Clustering Methods

A Case Study on Student Profiles in Educational Datasets

Seyed Parsa Neshaei - June 8, 2023

Agenda

- Introduction to student profiles
  - With an example of a sample dataset
- Clustering
  - Why needed?
  - A practical introduction to some clustering methods
- 2-step clustering for extracting student profiles
- Live coding: clustering + neural network using PyTorch
- Challenge: at home!

Attribution

The contents of this talk are mainly based on the findings from the following publications:

- [1] Paola Mejia-Domenzain, Mirko Marras, Christian Giang, and Tanja Käser. *Identifying and Comparing Multi-dimensional Student Profiles Across Flipped Classrooms*. AIED 2022, Durham, UK.

- [2] Paola Mejia-Domenzain, Mirko Marras, Christian Giang, Alberto Cattaneo, and Tanja Käser. *Evolutionary Clustering of Apprentices' Self-Regulated Learning Behavior in Learning Journals*. IEEE Transactions on Learning Technologies, vol. 15, no. 5, October 2022.

- [3] Paola Mejia-Domenzain, Eva Laini, Seyed Parsa Neshaei, Thiemo Wambsganss, and Tanja Käser. *Visualizing Self-Regulated Learner Profiles in Dashboards: Design Insights from Teachers*. AIED 2023, Tokyo, Japan.

- [4] Eva Laini. *Co-Designing a Teacher Tool for Visualizing Self-Regulated Learning Behaviors*. EPFL Master Thesis, 2023.

Student Collected Features

Dimension[a]	Feature	Description
Effort	Total time online [8]	Sum of session durations
	Total video clicks [8]	Video events (play, pause, stop, seek)
Consistency	Mean session duration [8]	Time measured in minutes
	Relative time online	Unit vector of total time online
	Relative video clicks	Unit vector of total video clicks
Regularity	Periodicity of week day [6]	Studying on certain day(s) of the week
	Periodicity of week hour [6]	Studying at certain hours of the day
	Periodicity of day hour [6]	Studying on certain day(s) & hours of the week
Proactivity	Content anticipation [17]	Fraction of videos (from subsequent weeks) watched before the scheduled due date
	Delay in lecture view [6]	Time interval between the first views and the due date of videos of prior weeks
Control	Fraction spent [20]	Real time spent watching the video divided by its duration, averaged across videos
	Pause action frequency [15]	Mean number of pauses divided by the time spent watching a video per video
	Average change rate [20]	Mean playback speed used to watch videos
Assessment	Competency strength [17]	Highest grade achieved by the student on a quiz divided by the number of attempts
	Student shape [17]	Student's tendency of obtaining the maximum grade in a quiz in the first attempt

From [1]

Student Collected Features (cont.)

Profile	%			Dimension					
	LA	FP	PC	*Effort*	*Consistency*	*Regularity*	*Proactivity*	*Control*	*Assessment*
A	24			Lower	Uniform	Lower Peaks	Delayed	Lower	Lower
B	18	28	35	Lower	Uniform	Lower Peaks	Delayed	Higher	Higher
C	19		18	Higher	Uniform	Higher Peaks	Anticipated	Higher	Higher
D	21			Lower	Uniform	Higher Peaks	Delayed	Higher	Higher
E	18			Lower	Uniform	Higher Peaks	Anticipated	Higher	Higher
F		15	27	Higher	Midterm	Higher Peaks	Delayed	Higher	
G		25		Higher	Midterm	Lower Peaks	Anticipated	Higher	
H		14		Lower	Midterm	Lower Peaks	Delayed	Lower	
I		18		Higher	Midterm	Higher Peaks	Anticipated	Higher	
J			20	Lower	Midterm	Lower Peaks	Anticipated	Lower	

From [1]

Student Behavior	Profile A	Profile B	Profile C	Profile D
Proactivity	More up-to-date	More up-to-date	Less up-to-date	Less up-to-date
Effort	Higher intensity	Lower intensity	Higher intensity	Lower intensity
Consistency	Constant work	Work before exams	Constant work	Work before exams
Control	Fast with pauses	Fast with pauses	Slow watchers	Slow watchers
Regularity	Peak before class	Peak before class	Peak before class	No peaks

From [4]

Sample Dataset

Computer-generated, for this talk

dataset

id	time_online	video_clicks	content_anticipation	delay_lectures	fraction_spent	average_grade
0	140	605	21	3	29	18
1	179	523	21	5	92	16
2	247	573	28	6	34	15
3	492	1054	77	1	82	19
4	193	716	29	6	86	17
5	486	986	79	4	92	19
6	270	663	24	5	107	16
7	98	644	28	3	59	16
8	252	680	20	7	104	17
9	279	502	29	5	27	15

Clustering

Why needed?

