Introduction to Machine Learning with Python

David Schaupp | WS2025

Content

```
Introduction to Machine Learning:
  - What is Machine Learning?
  - Machine Learning Project Checklist
Intro´s:
  - Colab
  - Python
  - Numpy
  - Matplotlib
Supervised Learning:
  - Classification (Binary|Multiclass)
  - Regression
  - Support Vector Machines
  - Decision Trees (Random Forest)
Unsupervised Learning:
  - Dimensionality Reduction
  - Clustering (k-means, DBSCAN)
Performances Measures:
  - Accuracy
  - Precision
  - Recall
  - F1-Score
  - ROC Curve
  - Confusion Matrix
```

Supervised vs Unsupervised Learning

```
- Supervised Learning:
  - Dataset is labeled
  - Labels:
    - Spam/Not Spam
  - Applications:
    - Spam filter
    - Image classification
  Pro:
    - Easy to understand and interpret
  Con´s:
    - Dependency on high quality labeled data
  Algorithms:
    - Decision Trees
    - Random Forest
    - Support Vector Machines
    - ...
```

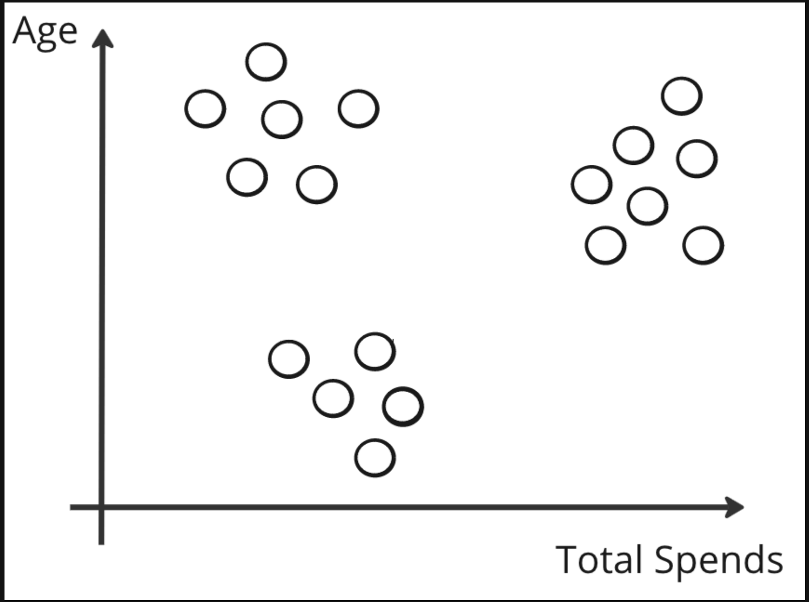
```
- Unsupervised Learning:
  - Training data is unlabeled
  - Applications:
    - Grouping customers by purchasing behavior
    - Recommender systems f.e. Netflix, Spotify
  Pro:
    - No need for labeled data
  Con´s:
    - More difficult to understand and interpret
  Algorithms:
    - Clustering
    - Dimensionality Reduction
    - ...
```

Clustering

```
- Method to group similar objects together
- Goal: Objects in the same group are more similar to each other than to those in other groups
- Characteristics:
  - Unsupervised Learning
  - No predefined classes
  - No labeled data
- Applications:
  - Market Segmentation: Grouping customers by purchasing behavior
  - Social Network Analysis: Identifying communities of interest
  - Anomaly Detection: Identifying unusual patterns f.e. fraud detection
- Typers of Clustering:
  - Hierarchical Clustering
  - K-Means Clustering
  - DBSCAN
```

