HOGESCHOOL UTRECHT

Fundamentals of Machine Learning
Week 7: Unsupervised learning
Working on final project

Jonas Moons

Evalytics survey

- https://app.evalytics.nl/#/login
- Choose: Evaluate with code
- Code: rkd-256

Program

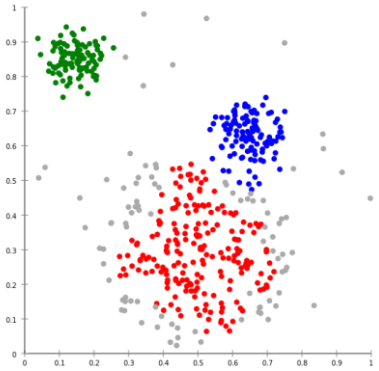
- 0-1.5 hours: short lecture
  - Clustering
  - Presentations
- >1.5 hours: project supervision / help in small groups
- Participants in AI research project get separate supervision next week

Supervised vs. unsupervised learning

- Supervised: use known patterns to predict new cases
  - Handwriting recognition

MNIST data set

- Unsupervised: you let the algorithm discover patterns/clusters on its own
  - Spotify Radio / Discover Weekly



Spotify®

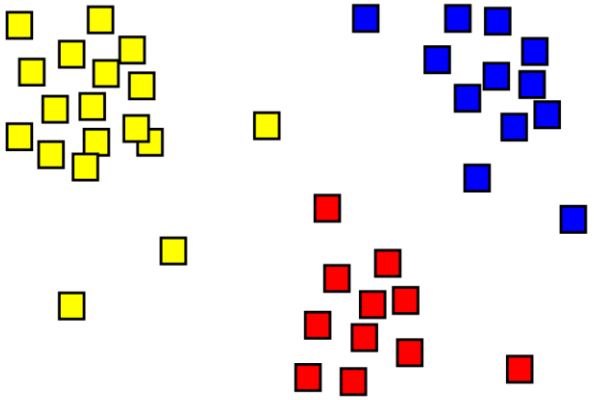
Why clustering?

- To discover interesting and useful clusters in the data and adapt our content or marketing strategy

- For instance, different user types based on behavior (user profiling), e.g.
  - Explicit: likes, favorites, ratings, comments, etc.
  - Implicit: pages visited, content seen, mouse movements, etc.

What is clustering?

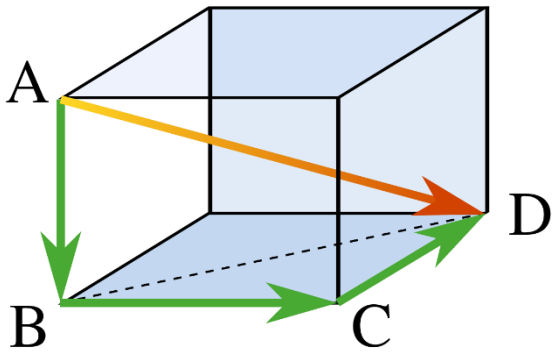
- Assign observations to several clusters
- It's not always clear which solution is 'correct'
- There are many algorithms; we will use *k*-means (simple to understand but often suboptimal)
- Remember: shown as 2-dimensional here but *n*-dimensional in reality, *n* being the number of variables used



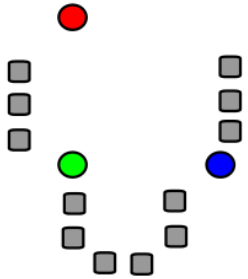
Distance

- Each user is represented by a point in space
- Each item (movie) is a dimension
- The distance between two users can be calculated for *any* number of dimensions/movies (shown: 3)

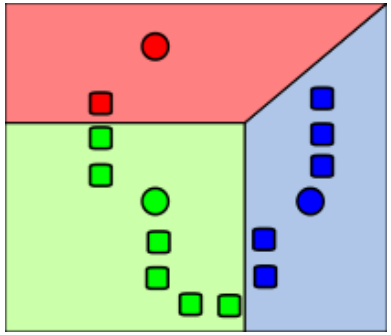
User	Movie 1	Movie 2	Movie 3
1	5	3	5
2	4	1	2