- We want to group students (unsupervised) into several profiles, such that we can predict the average grade for each student
- Data is given as numbers, but desired as discrete outputs
- We can group the number points close to each other as one label
- Research [3, 4] has shown that clustering student attributes has a positive effect on teachers to spark a range of potential interventions for students and course modifications, if presented in a correct way

Clustering

- Good clustering: high intra-cluster and low inter-cluster similarity
- Clustering is a data segmentation (partition) method
- Should be robust to outliers
- We discuss three methods briefly
  - K-Means
  - K-Modes
  - DBSCAN

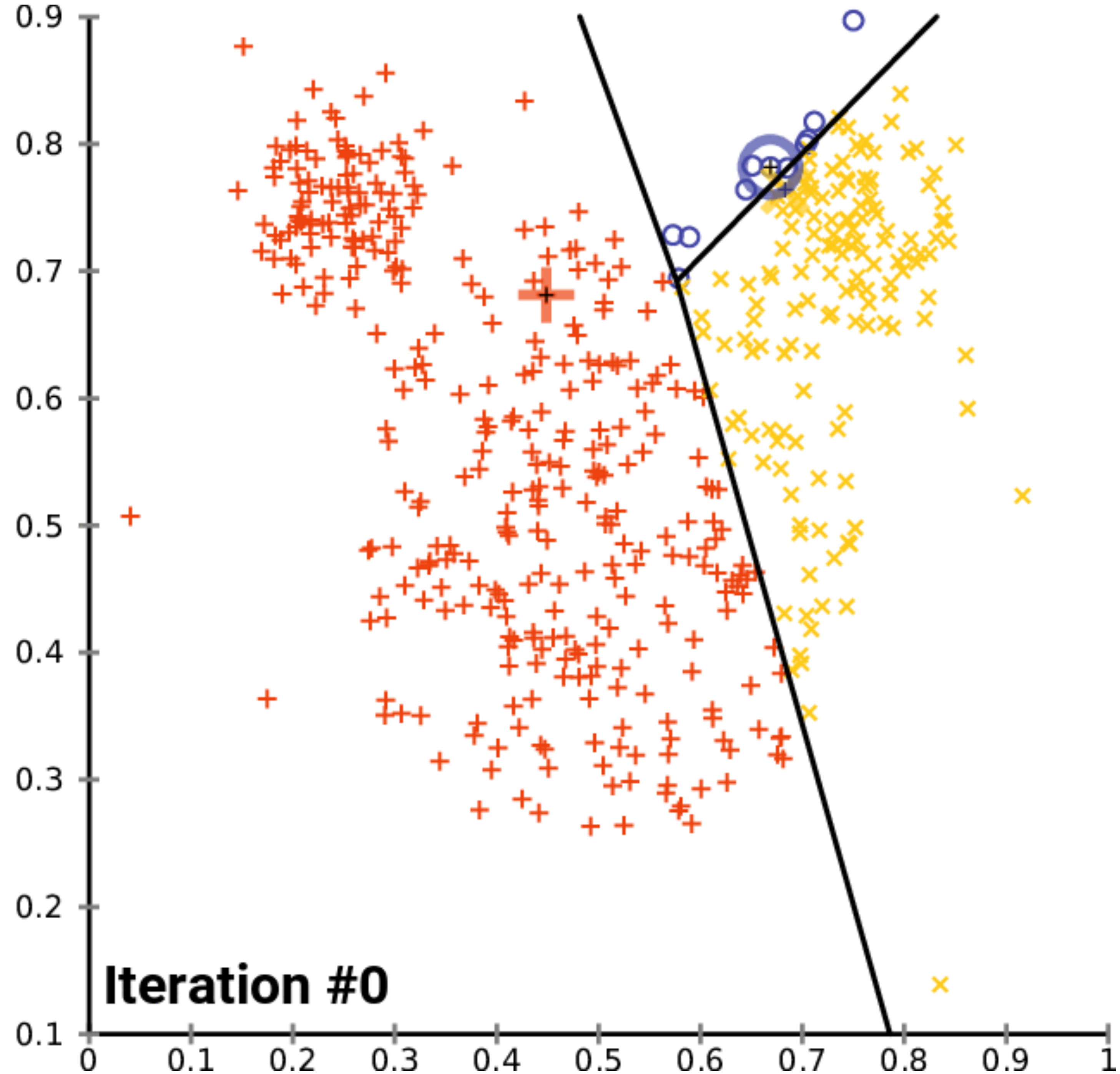
Clustering

K-Means

- Used for continuous data
- Choose k centers randomly
- Find euclidean distance of all data points to all centers
- Assign each data point to a cluster to which it is closer
- Change the cluster center to the mean of all data points of that cluster
- Continue until the centers don’t change further

Clustering

K-Means (cont.)



Source: https://upload.wikimedia.org/wikipedia/commons/e/ea/K-means_convergence.gif

Clustering

K-Modes

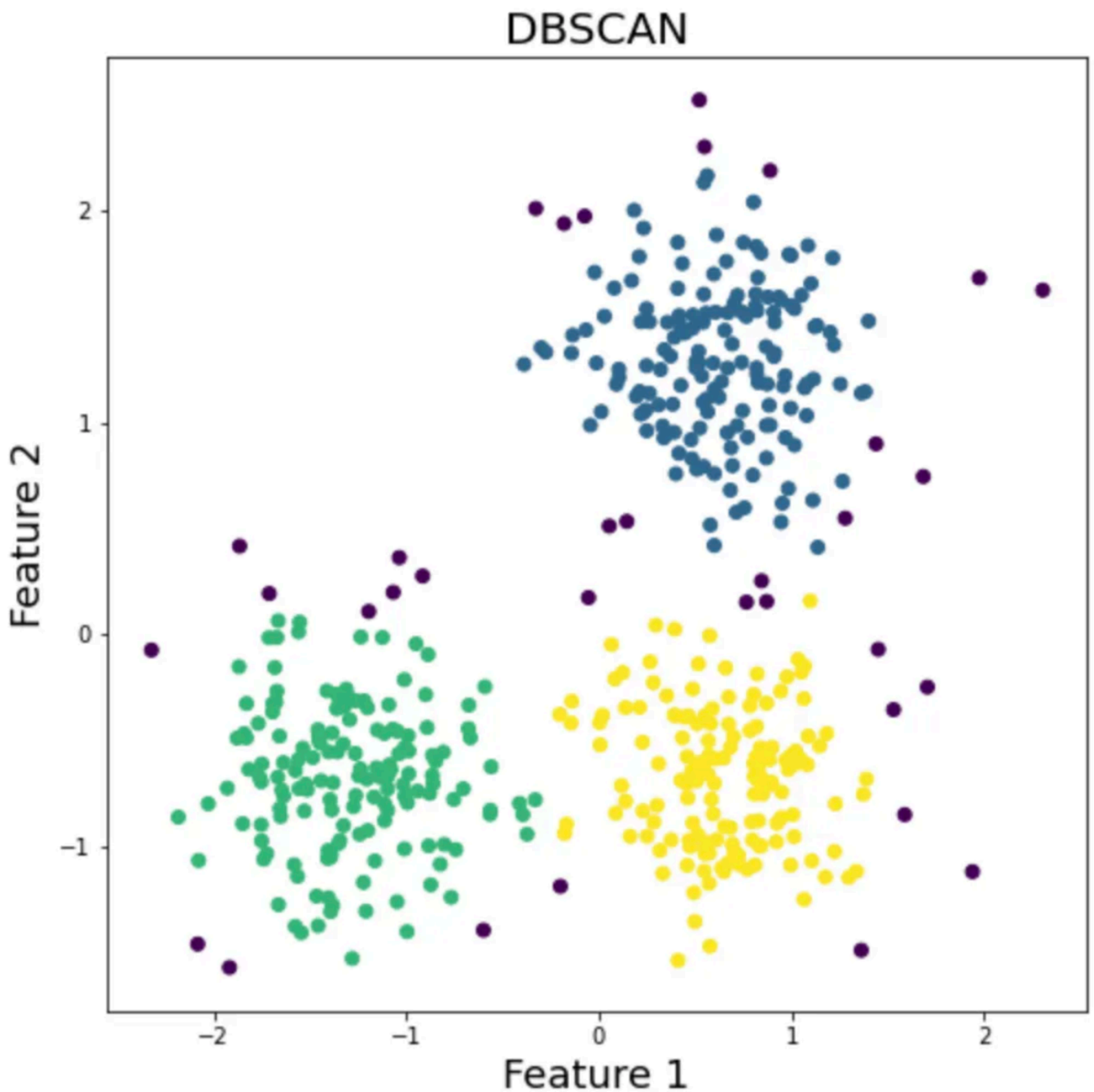
- The same, but calculates the “mode” instead of the mean
- K-modes method is usually applied to categorical data, while K-means method is applied to numerical data

