DS 5230

Unsupervised Machine Learning and Data Mining

Project Description

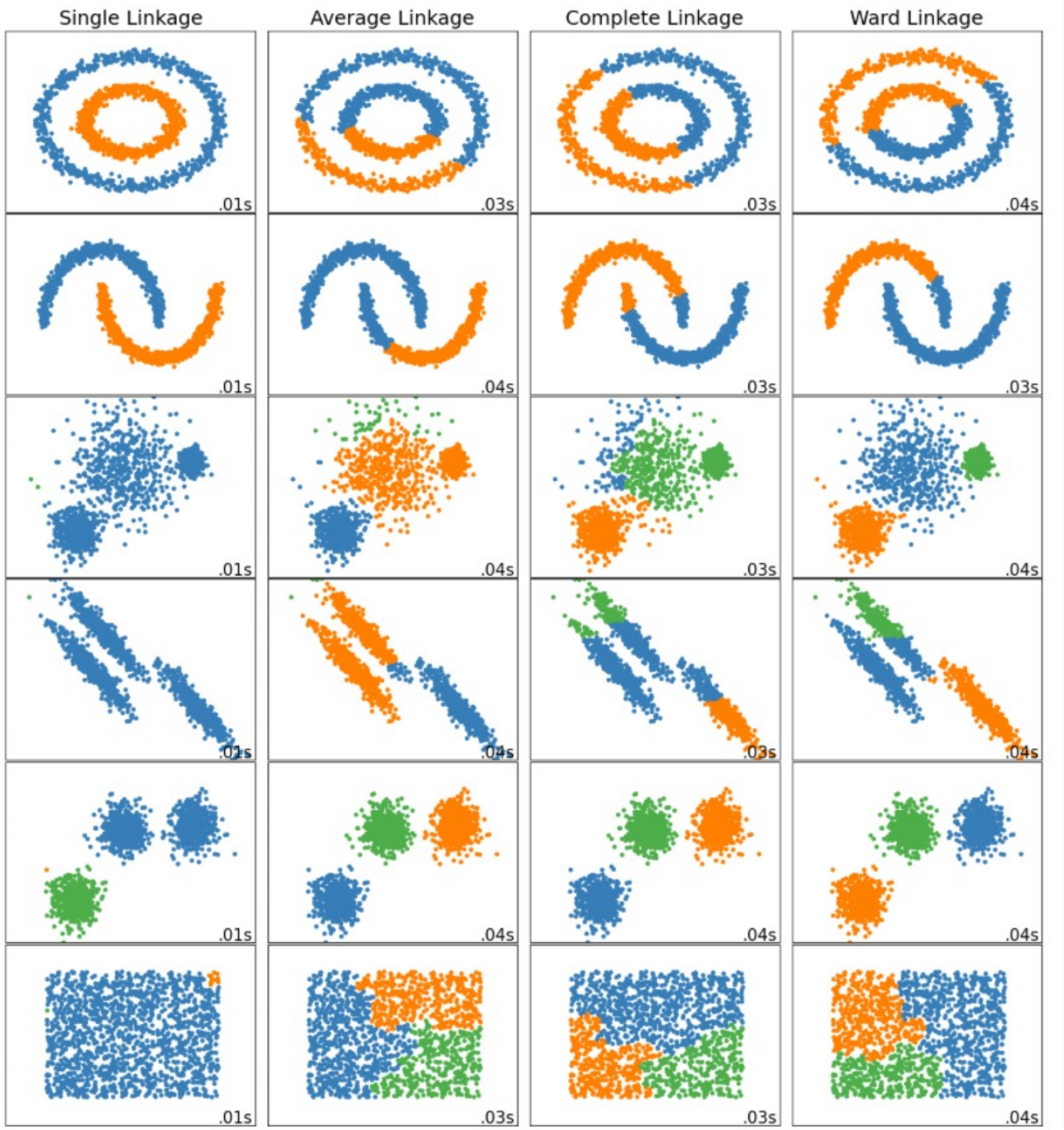
Steve Morin

Goal and Description

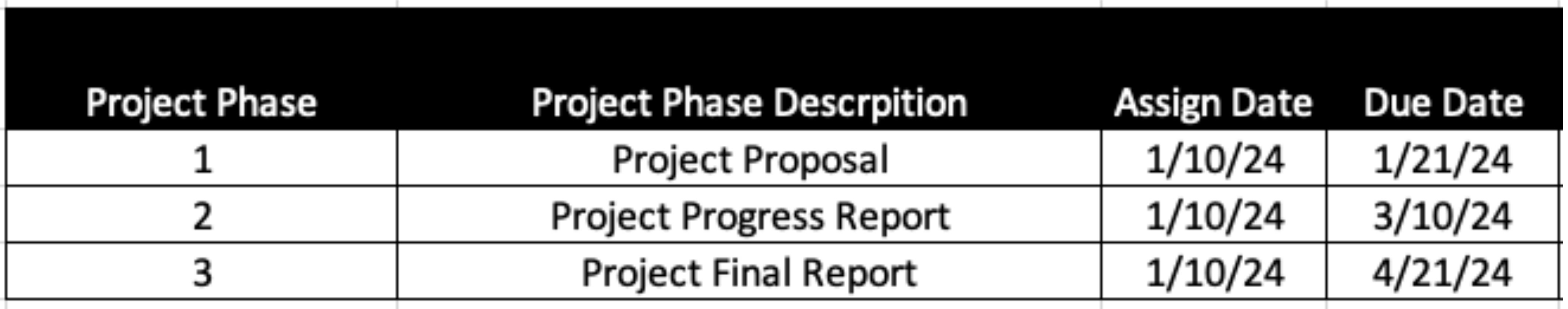
The goal of this project is to develop an appreciation of the challenges and pitfalls of clustering.