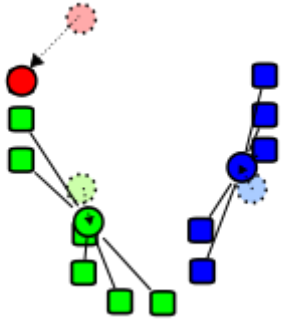
k-means algorithm



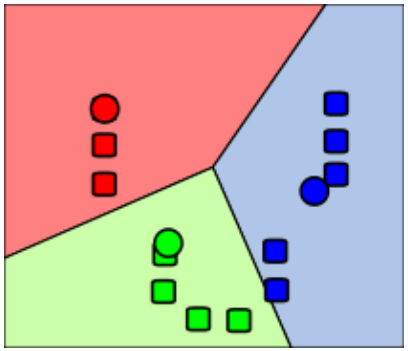
1. Start with *k* cluster centers at random



2. Assign each observation to nearest center

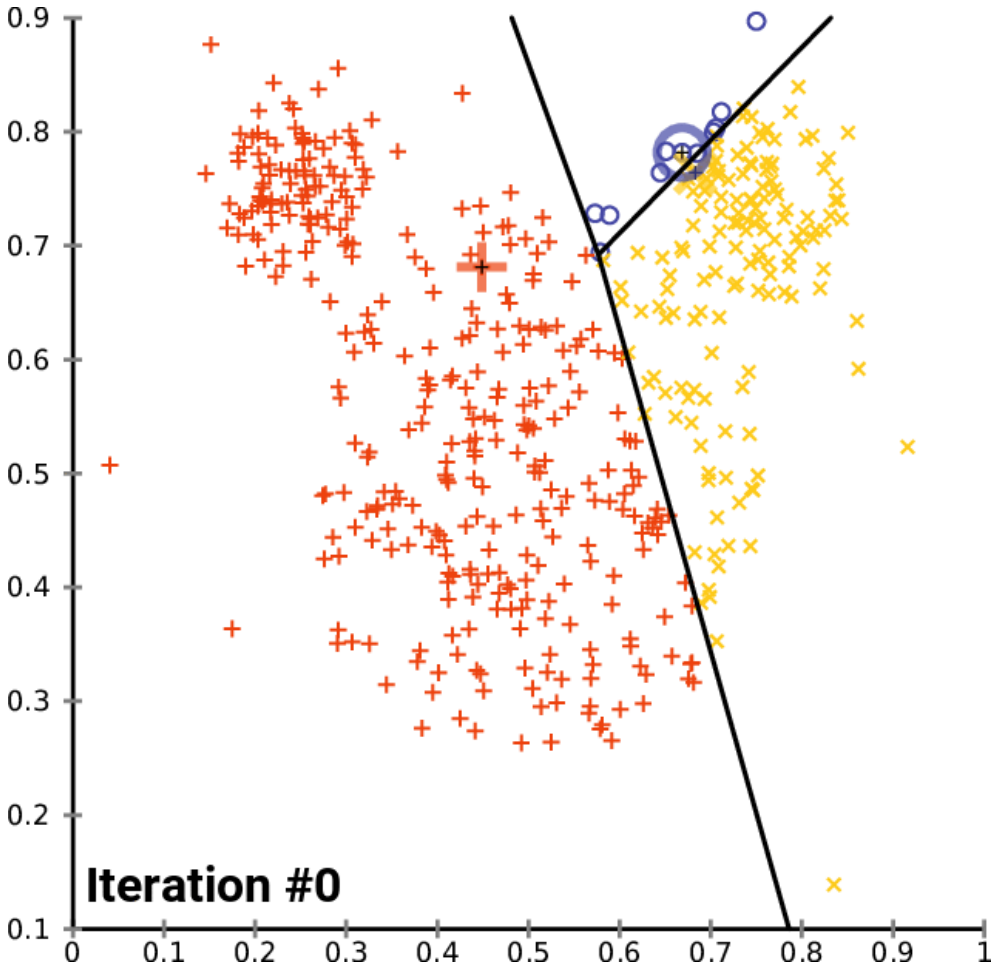


3. Move cluster centers to center of observations



4. Repeat until stable

K-means clustering



Exercise: clustering *Iris* data set (1)

In this exercise, you will explore the logic of clustering using the famous *Iris* data set. See the Exercise folder for the data set (*iris.csv*), and the Notebook *k-means* in the Examples folder.