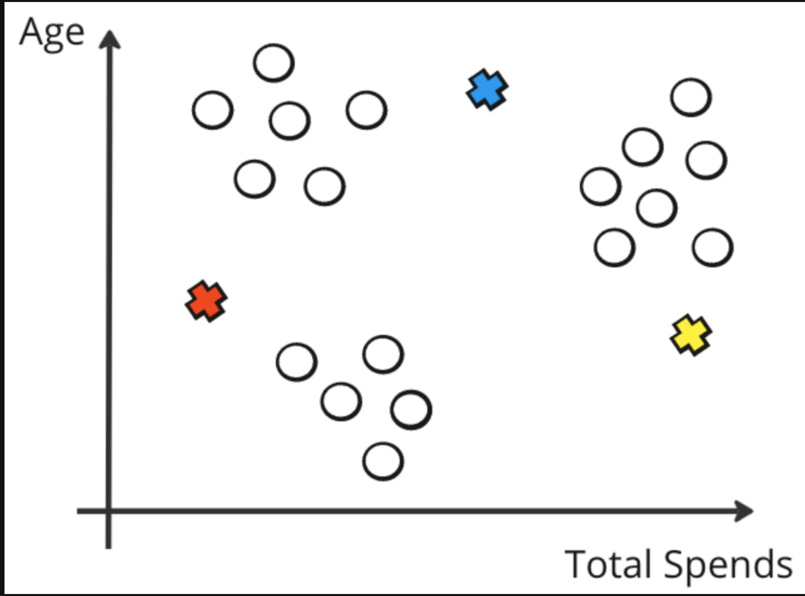
k-means Clustering

```
- Steps:
  - 1. Pick a number of clusters (k)
  - 2. Initialize centroids (center of clusters)
  - 3. Calculate the distance between each data point and the centroids
  - 4. Assign each data point to the closest centroid
  - 5. Move the centroids to the center of the points assigned to it
  - 6. Repeat steps 3, 4 and 5 until the centroids don´t move anymore
```



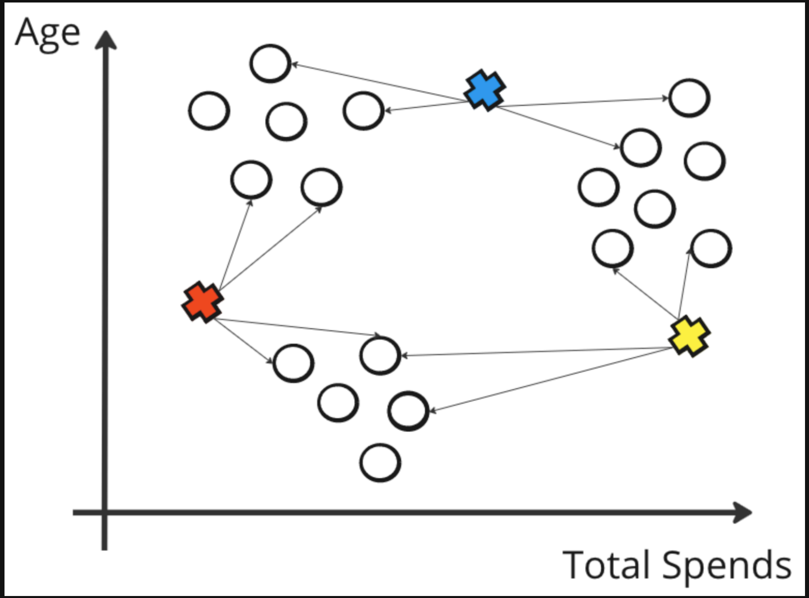
k-means Clustering

```
- Example
  - Step 1: Pick a number of clusters (k)
    - k = 3
  - Step 2: Initialize centroids (center of clusters)
```



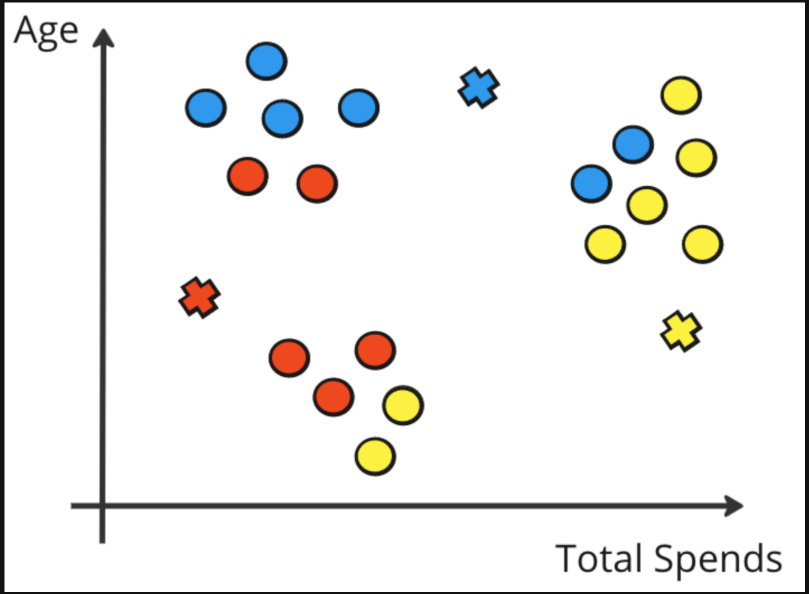
k-means Clustering

```
- Example
  - Step 3: Calculate the distance between each data point and the centroids
```