Clustering, like most other machine learning algorithms, will dutifully return an answer without regard for its usefulness or meaningfulness.

This feature of machine learning algorithms is especially problematic in unsupervised learning due to the lack of a target attribute to supervise the training and evaluate the resulting model.



Project Phases and Due Dates



Project Phase	Project Phase Descrpition	Assign Date	Due Date
1	Project Proposal	1/10/24	1/21/24
2	Project Progress Report	1/10/24	3/10/24
3	Project Final Report	1/10/24	4/21/24

A description of each project phase is provided below.

Teams/groups are allowed.

There is a maximum of 3 people per team/group.

Summary of Deliverables

Do not combine required documents.

Phase	Phase Name	Deliverables	Notes
1	Data Set Selection and Preparation	• A PowerPoint document submitted as a .pdf file with the string ‘phase_1’ in the fie name. • A .ipynb and .html that demonstrates the data set preparation outlined on the previous slide. • A .yml file documenting the python environment.	See phase 1 discussion for required content of deliverables.
2	Preprocessing Pipelines and EDA	• A PowerPoint document submitted as a .pdf file with the string ‘phase_2’ in the fie name. • A .ipynb and .html that demonstrates the data set transformations with the string ‘transformation’ in the fie name. • A .ipynb and .html that demonstrates the data exploration with the string ‘eda’ in the file name. • A .yml file documenting the python environment.	See phase 2 discussion for required content of deliverables.

Notes:

1. Include any .py files that are imported by the .ipynb files for all submissions.

Summary of Deliverables (continued)

Do not combine required documents.

Phase	Phase Name	Deliverables	Notes
3	Dimensionality Reduction, Clustering and Evaluation	• A PowerPoint document submitted as a .pdf file with the string ‘phase_3’ in the fie name. • A .ipynb and .html that demonstrates a working clustering pipeline. The string ‘clustering_pipeline’ should be in the file name. • A .ipynb and .html that demonstrates the optimization of the clustering pipeline. The string ‘optimization’ should be in the file name. • A .ipynb and .html that demonstrates checking for false discoveries. The string ‘afd’ should be in the file name. • A .ipynb and .html that demonstrates the validation of the clustering pipeline using external indices. The string ‘validation’ should be in the file name. • A .yml file documenting the python environment.	See phase 3 discussion for required content of deliverables.

Notes:
1. Include any .py files that are imported by the .ipynb files for all submissions.

Project Phase 1
Data Set Selection and Preparation

Phase 1 Outline

Select a multi-class classification data set for the project.

Many of the methods we will work with in unsupervised learning rely on attributes having an interval or ratio measurement scales.

As a result, nominal and ordinal attributes are not supported.

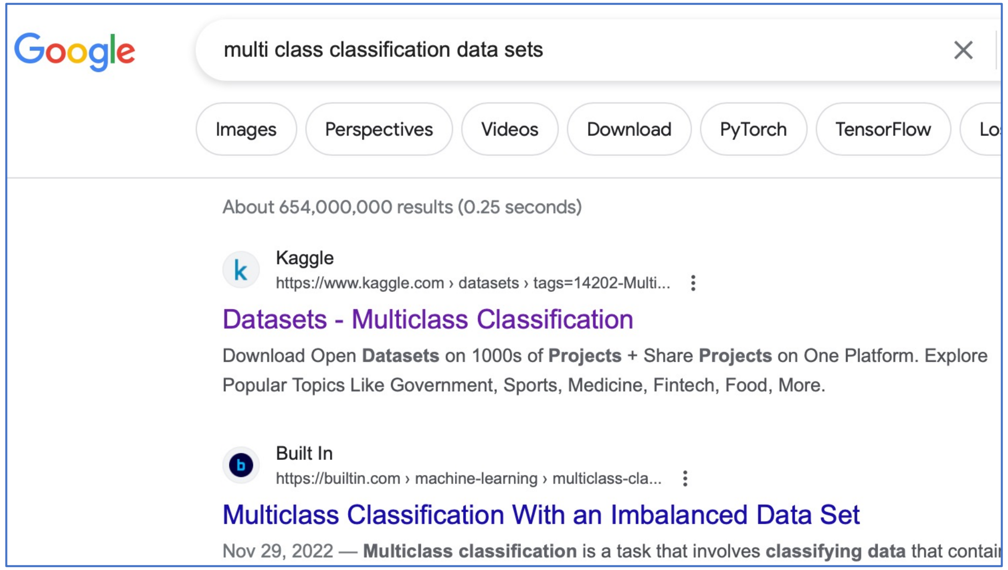
Select a data set with attributes that are predominately interval and/or ratio measurement scales.

