Clustering Assignment – Recession Months

Scenario

Are there obvious phases in the US economy that can be learned and help us to see when a recession is eminent or not? In this model you will take a big-data view of the US and apply different clustering algorithms to a database of over 100 economic variables spanning 50+ years. While the algorithms will have no knowledge of date nor what the phase of the business cycle (unsupervised learning), you have the luxury of knowing when the US has experienced a recession. Therefore, you will be able to evaluate the accuracy of your clustering algorithms to predict if a month was reflective of a recession or not. And since the data set is current, perhaps you can peer into the future to see if a recession is likely.

Data

Data available for use in the analysis is available in the file **FREDWeighted.csv** for download at the course GitHub site - https://raw.githubusercontent.com/SueMcMetzger/MachineLearning/main/chpt8/

The highly curated monthly economic indicators in this data set come from FRED (Federal Reserve Bank of St. Louis). Rather than using the raw data points provided by FRED, context has been added. The data values reflect a percentage change using a 3-month weighted average. Without weighting the change, a machine learning model would simply look at data over time and group it based on relative amounts. For example, one of the features is UEMPLT5 (number of unemployed for less than 5 weeks) value which is only valuable if compared against the population at that time *or* compared against prior months to see how much of a change differential. Like all other macro indicators provided, the % change in employment, rather than the actual value, is what will help a model to detect patterns.

In addition to the FRED macro indicators, the GitHub file **USRecessionData.csv** indicates if a given month was deemed a recession or not by NBER (National Bureau of Economic Research). USREC with a value of 1 indicates a month with recession.

Documentation on FRED outlines the details of each feature provided; there is no expectation that you dive deep into any one feature but leverage all that are viable.

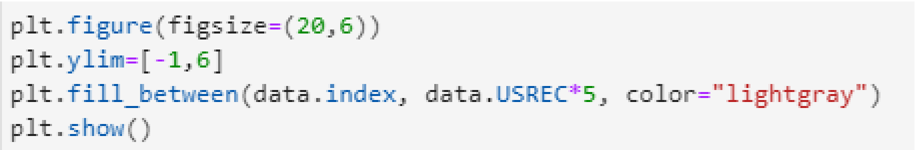
Considerations for data preparation:

- Merge together the FRED data with the US Recession data.
- Set the index for the data set to be the date field; this will be very supportive when visualizing results.
- Date should not be a feature in the models (and will not be if you make it the index).
- Delete any rows for which there are more than 5 columns of missing data.
- Delete any columns for which more than 10% of the data is missing.
- Replace any value in the dataframe that is inf (for infinity) with NaN (not a number).
- Because of the potential skewedness of data points, scale data using `PowerTransformer()` instead of `StandardScaler`()

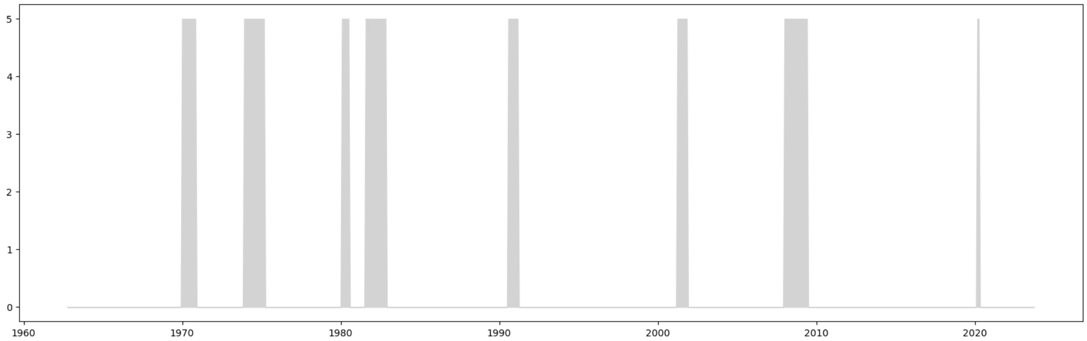
Instructions

Complete the following steps. Note that all analysis is to be performed using Python.

1. Prepare the data provided for analysis taking into consideration the prior recommendations.
2. Create a tSNE plot and if it is or is not a recession. This visual can be helpful to see if there are any obvious clustering of data.
3. For each clustering technique (K-Means, DBSCAN, and Gaussian Mixture) do as follows:
    a. Leverage the different approaches to determine the ideal number of clusters for the algorithm and select a cluster number to evaluate. See that no algorithm creates more than 6 clusters.
    b. Add the results of the algorithm to the original dataframe and then analyze how well these clusters identify with recession months or not.
    c. To assist you with your clustering analysis, the visual below can be helpful. This plot shows in gray the months for which the US was in a recession and uses the original dataframe called `data` (prior to the pipeline). Add to this plot a scatter plot (`plt.scatter()`) of your predicted clusters to visualize if any one cluster (or combination of clusters) helps to predict a recession. In addition, tools like `classification_report()`, `groupby()`, `crosstab()`, or `pivot_table()` can be helpful to quantify your findings.



```
plt.figure(figsize=(20,6))
plt.ylim=[-1,6]
plt.fill_between(data.index, data.USREC*5, color="lightgray")
plt.show()
```



4. Compare the results of the clustering algorithms. Which algorithm does a better job of identifying months of recession or not?
5. This dataset is very imbalanced. Try applying one of the 2 techniques (random under sampling OR over sampling using SMOTE) demonstrated in class. Re-run the clustering techniques, making adjustments to the algorithms & cluster sizes as needed. Does the data sampling help or hinder the model's ability to cluster months of recession or not? NOTE: This step may be easiest if you make a copy of your original notebook.

Submission

Complete the presentation AssignClustering.ppt and document your clustering algorithms and analysis. This format structure will require that your analysis be concise and direct (bullet points rather than paragraphs). Save this **presentation as PDF** to ensure it preserves all formatting and visuals accordingly. Submit to Blackboard your presentation, a **link to your Google Colab notebook** file(s), as well as a **PDF of your notebook**(s) using File -> Print as PDF. See that the notebook(s) includes the most current output for each cell (this aids the instructor).

You will be evaluated on your ability to prepare the data, implement unsupervised machine learning algorithm using Python, and the analysis & assessment of the model results. Professionalism of code and presentation materials will also be evaluated.