ESE 105 FL 2023: Case Study 1: Clustering and Classification

Due: 10/04/2023, 3:59PM

Preface

This case study is meant to allow you to work on a real engineering problem. There is no 'right' approach, nor is there a 'secret trick' that will achieve perfect results. Feel free to be creative, but also try and ground any decisions you make in terms of fundamental concepts that we have discussed in class. Using pre-built techniques that you find online, without any grounding or justification, is not in the spirit of this case study.

1 Introduction

In this case study you will use the skills and methods you have learned in linear algebra and MATLAB to analyze and classify a dataset containing information on COVID19 cases across the United States. The main idea of classification is to be able to determine the 'meaning' of an 'unlabeled' input. For this case study, your goal is to determine the geographic division (according to the US census) of unknown counties, based on their trajectory, i.e., a time series, of COVID19 case counts over time. There are many ways to accomplish classification; here we will do it using **clustering** (see also Chapter 4.4 and 4.5). You will be using your knowledge of linear algebra and MATLAB to build a method, using clustering, that takes a trajectory of COVID19 case counts and determines where in the country this county is located. You will also attempt to use your developed methodology to analyze COVID19 case data and identify trends and relationships between counties that may not be apparent to the naked eye.

2 Description of data

You are given a dataset "*COVIDbyCounty.mat*" containing COVID19 case count data for a number of counties. Key arrays are:

1. *CNTY_COVID*: $m \times n$ matrix with m counties and n dates. Entries represent the new cases in that county each week, normalized by 100k population.
2. *CNTY_CENSUS*: table with m rows, one for each county that correspond to each row in *CNTY_COVID*. Columns give summary census data for each county, e.g., fips code, division code, division name, state name, county name, and estimated population in 2021.
3. *dates*: $1 \times n$ vector with dates corresponding to each column of *CNTY_COVID*

These data are labeled according to US census divisions, which describe sub-regions of the United States. These divisions are shown in Figure 1 and in the file *us_regdiv.pdf* provided with the case study materials. **Do not confuse regions and divisions.** The divisions are: Pacfic, Mountain, West North Central, West South Central, East North Central, East South Central, South Atlantic, Middle Atlantic, New England. Figure 2 shows an example of the COVID19 case data for 12 counties across the US, including St. Louis City and St. Louis County.

3 Tasks

3.1 Lab practice:

This weeks lab is designed to be a complement to your Case Study, helping you to practice certain aspects of the MATLAB workflow, as well as become comfortable working with the data. **You do not have to finish the lab before starting the Case Study!**