The number of classes in the target vector must be ≥ 2 and ≤ 10.

Do not select the MNIST Digits data set as we will be working with that data set in class.

The results of a Google search on the right might be a place to start your search.

Other machine learning data repositories are listed on the next slide.



multi class classification data sets

About 654,000,000 results (0.25 seconds)

Kaggle
https://www.kaggle.com › datasets › tags=14202-Multi...

Datasets - Multiclass Classification

Download Open **Datasets** on 1000s of **Projects** + Share **Projects** on One Platform. Explore Popular Topics Like Government, Sports, Medicine, Fintech, Food, More.

Built In
https://builtin.com › machine-learning › multiclass-cla...

Multiclass Classification With an Imbalanced Data Set

Nov 29, 2022 — **Multiclass classification** is a task that involves **classifying data** that contai

Phase 1 Outline (continued)

Popular data repositories:

- OpenML.org (https://openml.org)
- Kaggle.com (https://www.kaggle.com/datasets)
- PaperWithCode (https://paperswithcode.com/datasets)
- UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/datasets)
- Amazon’s AWS datasets (https://registry.opendata.aws)
- TensorFlow datasets (https://www.tensorflow.org/datasets)

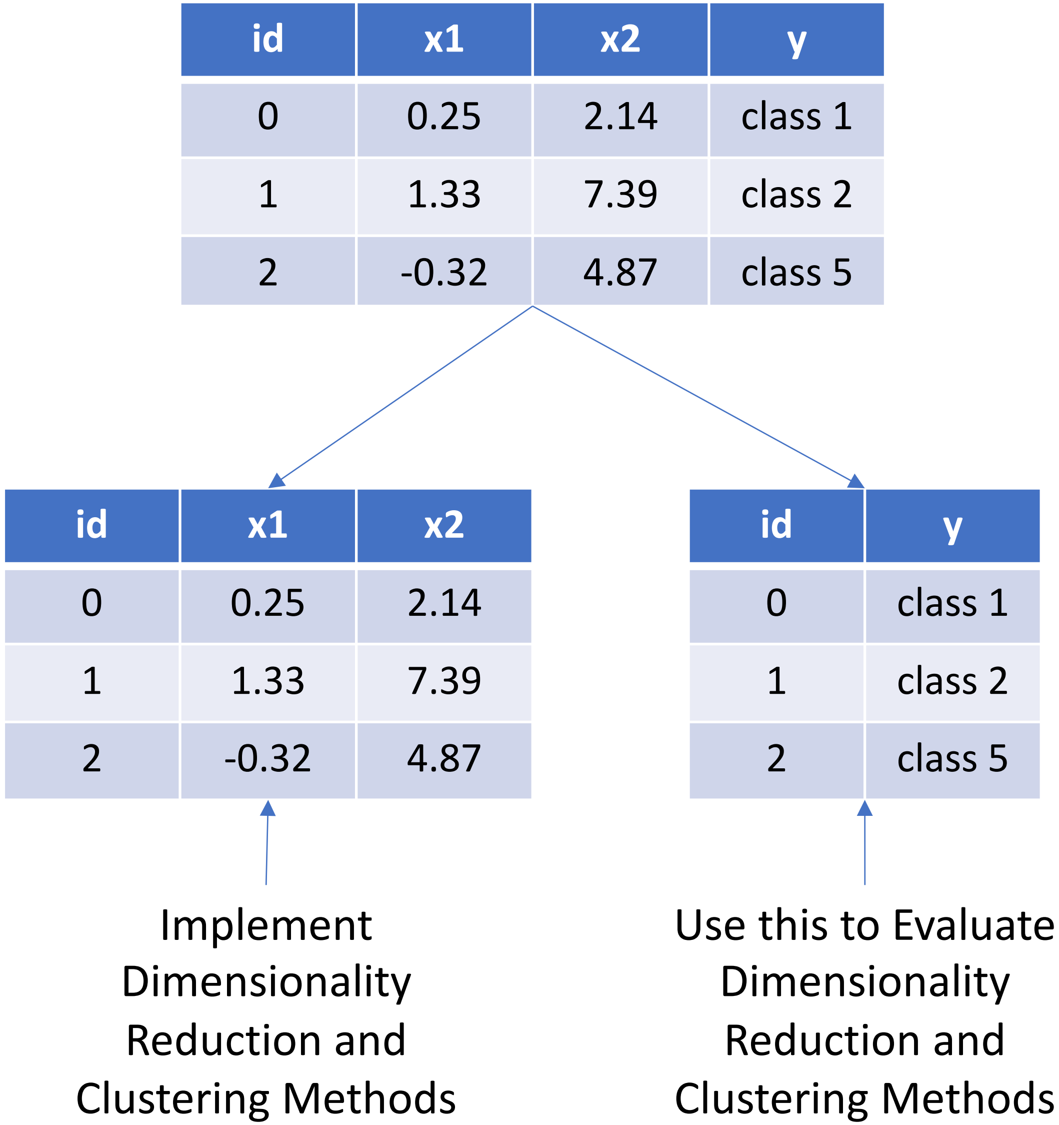
Phase 1 Outline (continued)

In this project you will select a multi-class classification data set.