1. Read in the csv file.

2. Using a Seaborn scatterplot, plot the *Iris* data set with petal length on the x-axis and sepal length on the y-axis. Plot the different Iris species with a different symbol (**see the Seaborn documentation**).

3. What possible clusters do you see with the naked eye? How do they relate to the three species?

4. Using *k*-means clustering with 3 clusters (using **only** petal length and sepal length as X variables), create a new variable *cluster* and use this to make another scatterplot, using color for *cluster*.

5. How does *k*-means perform?

6. Try out different numbers of clusters to see what happens. What seems like the 'natural' number of clusters for this data set?

Exercise: clustering *Iris* data set (2)

In this exercise, you will explore the logic of clustering using the famous *Iris* data set. See the Exercise folder for the data set (*iris.csv*), and the Notebook *k-means* in the Examples folder.

1. Read in the csv file.

2. Using a Seaborn scatterplot, plot the *Iris* data set with petal length on the x-axis and sepal length on the y-axis. Plot the different Iris species with a different symbol (**see the Seaborn documentation**).

3. What possible clusters do you see with the naked eye? How do they relate to the three species?

4. Using *k*-means clustering with 3 clusters (using **only** petal length and sepal length as X variables), create a new variable *cluster* and use this to make another scatterplot, using color for *cluster*.

5. How does *k*-means perform?

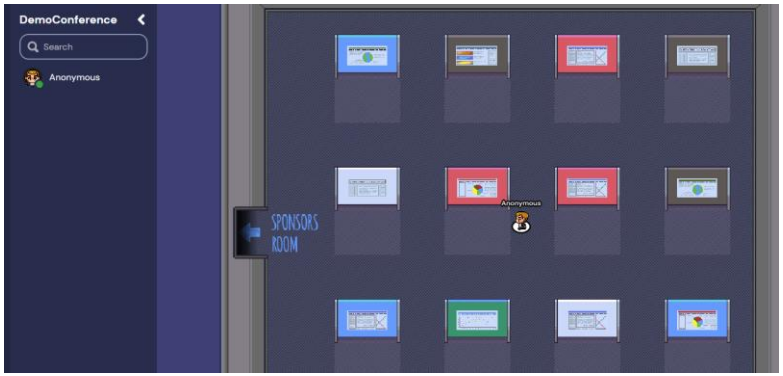
6. Try out different numbers of clusters to see what happens. What seems like the 'natural' number of clusters for this data set?

Problems with *k*-means

- *k*-means can only make Voronoi cells (straight lines)
- Suitable number of clusters often hard to determine in practice

Presentation

- Week 9
  - D02: Wednesday 20th January
  - D01: Friday 22nd January
- You get short feedback from peers and from me via e-mail
- Physical or online → what is your preference? Definitive details end of next week
- **If physical:**
  - Presentations of 5 min. with max. 5 slides
  - In groups of 6-8 (schedule to be sent)
  - Streamed online via Teams (unless you object) for peers, family and friends
  - Possible to participate online
- **If online:**
  - Poster presentation in Gather
  - PDF with simple poster



Format

- **Introduction**: in which you define the context, the research question and the practical relevance
- **Data set**: in which you explain how you acquired the data, and show your data cleaning steps
- **Feature engineering**: in which you explain which transformations you have made to make your variables more informative (e.g., calculating number of days from a starting date)
- **Descriptive analysis**: in which you show *relevant* graphs, tables and numbers with respect to your problem
- **Predictive model**: in which you explain which analysis you have chosen and why. In which you build a relevant statistical model or train a machine learning algorithm.
- **Evaluation**: in which you evaluate the model: numerically, qualitatively and in terms of practical value.

Peer groups

- This week (w7) & next week (w8)
- Discussion and Q&A on final assignment & weekly assignments
- Topic is up to you: present your work, ask questions
- Max. 4 minutes per person
- Maybe you can form a Whatsapp / Teams chat to encourage and help each other?

Image credit

- Pythagoras by CheCheDaWaff (CC-BY-SA)
- K-means algorithm pictures by Weston.pace (CC-BY-SA)
- K-means gif by Chire (CC-BY-SA)