Resit exam questions (2022-07-13 / 2022-07-18)

Resit exam (July 2021)

- This was an oral resit exam
- Students had to answer ~20 questions in ~20-25 minutes
- Questions were chosen at random by throwing a dice, advancing the number shown on the dice, and asking that question, then throwing the dice again, and so on ...

Exam protocol

- Choose language es/ca/en
- We are recording now, the recording will stay in the platform with access only to me, me unless the university authorities request it for some reason
- Please place your mobile in airplane mode (unless you’re using it for communicating with me)
- Please briefly show me the room where you are giving your exam
- Please briefly share with me (“present”) your entire computer screen
- We will start with a topic you think you’ve studied more, then we will go back to slide #3 and roll the dice to determine each question; if we land on a question you’ve already answered or a non-question slide, I ask you the next one; if we get to the end we restart
- I’ll ask you questions for 20 minutes starting now – pick the initial topic please

TT02 Data, methods, scenarios

Is ordinal a type of categorical data?

Why or why not?

TT02 Data, methods, scenarios

If you use one-hot encoding for “Source” and remove the original column, how many columns will the resulting table have?

Dataset	Features	Classes	Samples	Source
Gesture Phase	32	5	9.8k	OpenML
Gas Concentrations	129	6	13.9k	OpenML
Eye Movements	26	3	10.9k	OpenML
Epsilon	2000	2	500k	PASCAL Challenge 2008
YearPrediction	90	1	515k	Million Song Dataset
Microsoft (MSLR)	136	5	964k	MSLR-WEB10K
Rossmann Store Sales	10	1	1018K	Kaggle
Forest Cover Type	54	7	580k	Kaggle
Higgs Boson	30	2	800k	Kaggle
Shrutime	11	2	10k	Kaggle
Blastchar	20	2	7k	Kaggle

Table 1: Description of the tabular datasets

TT03. Data preparation: data types

How would you convert a **numerical** variable into an **ordinal** variable?

TT03. Data preparation: data types

Suppose data are:
-10, -6, -1, 4, 20, 30

Divide into three equi-width bins
(none of the bins can be empty).

Divide into three equi-depth bins.

TT04. Data prep.: integration & cleaning

What is a **set membership constraint**?

Give an example

TT04. Data prep.: integration & cleaning

Name two reasons why we perform data cleaning

TT04. Data prep.: integration & cleaning

Give an example of **cross-field validation**

TT04. Data prep.: integration & cleaning

Suppose in a sample of people from Catalonia the city is missing for 10% of the people

How would you decide whether to impute city=Barcelona or drop the rows without city?

TT04. Data prep.: integration & cleaning

What is **standardization**?

TT04. Data prep.: integration & cleaning

We have a variable taking values {-3, 0, 3}

Normalize by using min-max scaling

TT04. Data prep.: integration & cleaning

Describe clearly how to perform seasonal standardization

TT05. Reduction and transformation

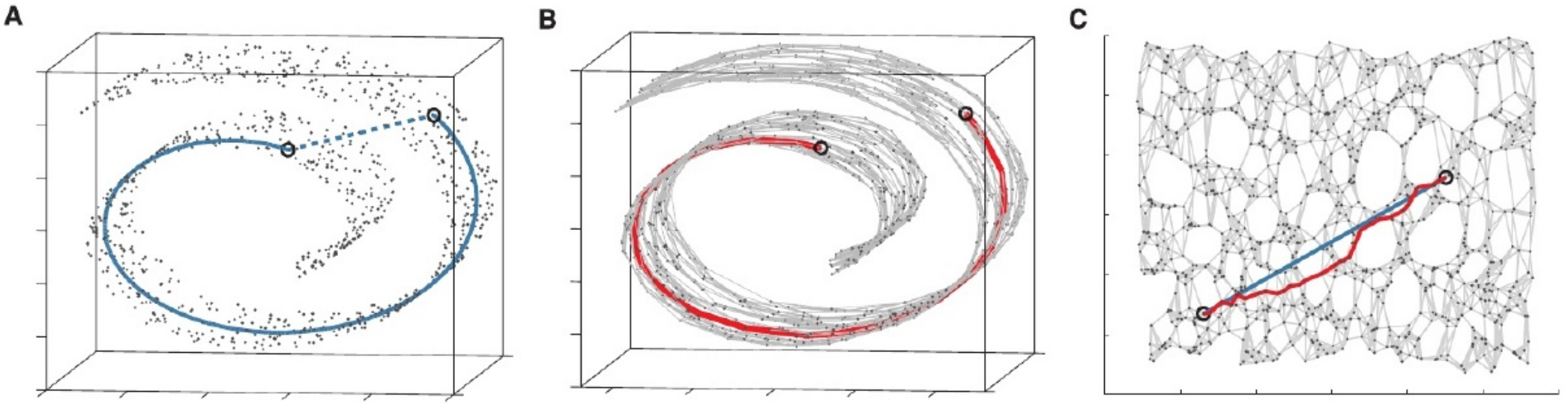
Suppose a population contains 10 people from Barcelona and 5 from Madrid. Describe how to sample 4 elements using **stratified sampling by city**