The target vector will be dropped from the data set and saved to the .csv file for later use.

You will implement dimensionality reduction and clustering methods on the data set.

Once clustering is complete you will use the reserved target vector to evaluate the clustering.



id	x1	x2	y
0	0.25	2.14	class 1
1	1.33	7.39	class 2
2	-0.32	4.87	class 5

id	x1	x2
0	0.25	2.14
1	1.33	7.39
2	-0.32	4.87

id	y
0	class 1
1	class 2
2	class 5

Implement Dimensionality Reduction and Clustering Methods

Use this to Evaluate Dimensionality Reduction and Clustering Methods

Phase 1 Outline - Lecture 1 Lab

Once you have selected and downloaded a data set complete the following.

1. Drop data objects that have a missing value in the target attribute.
2. If an ID attribute is not present in the data set add one.
3. Copy the ID attribute and the target vector attribute into a second target data frame.
4. If the target vector is a string vector encode it to a numerical vector. Be sure to document the mapping in the jupyter notebook.
5. Save the second data frame in a .csv named the xxx_target.csv where xxx is the name of the data set.
6. Drop the target vector attribute from the original data frame and save the resulting data frame to a .csv named xxx_design.csv where xxx is the name of the data set.

An ID attribute is an attribute with the number of unique values equal to the number of rows in the data frame.

An ID attribute allows us to uniquely identify data objects in a data frame.

Phase 1 Deliverables

A PowerPoint document submitted as a .pdf file that addresses the following points:

1. If team or group – list members.
2. Data set name.
3. Data set description.
4. Internet link to the data set.
5. Data set size (rows by columns).
6. Attribute meanings (what is being measured).
7. Attribute types (dtypes).
8. Attribute measurement scale.
9. Attribute missingness (numerical and visual - missingno).
10. The number of classes in the target vector.

A .ipynb and .html that demonstrates the data set preparation outlined on the previous slide.

Include any .py files that are imported by the .ipynb.

A .yml file documenting the python environment.

Deliverables
.pdf
.ipynb (plus any imported .py files)
.html
.yml

Project Phase 2
Preprocessing Pipelines and EDA

Phase 2 Outline - Lecture 2 Lab

I. Prepare the data to better expose the underlying data patterns to machine learning algorithms.

   A. Identify attributes with missing value count greater than a predefined threshold. I typically use 20% as a threshold.

   B. Identify non-machine learning attributes.

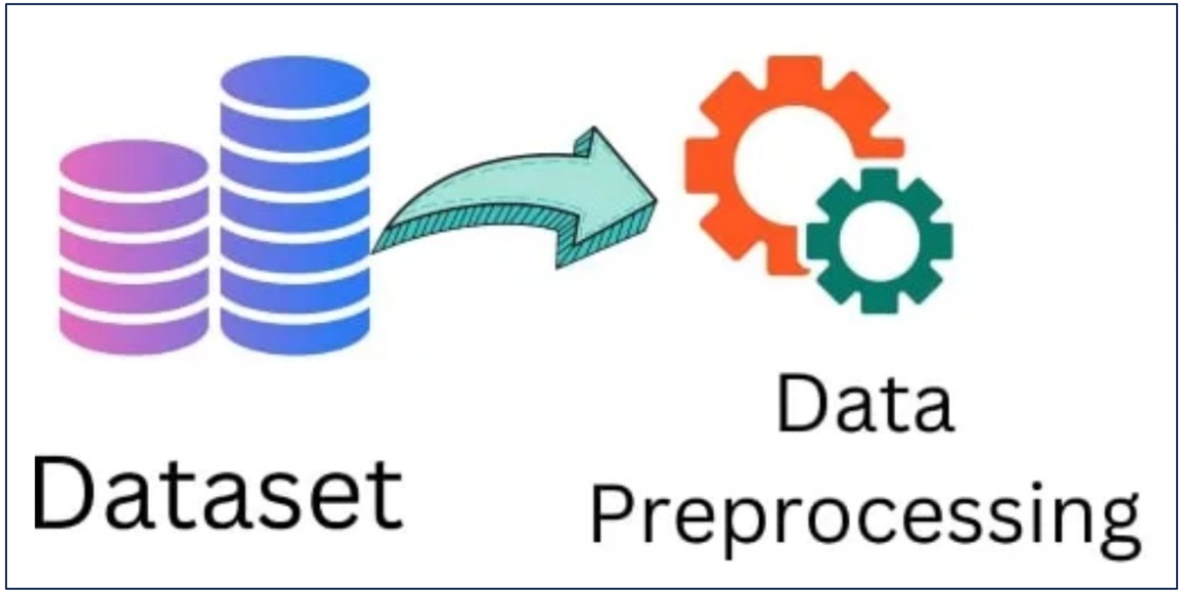
   C. Identify attributes to drop from machine learning.