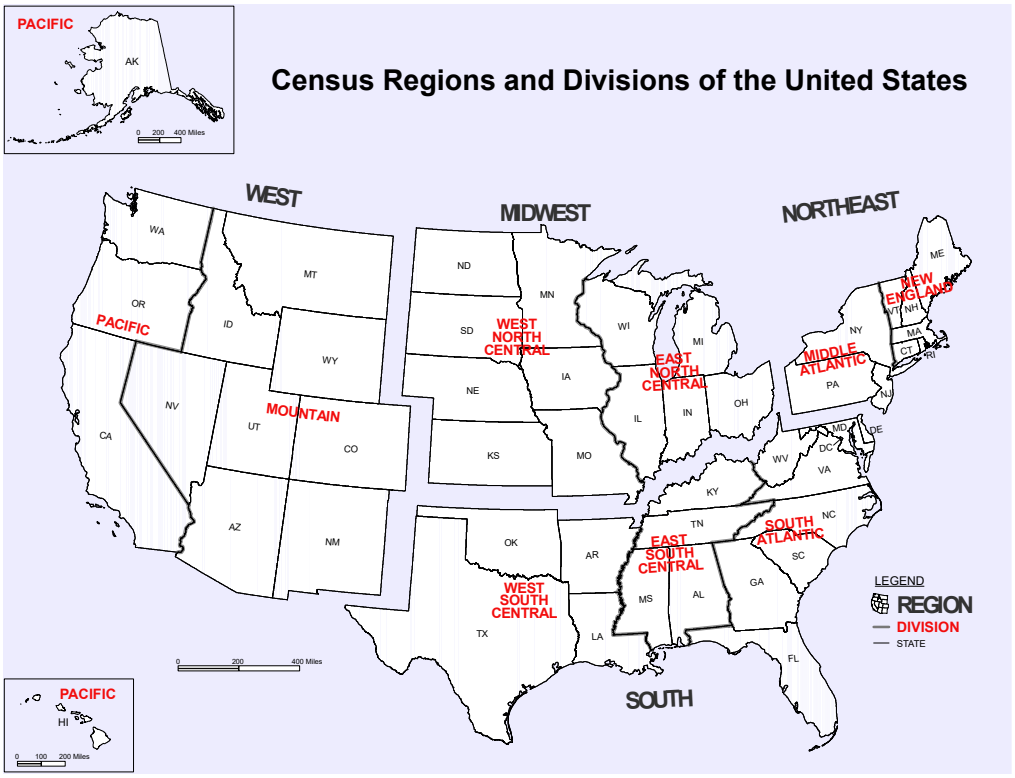

Figure 1: US Census Divisions

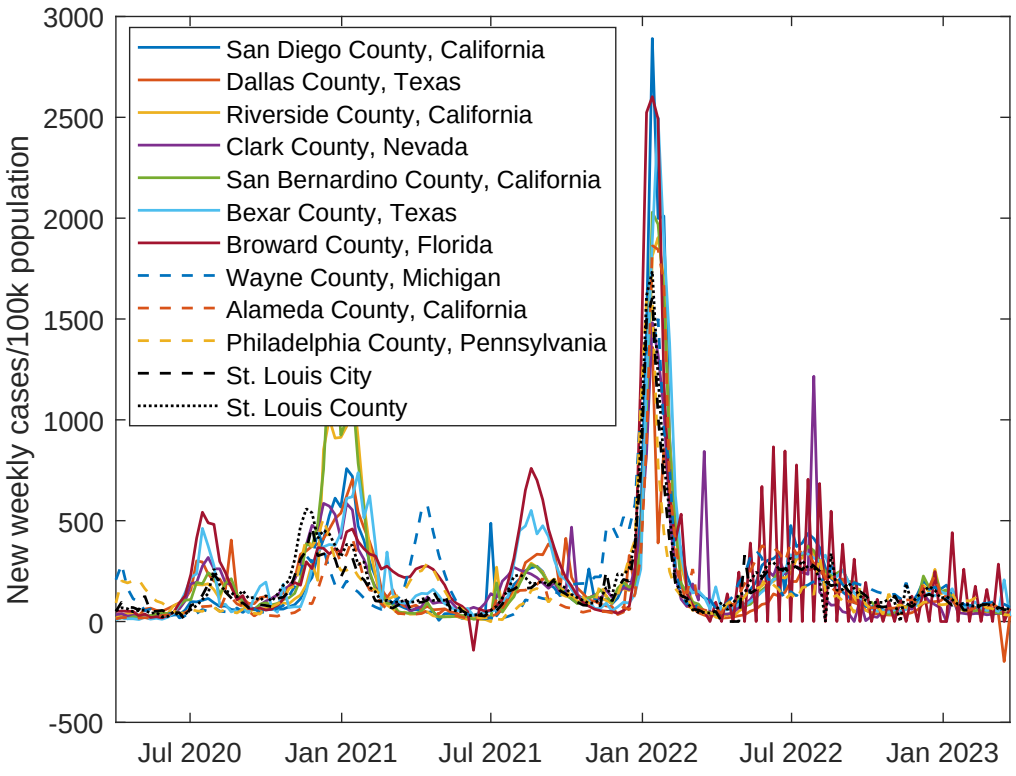

Figure 2: New weekly cases in 12 selected counties across the US. View an animated version on Canvas.