TT06. Similarity on numerical data

Explain what is **the curse of dimensionality**

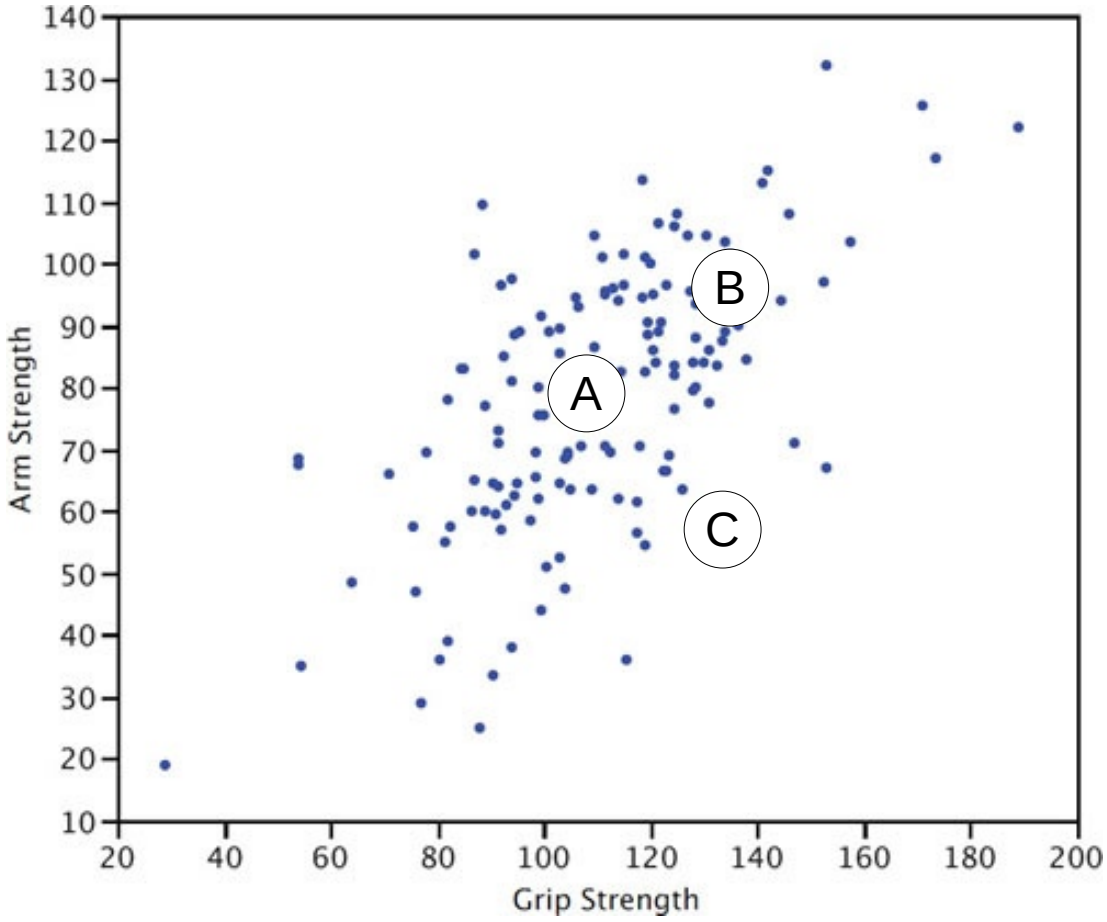
TT06. Similarity on numerical data

Explain how ISOMAP works



TT06. Similarity on numerical data

The Mahalanobis distance would consider that d(A,B) < d(A,C) or d(A,B) > d(A,C) ?