   D. Establish numerical and nominal attributes.

   E. Establish machine learning attribute configuration using the A, B, C and D.

   F. Check out the missingness of the machine learning attributes.

   G. Check out the types of the machine learning attributes.



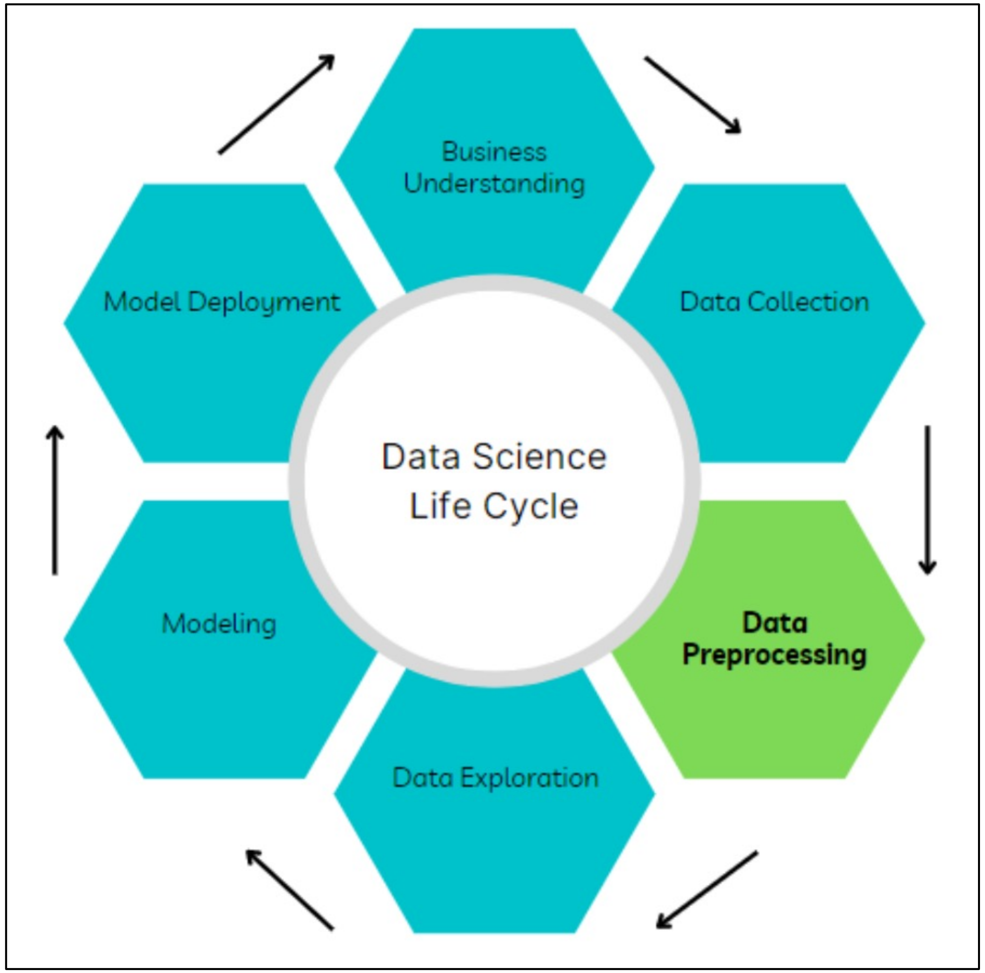
Phase 2 Outline - Lecture 2 Lab (continued)

I. Prepare the data to better expose the underlying data patterns to machine learning algorithms. (continued)

   H. Use Scikit-Learn pipelines and a column transformer to build a preprocessing pipeline that will complete the following:

      i. impute missing values
      ii. transform nominal attributes
      iii. scale the attributes

   I. Appy the preprocessing pipeline to the data frame to create a transformed data frame ready for EDA and machine learning.



Phase 2 Outline (continued)