1. Identify the most populous county in each division (as per 2021 census data). This should result in 9 counties. Plot the COVID case data for these 9 counties.

2. Check whether case trajectories for the 9 counties above are linearly independent. To do this it is sufficient to verify that the angle between these trajectories (i.e., the vectors containing the case data) is not zero.

   Hint: Think carefully about how many angles you need to calculate to show linear independence.

3. Normalize the vectors you have found. (Normalizing a vector means scaling it so that its norm is then 1). Let's denote these as $d_1, d_2, ..., d_9$.

4. Find the case data for St. Louis City in Missouri. Denote it as c. Obtain the vector:

$$r_i = c - (c^T d_i)d_i \tag{1}$$

   and the quantity $\|r_i\|_2$, for $i = 1, ..., 9$.

   Interpret r_i and its norm. What do these describe?

   What might they indicate about St. Louis City relative to the 9 census divisions?

   Hint: (i) How does this expression compare to things we discussed regarding basis and basis expansion? (ii) This term r_i is sometimes referred to as a 'residual'.

5. **Submit your .m code and published PDF at the end of your lab time on Friday, 9/29**.

3.2 Clustering

Your goal is to cluster the COVID case data into k group representatives, where each centroid may be interpreted as belonging to a unique census division. You should use k-means in order to perform clustering, and may deploy the MATLAB function *kmeans* for this purpose. You may also develop your own implementation of kmeans, if you wish.

Further, you may cluster the data in their native form, or you may deploy a linear transformation of the form:

$$y = Ax \tag{2}$$

where x is the native COVID19 time-series given to you in *COVIDbyCounty.mat*, and y is a transformed version of this time-series. Note that if your clustering is performed on the native time-series, then the group representatives would be centroids of the same dimensionality as x. If you use a transformation, then your centroids will be the same dimensionality as y. Whether or not to use a transformation is a design choice that is entirely yours.[1]

3.2.1 Design your clustering method and interpret your centroids

Remember, each time you run *kmeans*, you may get a different outcome depending on your initialization. There are also many parameters that adjust the behavior of the MATLAB *kmeans* function. Feel free to experiment and explore, and make sure to justify your final design choices in your final report.

3.3 Classification

In order to use clustering for the purposes of classification, you must assign each centroid a 'meaning'. For example, you may decide that a particular centroid is representative of counties in New England. Another may be representative of counties in the Pacific. You can then test the performance of your classifier as follows:

1. Take a particular test county's trajectory and use a nearest neighbor criteria to assign it to a centroid.

2. Check whether the meaning of that centroid matches the division label of the test county.

A correct classification occurs when the true division label of the test county is consistent with the meaning of the centroid to which it is assigned.

[1] Following from the preface: the option to use a linear transformation is meant to allow for some flexibility in your design. It is not an indication that this strategy is preferred or that this will unlock superior performance.

3.3.1 Training and Testing

An important aspect of classification is the separation between training and testing sets. To test a classifier, it is often preferable to use data that were not used in the 'training' of that classifier. For this case study, this means that you should establish your clusters using one set of data, then test the 'performance' of these clusters using a different set of data.

For this purpose, you should separate your case study into training and testing sets, where clusters are obtained only from the training set. Make sure to justify your decisions regarding how you split your data for training and testing purposes in your final report.

3.3.2 Classifying the test set

Write a MATLAB script that uses your final set of k group representatives to classify your test set of counties. Your script should use a matrix called *centroids*, whose k rows contain the relevant centroids, and an array called *centroid_labels* that contains the division code that you have ascribed to that centroid.

3.4 Competition

We will take your centroids and labels and use them to classify a competition dataset that we have established but have not provided to you. Competition performance will be ranked based upon the accuracy of assigning the correct division label to each competition county, in the following manner:

$$J = N_{correct} - 0.5N_{centroids}, \tag{3}$$

where $N_{correct}$ is the number of correct regions, and $N_{centroids}$ is the number of centroids used in your classifier.

