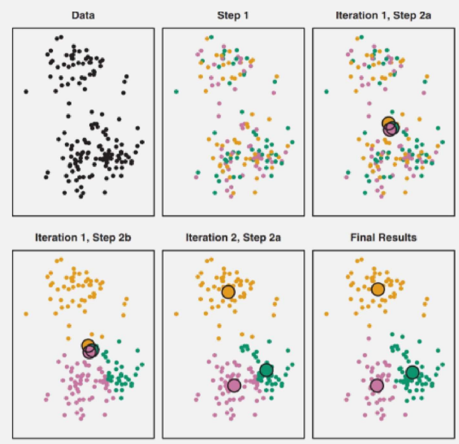


Figure 6: The progress of the K-means algorithm

cbt

The K-means clustering algorithm starts by **randomly assigning points** to initial clusters. It then alternates between **calculating centroids** for each cluster and reassigning points to cluster which is **closest centroid of that cluster**. This process repeats until the clusters **stabilize** and no points change clusters.
Key considerations include the **initial cluster assignments** and the choice of **K**, as both can affect results. Running the algorithm multiple times with different initial setups helps ensure a **strong, stable grouping**. Choosing the right number of clusters, K, is also crucial for **accurate and meaningful clusters**.

Hierarchical Clustering

- Flexible clusters
- Tree structure (dendrogram)
- Types:
  - Bottom-up
  - Top-down

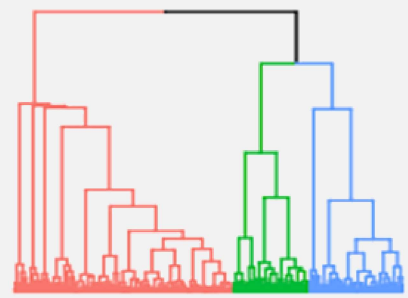


Figure 7: Hierarchical clustering display via dendrogram

cbt

Algorithm Steps

- Define dissimilarity measure
- Each observation starts as own cluster
- Iterative Merging:
  - Find closest clusters by dissimilarity
  - Fuse clusters to form new n-1 clusters
  - Repeat until one cluster remains

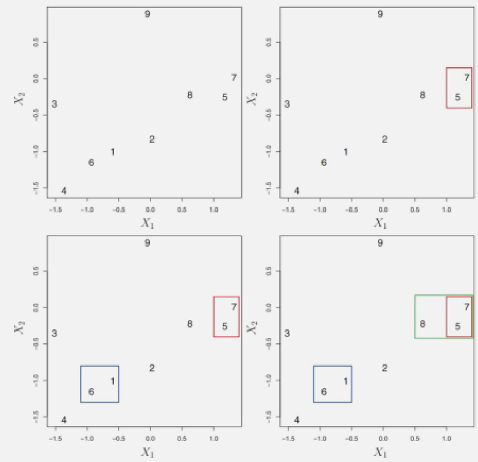


Figure 8: Some steps of hierarchical clustering algorithm

In hierarchical clustering, we start by defining a **dissimilarity measure** like Euclidean distance, with alternatives like correlation or Manhattan distance. **Each data point begins as its own cluster**. The algorithm then iteratively **finds the closest clusters** based on this measure and merges them, reducing the number of clusters by one at each step. This continues until all observations form a single cluster. During the process, we record **fusion heights** to build a dendrogram, where each height reflects the **inter-cluster dissimilarity** at that merge point. This provides a visual map of clustering steps and distances.

Considerations

- Linkage types:
  - Complete
  - Single
  - Average
  - Centroid
- Scaling variables