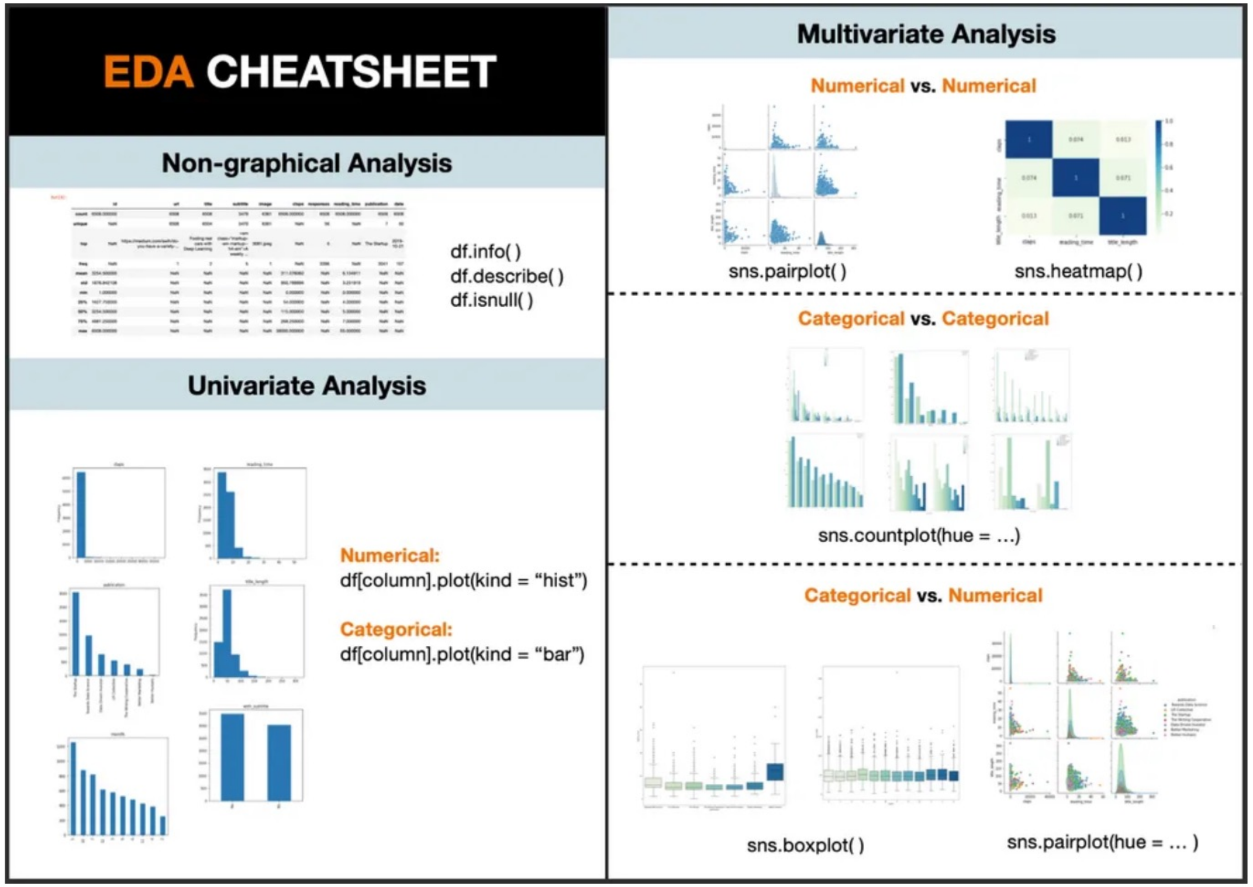
II. Explore the data to get insights (EDA).

A. Check out https://www.codecademy.com/article/eda-prior-to-unsupervised-clustering to review EDA options in preparation for clustering.

B. Conduct the following analysis on the numerical (interval and ratio attributes):

i. Pair plots to look for possible clusters.

ii. Histograms to look for possible clusters.



EDA CHEATSHEET

Non-graphical Analysis

df.info( )
df.describe( )
df.isnull( )

Univariate Analysis

Numerical:
df[column].plot(kind = "hist")

Categorical:
df[column].plot(kind = "bar")

Multivariate Analysis

Numerical vs. Numerical

sns.pairplot( )

sns.heatmap( )

Categorical vs. Categorical

sns.countplot(hue = ...)

Categorical vs. Numerical

sns.boxplot( )

sns.pairplot(hue = ...)

https://towardsdatascience.com/semi-automated-exploratory-data-analysis-eda-in-python-7f96042c9809

Phase 2 Deliverables

A PowerPoint document submitted as a .pdf file that addresses the following points:

1. Pairwise plots of all numerical pairs (interval and ratio measurement scale) attributes with a discussion of possible clusters discovered.

2. Histograms of all interval and ratio measurement scale attributes with a discussion of possible clusters discovered.

A .ipynb and .html that demonstrates the data set transformations outlined on the previous slides. The word ‘transformation’ should be in the notebook name.

A .ipynb and .html that demonstrates the data exploration described in steps 1 and 2 above. The acronym ‘eda’ should be in the notebook name.

Include any .py files that are imported by the .ipynb.

A .yml file documenting the python environment.

Deliverables
.pdf
.ipynb (plus any imported .py files)
.html
.yml

Project Phase 3

Dimensionality Reduction, Clustering and Evaluation

Phase 3 Outline - Lecture 3 and 4 Labs

I. From this point forward you should only be working with the numerical (interval and ratio measurement scale) attributes of your data set.

II. Add a default UMAP manifold learning / dimensionality reduction stage to the preprocessing pipeline built in Phase 2. Test the new pipeline.

III. Add a default K-means clustering stage to the pipeline built in step II above. Test the new pipeline.

IV. Optimize the clustering pipeline built in steps II and III above by searching over hyperparameter values and clustering algorithms.

