UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

Dr Xuefei Lu
The University of Edinburgh Business School

Quiz

In which of the of following instances does the Euclidean distance suffers from the curse of dimensionality?

- [✓ A. When there is a high number of dimensions.]
- B. When there is a low number of dimensions.
- C. It is unrelated to the number of dimensions, but is related with the size of the dataset.
- D. It does not suffer from the curse of dimensionality.

Quiz

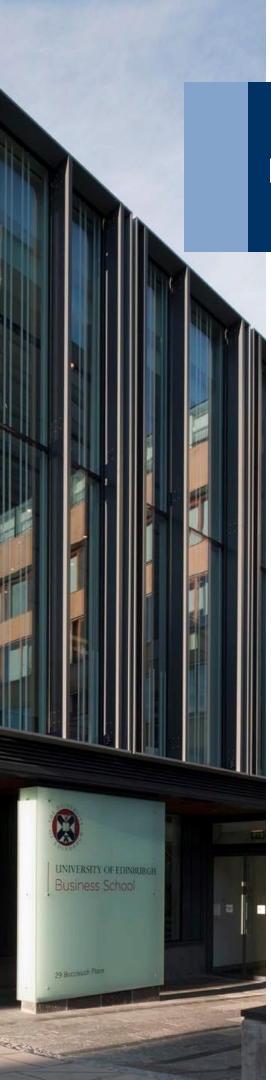
A bad initialisation of K-means can lead to:

- A. More clusters.
- B. Fewer clusters.
- [✓ C. Longer running times.]
- D. All of the above.

Quiz

In what case would you convert a numeric variable to a different type?

- [✓ A. To convert a regression into a classification.]
- B. To convert a regression into a time series analysis.
- C. To convert a classification into a regression.
- D. None of the above



Quiz

When is standardisation preferred over min-max scaling?

- A. When there are a lot of negative values in the dataset.
- B. When there are no negative values in the dataset.
- [✓ C. When outliers are present.]
- D. When it performs similarly to min-max scaling.

Quiz

Select the best TWO options for completing the following sentence:

“Normalisation is required because...”

- [✓ A. Predictive algorithms require particular input ranges.]
- B. It can transform continuous variables into a normal distribution.
- C. It can transform categorical variables into continuous ones.
- [✓ D. Variables have very different scales.]

Quiz (optional)

Select the best TWO options for completing the following sentence:

“When converting the dependent categorical value of a dataset,…”

- A. We can obtain regression
- B. We should use dummies
- [✓ C. We should use label encoding]
- [✓ D. We can do time series analysis]

Quiz

A company wants to perform the analysis of its yearly sales data to predict which customers will leave the company.

What approach should the company use?

- A. Time series analysis
- [✓ B. Classification]
- C. Regression
- D. All of them can be used

Quiz

Which of the following techniques is/are normalisation used for?

- A. Regression
- B. Classification
- C. Pre-processing
- [✓ D. All of them above]

Quiz

Which of the following parts of the data mining process can clustering be used for?

- [✓ A. Data selection]
- [✓ B. Data pre-processing]
- C. Data transformation
- D. All of them above

Quiz

When can predictive modelling achieve the best results?

- A. When the data is normalised.
- B. When the model is performing very well.
- C. When the mean and median of all variables are close together.
- D. When the mean and median of all variables are close together.
- [✓ E. None of them above]

Homework

Based on '13 - Assessment_can_you_spot_the_fraudsters.ipynb' and dataset 'credit.csv', answer the following question:

Q1. Which of the two variables did you remove from the dataset?