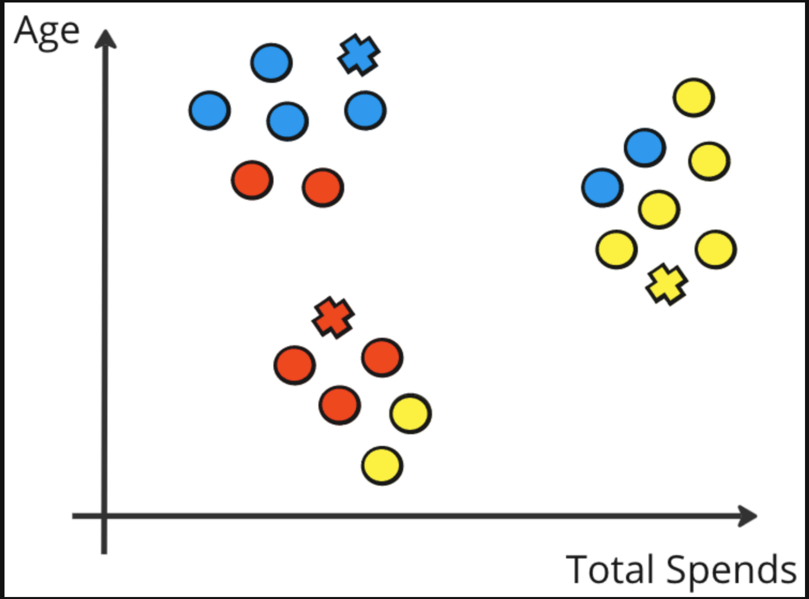
k-means Clustering

- Example
  - Step 4: Assign each data point to the closest centroid



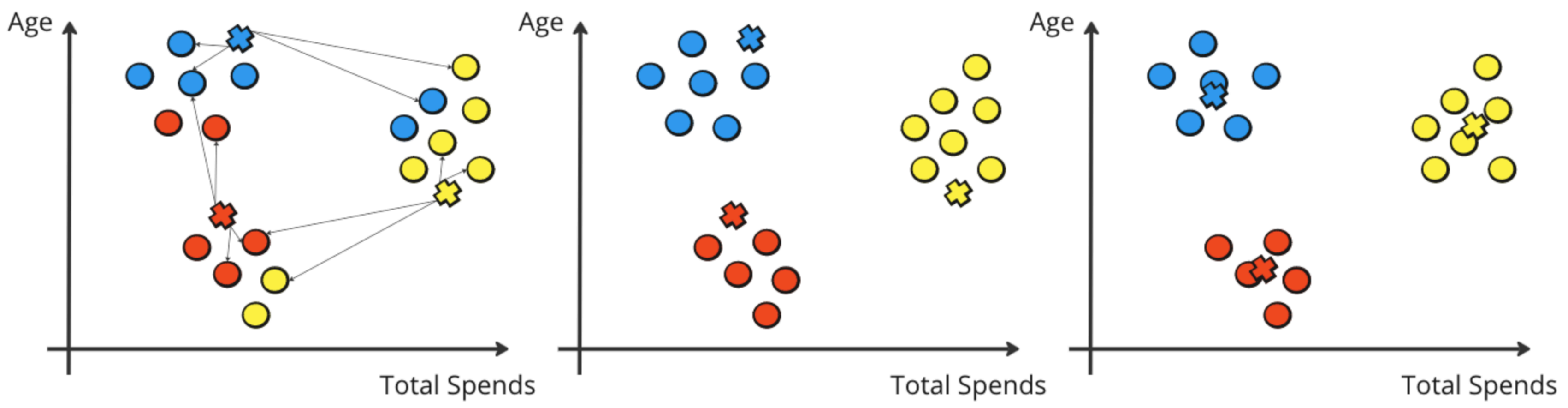
k-means Clustering

- Example
  - Step 5: Move the centroids to the center of the points assigned to it



