COMP20008 Elements of Data Processing

ORAL PRESENTATION

Group members:

- Ngoc Thanh Van Tran ID: 980818
- Edward Marozzi ID: 910193

Outline and corresponding questions:

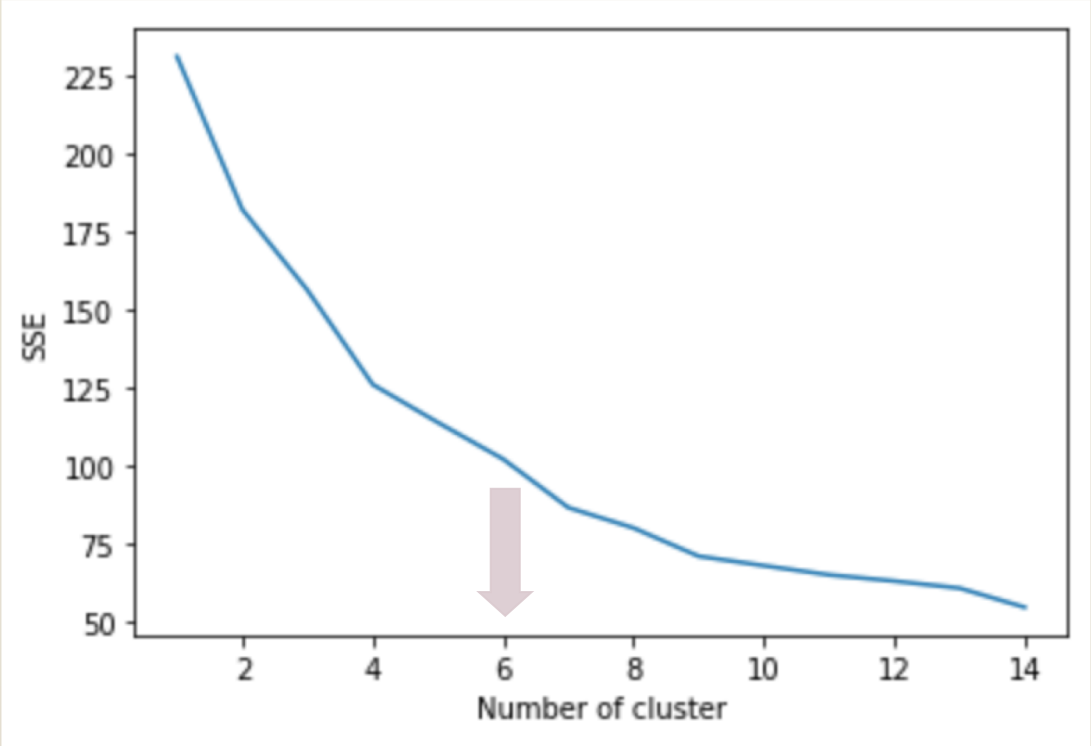
- Pre-processing tasks — 1
- Designing classifier & Methodology — 2 3 4
- Performance & Limitations — 5 6
- Improvement & Alternatives — 7

1 Pre-processing tasks:

- 1) Impute missing values
  - Median
  - Mean
- 2) Scale features
  - Mean centering
  - Standardisation

2 Different classification algorithms:

Algorithm	k-NN		Decision tree
	k = 5	k = 10	
Accuracy on the test data *(based on the first 8 features)*	66.14%	69.06%	63.68%



3 4 Classifier methodology and parameters manipulation:

Feature Engineering

- **Interaction term pairs**
  (New features are products of each pair of initial features)
- **Clustering labels**
  (Applying k-means clustering to create an extra cluster label feature)

k-NN classification

5 Performance:

Both gives accuracy of 70.71% with:

- Top number of features to select: 14
- Number of nearest neighbours: 2

6 Generalization & Limitation:

Highly generalised: Our classifier will always improves or maintains accuracy as it takes a brute force approach to the problem by testing all possible combinations of k in k-NN and n in top-n neighbours. Then the classifier outputs the two parameters that produced the highest accuracy.

Limitation: Performance is the biggest issue, by testing all possible combinations the classifier is computationally complex.

Currently sensitive to random state changes.

7 Improvement & Alternatives:

Performance is the main area that our classifier needs improvement on. Faster performance would mean we could iterate over a few random states to find the mode of our k and n parameters.

This would reduce the sensitivity to different random states and find a more accurate model of the data overall.

To improve performance we likely would have to re-write the k-NN classifier built into sk-learn and tailor it to our data for faster performance.

Q&A