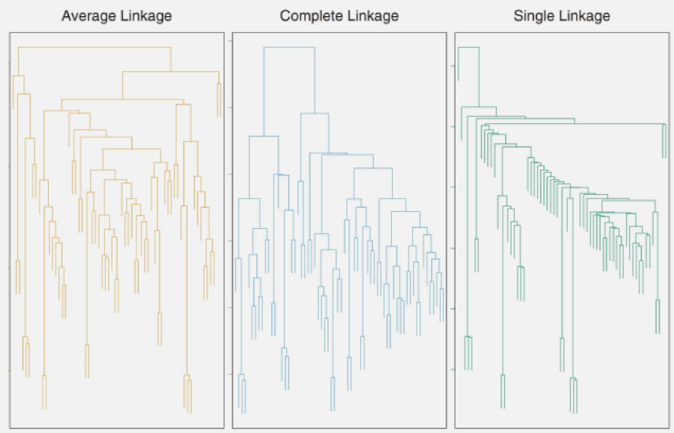


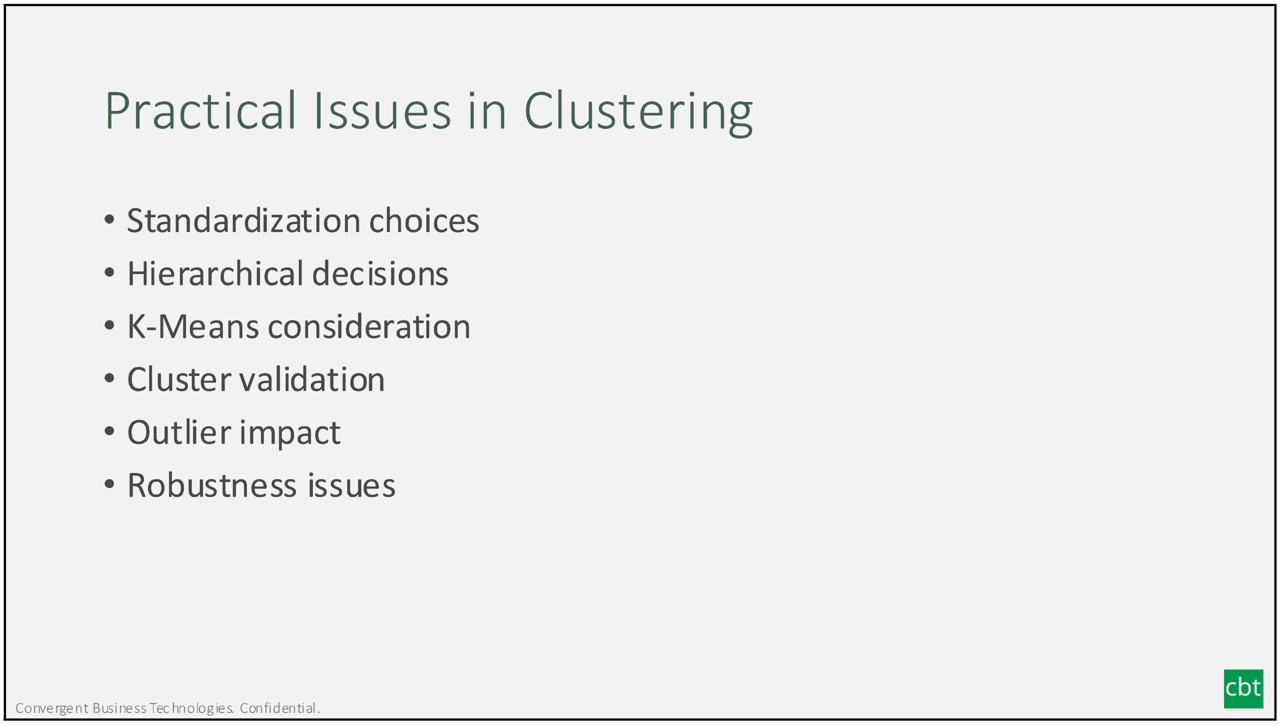
Figure 9: Average, complete, and single linkage

cbt

When performing hierarchical clustering, there are several **key considerations**.
First, the choice of **linkage type** influences the clustering results:

- **Complete linkage** considers the largest distance between clusters, promoting compact clusters.
- **Single linkage** uses the smallest distance, which can lead to long chains or 'chaining' effects.
- **Average linkage** balances the clusters by using the mean distance.
- **Centroid linkage** calculates the distance between cluster means, offering a different perspective on cluster formation.

Next, the **dissimilarity measure choice** is crucial and should align with your data type and goals. For instance, using **correlation distance** can be beneficial when analyzing similarities in preferences, regardless of their frequency.
Finally, it’s important to **scale variables** appropriately. Standardizing features with different units ensures that no single variable, especially those with high frequency, disproportionately influences the clustering results. These considerations are vital for effective clustering outcomes.



When utilizing clustering techniques for data analysis, it's essential to address several **practical issues** that can significantly influence your results. Decisions regarding **standardization** are crucial, as properly centering and scaling your data can enhance clustering effectiveness.

In hierarchical clustering, selecting the appropriate **dissimilarity measure** and **linkage type** is vital, as is determining where to cut the dendrogram for meaningful clusters.

Similarly, with K-means clustering, one must carefully decide on the **number of clusters** to analyze.

Validation of clusters is another key consideration: we must question whether the identified clusters reflect genuine subgroups or merely result from random noise. Techniques to assess p-values can help, though there is no consensus on a single best method.

Moreover, be mindful of **outliers**, as they can distort clusters. Utilizing **mixture models** may help account for such anomalies.

Clustering methods can also lack robustness, leading to different cluster formations with slight variations in data.

Conclusion

- Learning without labels
- Principal component analysis
- Clustering methods
- Practical issues

cbt

In this overview of **unsupervised learning**, we explore its definition as a method of analyzing data without labeled outcomes. The two key techniques highlighted are **Principal Component Analysis (PCA)** and various **clustering methods**.
PCA is essential for **dimensionality reduction**, allowing for easier interpretation of data through the identification of principal components and their alternative interpretations.
When examining **clustering**, we focus on **K-means** and **hierarchical clustering** methods, emphasizing the importance of understanding practical issues as well.

Thank you.

cbt