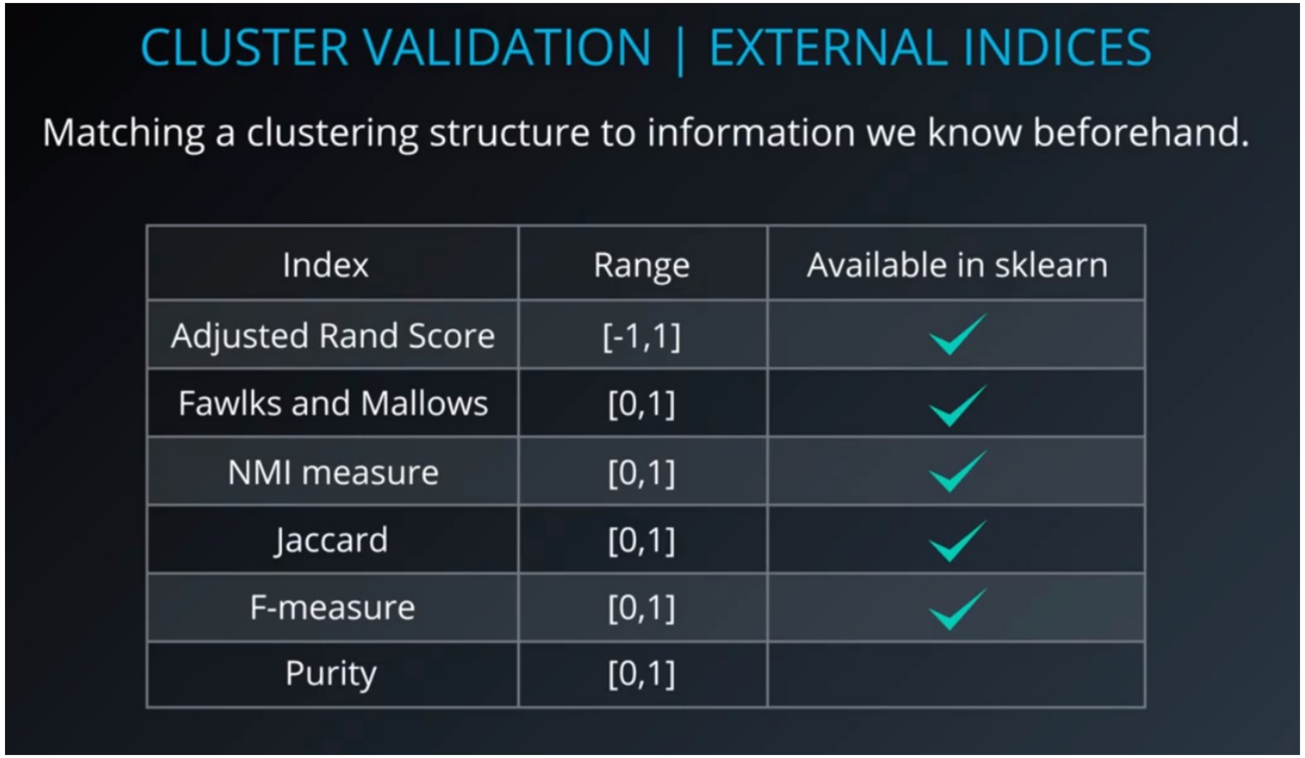
V. Check to see if the optimum solution is a false discoveries.

Note that if there is a random_state parameter available in an API you should always set the random_state for reproducibility.

This is true even in a default instantiation.

Phase 3 Outline - Lecture 3 and 4 Labs (continued)

I. Evaluate the clustering pipeline performance by comparing your clustering solution to the actual data object classes stored in the xxx_target.csv file. Use a variety of external indices for this evaluation.



CLUSTER VALIDATION | EXTERNAL INDICES

Matching a clustering structure to information we know beforehand.

Index	Range	Available in sklearn
Adjusted Rand Score	[-1,1]	✓
Fawlks and Mallows	[0,1]	✓
NMI measure	[0,1]	✓
Jaccard	[0,1]	✓
F-measure	[0,1]	✓
Purity	[0,1]	

https://julienbeaulieu.gitbook.io/wiki/sciences/machine-learning/unsupervised-learning/cluster-validation

Phase 3 Deliverables

A PowerPoint document submitted as a .pdf file that addresses the following points:

1. The optimum UMAP hyperparameter values.

2. The optimum clustering algorithm and its optimum hyperparameter values.

3. The internal indices used to optimize the clustering pipeline.

4. A table listing hyperparameter value combinations and their resulting internal index values for all clustering algorithms.