TT07. Similarity: beyond numerical

Write and explain the formula for the Goodall measure

TT07. Similarity: beyond numerical

Compute the **Jaccard distance** between these two sets:

{apple, banana}

{tomato, banana}

TT08. Near duplicates

Explain the principle behind min hashing, and give an example

TT08. Near duplicates

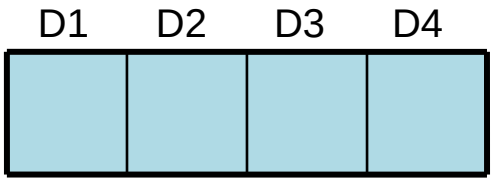
How many **different** 2-word-gram shingles are contained in the phrase “to be or not to be”?

TT08. Near duplicates

Permutation π **Rows=Shingles, Columns=Documents**

	D1	D2	D3	D4
2	0	0	0	1
3	1	0	1	1
6	1	1	0	1
4	0	1	0	1
1	0	0	0	1
5	1	0	0	0

Compute the signature vector under π



TT08. Near duplicates

What is the similarity between each pair of documents, in this signature matrix?

	D1	D2	D3	D4
π_1	1	1	2	2
π_2	2	3	3	2
π_3	5	5	4	5

TT11. Itemsets

Give 3 examples of transactions
from 3 different application domains

TT11. Itemsets

What is the maximum possible support of an itemset that exists in a database?

TT11. Itemsets

Indicate the support of the itemset “Scissor, Eraser”

tid	Set of items
1	Pencil, Eraser, Paper
2	Scissors, Eraser
3	Pencil, Scissors
4	Eraser, Paper, Scissors
5	Pencil, Highlighter, Eraser

TT11. Itemsets

Explain the support monotonicity property using as example the itemset {“Eraser”}

tid	Set of items
1	Pencil, Eraser, Paper
2	Scissors, Eraser
3	Pencil, Scissors
4	Eraser, Paper, Scissors
5	Pencil, Highlighter, Eraser

TT11. Itemsets

What is a closed itemset?

TT11. Itemsets

What is a maximal itemset?

TT11. Itemsets

What is a closed itemset in this database?

What is a non closed itemset in this database?

tid	Set of items
1	Pencil, Eraser, Paper
2	Scissors, Eraser
3	Pencil, Scissors
4	Highlighter, Paper, Scissors
5	Pencil, Highlighter, Eraser

TT11. Itemsets

Draw an itemset lattice for the following transactions:

TID	Items
1	Tomato, Pear
2	Strawberry, Pear, Apple
3	Apple, Pear
4	Apple, Strawberry, Tomato, Pear

TT12. Association rules

What does it mean if the **lift** of a rule is smaller than 1.0?

TT13. Association rule mining

Explain the apriori algorithm on this dataset, with minsup=0.5
Tip: first write a table with itemsets of size 1 (itemset, support)

tid	Set of items
1	x1 x2 x3
2	x2 x3 x4
3	x4 x5
4	x1 x2 x4
5	x1 x2 x3 x5
6	x1 x2 x3 x4

TT13. Association rule mining

Explain the confidence monotonicity property

Let X_S, X_L, I be itemsets; assume $X_S \subset X_L \subset I$

Then:

$$\text{conf}(X_L \Rightarrow I - X_L) \geq \text{conf}(X_S \Rightarrow I - X_S)$$

TT16. Recommender systems

What is the difference between explicit and implicit preferences in an utility matrix? Give some examples.

TT17. Recommender systems

What does it mean cold-start in a recommender system?

TT17. Recommender systems

$$\mathrm{sim}(u,v) = \frac{\sum_{i \in I_{u,v}} (u_i - \hat{u}) \cdot (v_i - \hat{v})}{\sqrt{\sum_{i \in I_{u,v}} (u_i - \hat{u})^2 \cdot \sum_{i \in I_{u,v}} (v_i - \hat{v})^2}}$$

$$\mathrm{score}(u,i) = \hat{u} + \frac{\sum_{v: v_i \neq \mathrm{NULL}} \mathrm{sim}(v,u) \cdot (v_i - \hat{v})}{\sum_{v: I_{u,v} \neq \emptyset} |\mathrm{sim}(v,u)|}$$

Consider the utility matrix of 3 users (u1, u2, u3) on 4 movies (m1, m2, m3, m4) as shown on the next table.

	m1	m2	m3	m4
u1	1	1	1	-2
u2	-1		-1	2
u3	2	-1		

What are the values of $\widehat{u1}$, $\widehat{u2}$, $\widehat{u3}$*?*

$\widehat{u1} =$ $\widehat{u2} =$ $\widehat{u3} =$

What is $I_{u1,u3}$*? What is* $sim(u1, u3)$*?*

$I_{u1,u3} = \{ \quad \}$ $\mathrm{sim}(u1, u3) =$

TT18. Factorization-based recsys

What is the objective function in an factorization-based recommender system?