Clustering
DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
- Finds data points that are “dense” and assigns them to a cluster
- Useful for data points with arbitrary shape

Clustering

DBSCAN (cont.)



Source: https://towardsdatascience.com/overview-of-clustering-algorithms-27e979e3724d

Two-step Clustering

dataset

id	time_online	video_clicks	content_anticipation	delay_lectures	fraction_spent	average_grade
0	140	605	21	3	29	18
1	179	523	21	5	92	16
2	247	573	28	6	34	15
3	492	1054	77	1	82	19
4	193	716	29	6	86	17
5	486	986	79			
6	270	663	24			
7	98	644	28			
8	252	680	20			
9	279	502	29			

Profile	%			Dimension					
	LA	FP	PC	*Effort*	*Consistency*	*Regularity*	*Proactivity*	*Control*	*Assessment*
A	24			Lower	Uniform	Lower Peaks	Delayed	Lower	Lower
B	18	28	35	Lower	Uniform	Lower Peaks	Delayed	Higher	Higher
C	19		18	Higher	Uniform	Higher Peaks	Anticipated	Higher	Higher
D	21			Lower	Uniform	Higher Peaks	Delayed	Higher	Higher
E	18			Lower	Uniform	Higher Peaks	Anticipated	Higher	Higher
F		15	27	Higher	Midterm	Higher Peaks	Delayed	Higher	
G		25		Higher	Midterm	Lower Peaks	Anticipated	Higher	
H		14		Lower	Midterm	Lower Peaks	Delayed	Lower	
I		18		Higher	Midterm	Higher Peaks	Anticipated	Higher	
J			20	Lower	Midterm	Lower Peaks	Anticipated	Lower	

From [1]

Two-step Clustering (cont.)

- We present a simplified model of the approach used in [1]
- Two-step clustering of the dataset
  - Finding categorical labels for the dimensions, from the raw numbers
  - Finding student profiles from the categorical labels of the dimensions
- **<u>Ideal outcome:</u>** finding student profiles with different ranges of average grade
  - <u>Conclusion 1:</u> the value of dimensions have correlation with grades
  - <u>Conclusion 2:</u> finding correct labels in an unsupervised manner
  - <u>Conclusion 3:</u> the possibility to predict student grades with ML
    - Predict profile, not the grade directly, for higher accuracy

Live Coding

Challenge: At Home!

- Train two similar ML models to predict the grade from a) the raw values of the dataset, and b) the value of the dimensions (effort, control, etc.) — one of them was trained in the talk!
    - Compare the accuracy of the two models (maybe you can reach 100%!) Do the accuracies differ? (As our sample dataset is computer-generated and not real, there is no guarantee on which model has a higher accuracy)
- Use at least one another different types of model (e.g. a classical model.) for each of the two training approaches, and compare the results by providing plots/graphs