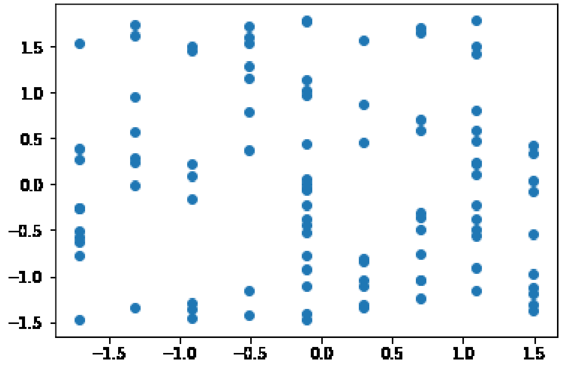
CC, NO, CITY, PHONE

Homework

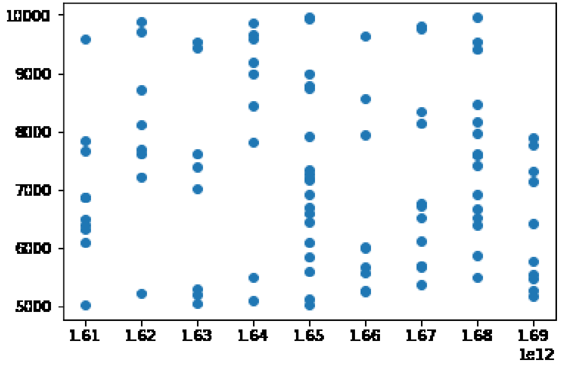
Based on '13 - Assessment_can_you_spot_the_fraudsters.ipynb' and dataset 'credit.csv', answer the following question:

Q2. Which graph contains the standardised representation of both No and Money?

Graph A



Graph B

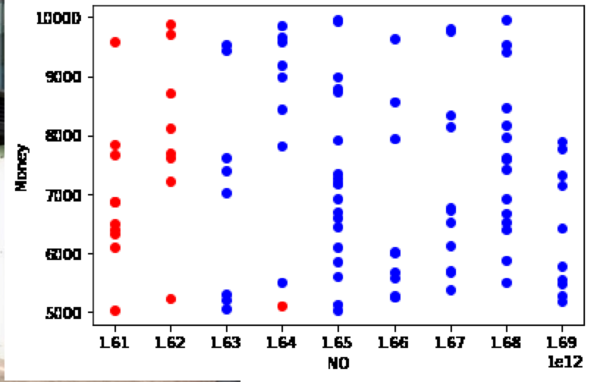


Homework

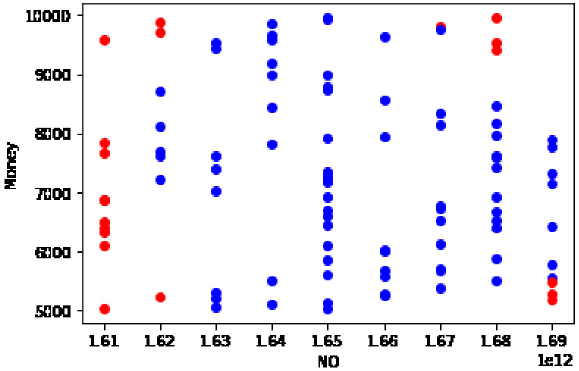
Based on '13 - Assessment_can_you_spot_the_fraudsters.ipynb' and dataset 'credit.csv', answer the following questions:

Q3. Which graph represents the result of LOF with 2 neighbours, 20 neighbours, and the elliptic envelope procedure, all at a contamination rate of 20%? The outliers are indicated in red.

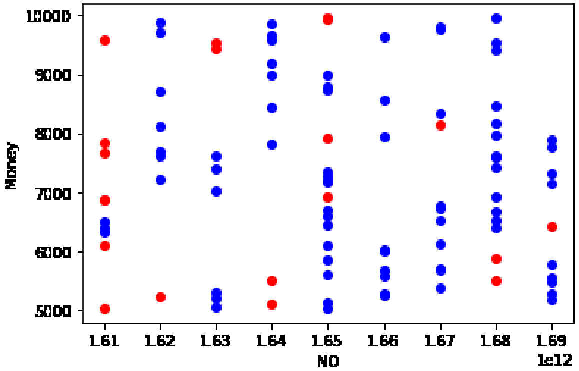
Graph A = LOF(20)



Graph B = EE



Graph C = LOF(2)


Money
NO
1e12