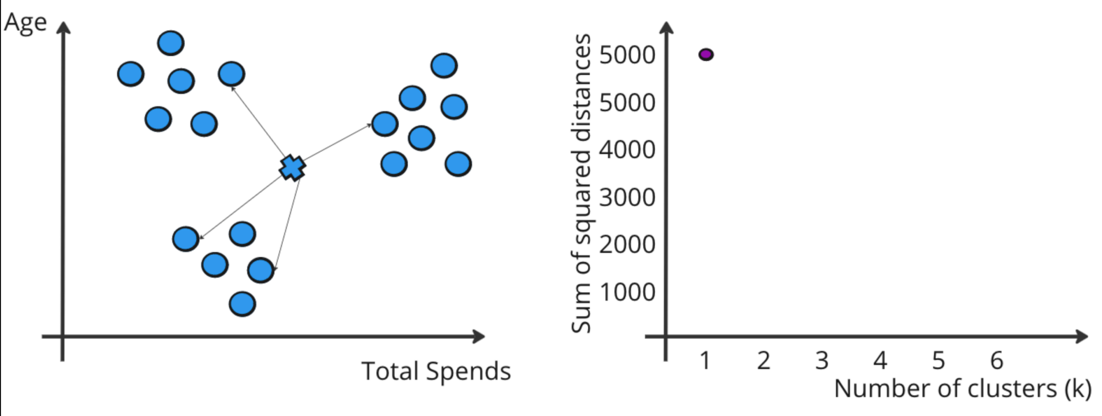
k-means Clustering

- Example
  - Step 3: Calculate the distance between each data point and the centroids
  - Step 4: Assign each data point to the closest centroid
  - Step 5: Move the centroids to the center of the points assigned to it



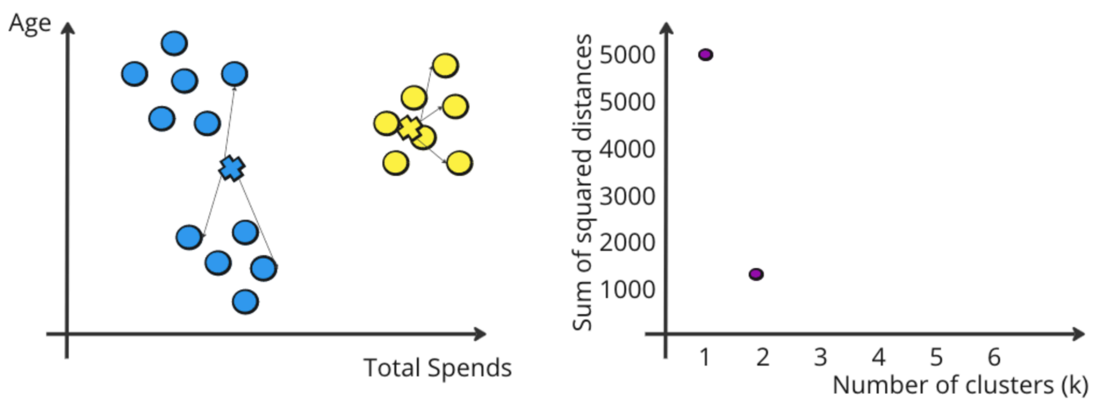
k-means Clustering

- Step 1: Pick a number of clusters (k)
  - Most important hyperparameter in k-means clustering
  - How to pick the right number of clusters k?
    - Elbow Method
      - Plot the number of clusters vs. the sum of squared distances
      - Pick the number of clusters where the sum of squared distances doesn´t decrease significantly anymore
    - Steps:
      - 1. Perfoms k-means with k=1, k=2, k=3, ...
      - 2. For each k, calculate the sum of squared distances
      - 3. Plot the number of clusters vs. the sum of squared distances
      - 4. Pick the number of clusters where the sum of squared distances doesn´t decrease significantly anymore