TT19. Outlier detection

Use the method based on z-scoring

	a1	a2	z-score of a1	z-score of a2
1	1	100		
2	2	50		
3	0	150		
4	1	30		
μ				
σ				

$\mu = (\sum x_i)/N$ and $\sigma = \sqrt{\sum (x_i - \mu)^2 / N}$

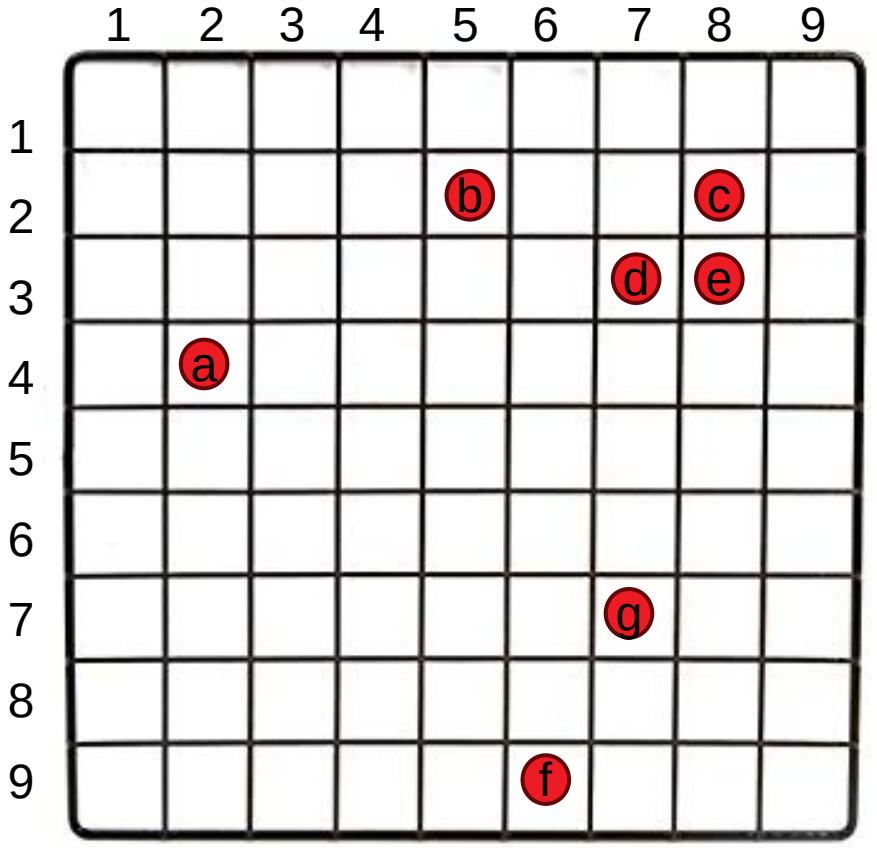
Complete μ, σ, and the z-scores of a1 and a2 in the above table, with 4 decimal digits

Which item is the outlier, and because of which attribute?

TT21. Outlier detection

Indicate how do you create an isolation forest over the graph on the right

Explain what the outlier score for a point depends on (no need to give a formula)



TT22. Streams

Suppose we have a stream of the type *(a, b)* indicating that user *a* purchased book *b*

Indicate how to sample 1% of the users and the books they have bought from this stream

TT22. Streams

Suppose we have a stream of photos from a photo sharing site

Indicate how to sample 10 photos from this stream **uniformly at random**

TT23. Streams

Explain how reservoir sampling works

TT24. Bloom filters

In a bloom filter, why would we sometimes want to increase the number of bits of the filter?

TT24. Bloom filters

When a bloom filter says an object **is <u>not</u> a member** of a set, it is:
(a) always right (b) sometimes right

When a bloom filter says an object **is a member** of a set, it is:
(a) always right (b) sometimes right

TT27. Time series

Interpolate the following time series using **linear interpolation** to obtain the values on Monday at midnight and Tuesday at noon

Monday 12:00 – 36ºC

Monday 23:59 – ???

Tuesday 06:00 – 30ºC

Tuesday 12:00 – ???

Tuesday 18:00 – 35ºC

TT27. Time series

Compute a moving average with k=3 in the following series:

t	1	2	3	4	5
y_t	1	5	67	10	23

TT27. Time series

Consider series
X = 1, 1, 2, 4, 3
Y = 1, 3, 3, 2

Draw a table like this on paper and calculate the alignment path

	X1=1	X2=1	X3=2	X4=4	X5=3
Y1=1					
Y2=3					
Y3=3					
Y4=2					