3.5 Unsupervised analysis: Finding trends and patterns in the data

Now, go further and use the methodologies you have developed to analyze your dataset and search for patterns, trends or relationships that go beyond the census divisions considered thus far. For example, the counties in your dataset may have many different aspects that are showing up in your clustering. These include census division (as you have already examined), but also specific the different COVID variants that were present; different policy decisions made by local governments; and urban, suburban, vs. rural environments, to name a few. **What do you think your clusters are picking up in terms of these distinctions?**

You might try clustering with different values of k or different norms. Intepret your findings. Such analysis can be thought of as 'unsupervised' because there is no prior imposition of labels on the data, i.e., *kmeans* is unaware of and does not consider pre-assigned labels when it clusters the COVID19 trajectories.

4 What to Turn In

1. A completed and signed honor code indicating your full and complete participation in the case study.
2. A MATLAB script (i.e., .m file) named *cluster_covid_data.m* and its published PDF that separates the data into "training" and "testing" groups, uses kmeans clustering on the "training" group, and results in the construction of k centroids.[2]
3. A MATLAB script (i.e., .m file) named *classify_covid_data.m* and its published PDF that labels the data in the "testing" group (based on nearest neighbors to a finalized set of centroids and centroid labels). Note that the finalized centroids should be designed through the use of the script *cluster_covid_data.m* from #2 above.

[2]Note that each run of this file may produce different centroids. Your published file does not need to produce your finalized centroids, i.e., the ones you use for classification, below.

4. A MATLAB data file named *competition.mat* containing:

   (a) *centroids*: a $k \times p$ matrix containing k centroids. Note that $p = n$ from the dataset if you cluster the COVID19 data directly, otherwise, p is the dimension of y in Eq. 2 above

   (b) *centroid_labels*: a $k \times 1$ vector with numerical labels that represent your interpretation of the division each centroid represents, matching the division codes within *CNTY_CENSUS* above

   (c) if applicable, the linear transformation matrix A

   NOTE: It is **important that your file and arrays are named and constructed exactly as specified**. Otherwise, you will not be included in the competition.

5. A detailed 3-6 page report, including your unsupervised analysis and interpretations of your results. Some specific considerations:

   (a) Be sure to include your interpretations regarding the quality of your centroids, and any design choices that you made in *kmeans*.

   (b) Be sure to include your interpretations regarding your testing and training separation, including your justification for how you established this separation.

   (c) Be sure to include the results of your testing, including commentary on limitations and challenges.

   (d) Do you see any disparities in data collection that would affect a public health official's ability to combat the pandemic effectively?

   (e) Do you see COVID impacting certain regions of the US more than others? How might this type of analysis help a health official or social engineer discover ways to combat systemic inequality?

   Use plots or diagrams to support your arguments. Make sure all plots, axes and axes labels are legible. Use the figure export tools within MATLAB (e.g., the function "exportgraphics()"), and avoid the use of screen capture tools.

   Your report will use the IEEE 2-column format. Templates for MS Word and Latex are provided.

5 Rubric

- Correctness of MATLAB code - 40%
  - Partitioning of data into "training" and "testing" groups
  - *kmeans* clustering of COVID19 trajectories
  - Classification of testing dataset
  - Items in *competition.mat*
- Presentation - 20%
  - Plots are easy to read and interpret, with appropriate font sizes, line widths, axis labels, etc.
  - Report should be well-organized, concise, and clearly written.
- Programming style - 20%
- Study design - 20%
  - Report addresses specific considerations noted in section 4, item 5 above.

6 Data sources

- Annual Resident Population Estimates (CO-EST2021-ALLDATA), United States Census Bureau.
- Coronavirus (Covid-19) Data in the United States. Data from The New York Times, based on reports from state and local health agencies.