Deliverables
.pdf
.ipynb (plus any imported .py files)
.html
.yml

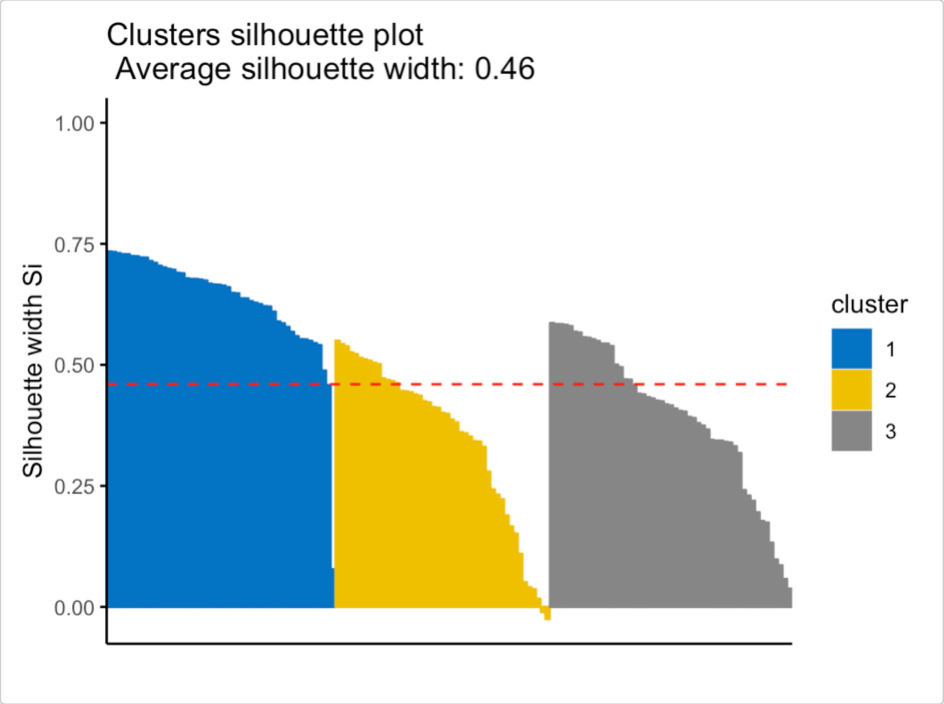
Phase 3 Deliverables (continued)

A PowerPoint document submitted as a .pdf file that addresses the following points: (continued)

5. A discussion of your work checking for a false discovery in the clustering solution.

6. The external indices used to validate the optimized clustering pipeline and the results of the validation.

7. A discussion of the performance of the clustering pipeline at clustering the data objects into their true classes.



https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/

Phase 3 Deliverables (continued)

A .ipynb and .html that demonstrates a working clustering pipeline (steps I through III above). The phrase ‘clustering_pipeline’ should be in the notebook name.

A .ipynb and .html that demonstrates the optimization of the clustering pipeline. The notebook should demonstrate a search over UMAP and Clustering hyperparameters. It should also demonstrate optimization over clustering algorithms (step IV above). The word ‘optimization’ should be in the notebook name.

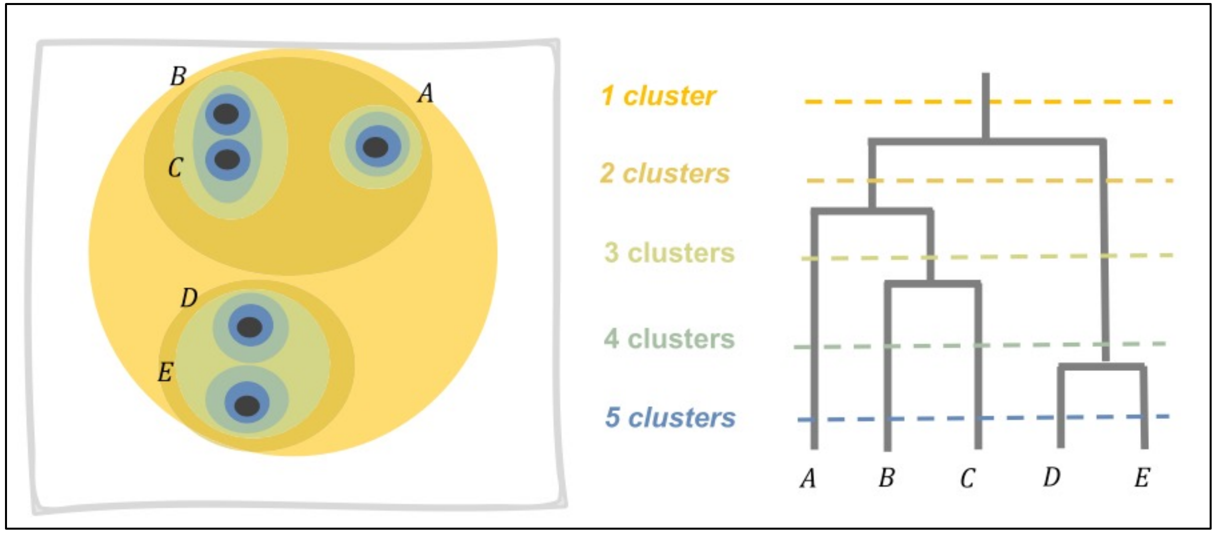
Phase 3 Deliverables (continued)

A .ipynb and .html that demonstrates checking for false discoveries. The acronym ‘afd’ should be in the notebook name.

A .ipynb and .html that demonstrates the validation of the clustering pipeline using external indices (step V). The word ‘validation’ should be in the notebook name.

Include any .py files that are imported by the .ipynb.

A .yml file documenting the python environment.


B
A
C
D
E
1 cluster
2 clusters
3 clusters
4 clusters
5 clusters
A
B
C
D
E