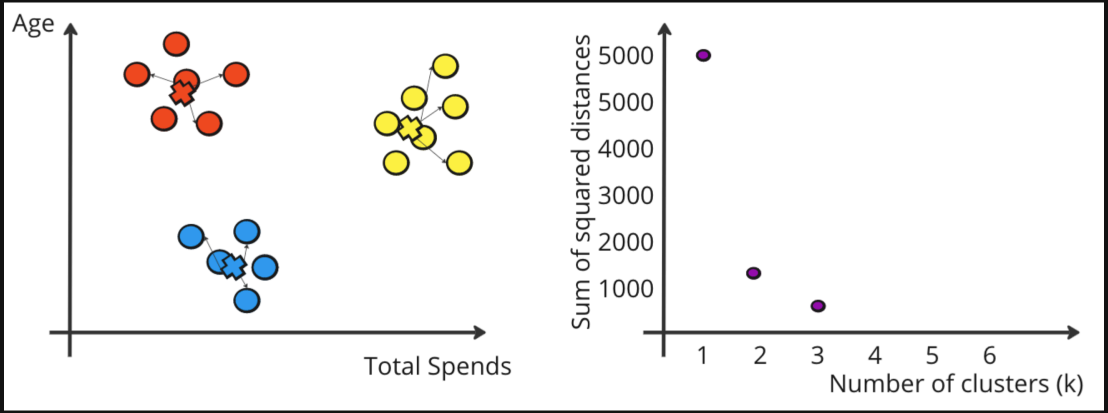
k-means Clustering

- Step 1: Pick a number of clusters (k)
  - Most important hyperparameter in k-means clustering
  - How to pick the right number of clusters k?
    - Elbow Method
      - Plot the number of clusters vs. the sum of squared distances
      - Pick the number of clusters where the sum of squared distances doesn´t decrease significantly anymore
    - Steps:
      - 1. Perfoms k-means with k=1, k=2, k=3, ...
      - 2. For each k, calculate the sum of squared distances
      - 3. Plot the number of clusters vs. the sum of squared distances
      - 4. Pick the number of clusters where the sum of squared distances doesn´t decrease significantly anymore



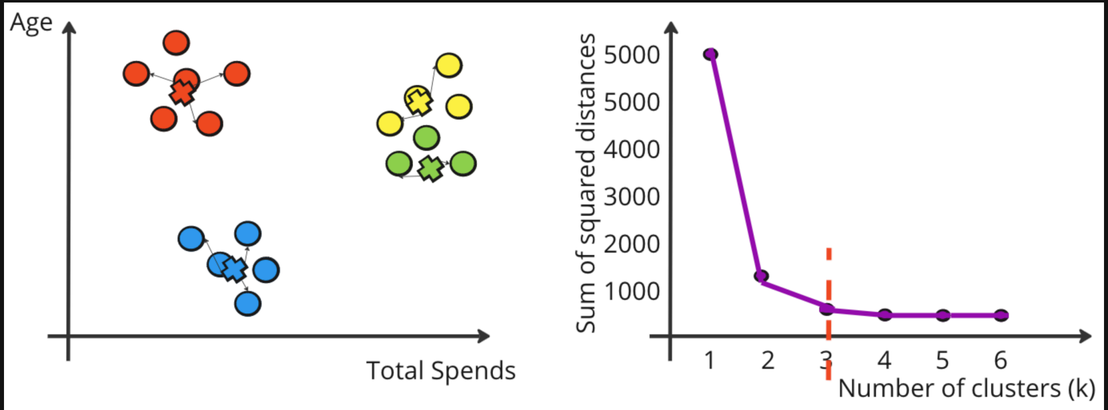
k-means Clustering

- Step 1: Pick a number of clusters (k)
  - Most important hyperparameter in k-means clustering
  - How to pick the right number of clusters k?
    - Elbow Method
      - Plot the number of clusters vs. the sum of squared distances
      - Pick the number of clusters where the sum of squared distances doesn´t decrease significantly anymore
    - Steps:
      - 1. Perfoms k-means with k=1, k=2, k=3, ...
      - 2. For each k, calculate the sum of squared distances
      - 3. Plot the number of clusters vs. the sum of squared distances
      - 4. Pick the number of clusters where the sum of squared distances doesn´t decrease significantly anymore



k-means Clustering

- Step 1: Pick a number of clusters (k)
  - Most important hyperparameter in k-means clustering
  - How to pick the right number of clusters k?
    - Elbow Method
      - Plot the number of clusters vs. the sum of squared distances
      - Pick the number of clusters where the sum of squared distances doesn´t decrease significantly anymore
    - Steps:
      - 1. Perfoms k-means with k=1, k=2, k=3, ...
      - 2. For each k, calculate the sum of squared distances
      - 3. Plot the number of clusters vs. the sum of squared distances
      - 4. Pick the number of clusters where the sum of squared distances doesn´t decrease significantly anymore



Content