Lab Package 2 Unsupervised Classification

Wednesday, October 6, 2021 1:42 PM

There are two objectives in the package:

1. Learn what geospatial research questions can be addressed by unsupervised classification
2. Learn some popular unsupervised classification methods in machine learning

- Scientific research aims to discover new knowledge from seemly disordered observations.
- A common first step is to figure out a way to organize observations so that we can summarize general characteristics in the organization: **classes and relationships.**

- The unit of analysis in geospatial research has **location** and **geometry**.
- **Geospatial classes** can be determined by locations, geometries, or properties measured at each spatial unit.
- **Geospatial relationships** can be about proximity, topology, mereology, and numerical relationships among properties.

Questions about finding and extracting spatial objects from geospatial data:

- Need to determine what geospatial classes of interest
- Need to determine what geospatial relationships of interest

In this package, our research question asks how demographic compositions vary across the city of Dallas. Do some census tracts share similar demographic compositions? If some do, how tracts of similar demographic compositions distributed across Dallas? Are there tracts representative or prototypical of different kinds of demographic compositions in Dallas? If so, what are these prototypical tracts? Are demographically similar tracts likely to be nearby each other? Are there geographic divides of tracts with distinctive demographic compositions?

-- note: Likely follow-up questions will be whether tracts with different demographic compositions exhibit differences in accessibility, opportunities, quality of life, etc.

We will use the Dallas Tracts data from Lab Package 1.

1. Create and set up a python notebook for the lab
2. Load in DallasTract2020.shp.zip
3. Prepare the data for analysis: only reserve the data (sf_xxxx) needed for analysis and make sure no null data
4. Find out what each attribute means
5. Check data and preprocess the data for analysis

Q#6: What do the attributes mean in DallasTract2020?
Sf_hh: _____
sf_hu: _____
sf_pph: _____

sf_occrate: _____

What we have are tracts with demographic data (concrete objets).
What we want to find are potential classes(or groupings) of tracts based on demographic similarity. These classes are abstract objects.
What we don't know is

1. How many different abstract objects
2. What demographic features are best for identifying these abstract objects.

The basic ideas:

1. Cluster concrete objects based on similarity among features
2. Each cluster becomes a class
3. Evaluate the goodness of each class

Analyze and select features (https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html)

1. Scale data

   Q#7: Why do we need to scale data?

   Q#8: If we want to scale data to values between 0-1, we should consider :
   A. StandardScaler
   B. MaxAbsScaler
   C. MinMaxScaler
   D. RobustScaler

   Q#9: In addition to scaler, we can use transformer to transform a skewed data distribution to approximate the normal distribution. Which of the following statement(s) is(are) correct?
   A. If the data distribution is negatively skewed, we should consider Box-Cox or quantile transformers to transform the data
   B. If the data distribution is uniform, we should consider Box-Cox or Yeo-Johnson transformers to transform data
   C. If the data is positively skewed, we should consider Yeo-Johnson or quantile transformers to transform the data
   D. If the data is bimodal, we should consider Box-Cox or Yeo-Johnson transformers to transform the data

2. Remove features with low variance

   Q#10: What are the two features with variance less than 0.5?
   A. Sf_totalpo and sf_pop_18p
   B. Sf_pacific and sf_someoth
   C. Sf_hh and sf_hu
   D. Sf_pph and sf_occrate

3. Identify highly correlated features and retain only one of the pairs
   The threshold value for high correlation is a subjective judgement. Usually consider correlation coefficients greater than 95% or 90%. In this exercise, we use 90% as the threshold

Identify clusters

There are many machine learning methods to identify clusters.
Check https://scikit-learn.org/stable/modules/clustering.html
We will discuss and practice the following methods

1. DBSCAN
2. KMeans
3. Agglomerative Clustering
4. Affinity Propagation
5. Self Organization Maps

DBSCAN

Q#11: What is a required parameter for DBSCAN clustering?

A. The minimum geographic distance between every pairs of samples
B. The minimum number of features for each sample
C. The minimum number of samples for clustering analysis
D. The minimum number of samples to make up a cluster

K-Means

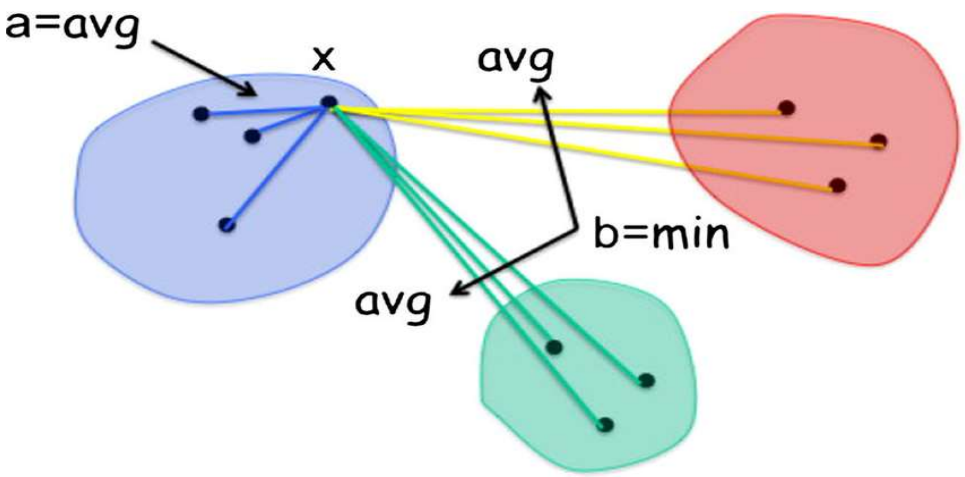
Q#12: Throughout the iterations in KMeans, the inertia value _____ (increases or decreases).

Silhouette to measure goodness of clusters

1. Silhouette_score: the mean Silhouette Coefficient of all samples.
2. Silhouette_samples: the Silhouette Coefficient for a sample

Q#13: Silhouette_samples compute the Silhouette Coefficient for each sample. The Silhouette Coefficient is a measure of how well samples are clustered with that are _____ to themselves. Clustering models with a high Silhouette Coefficient are said to be _____, where samples in the same cluster are similar to each other, and well _____, where samples in different clusters are not very similar to each other.

The Silhouette Coefficient is calculated using (a) the mean intra-cluster distance and (b) the mean nearest-cluster distance for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b).



Copy the codes for silhouette analysis at https://scikit-

learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py
and update the following:

- X
- ax1.set_xlim
- ax1.set_xticks
- ax2.scatter (check the correlation pairplot, what features may be better to see separation of clusters?)

Q#14: The silhouette plots of Dallas Tracts with demographic data show many classes have silhouette coefficient less than 0. What does it mean?

A. These are the samples with in-cluster distances less than the average distance of all samples in the same class.
B. These are the samples with in-cluster distances greater than the average distance of all samples in the same class.
C. These are the samples that have longer average distances to samples within the same class than their average distances to samples in other classes.
D. These are the samples that have shorter average distances to samples within the same class than their average distances to samples in other classes

Use parallel plots to inspect feature values in each class.

Q#15: Based on the parallel plots, which of the following statement is false:

A. There are significant overlaps of similar tracts in the two classes.
B. Comparatively, the total population is the best feature that separates the two classes.
C. The Asian population is relatively a good feature that separate the two classes.
D. The Pacific Islanders have limited numbers of population among all census tracts in Dallas.

Agglomerative Clustering

https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html#sphx-glr-auto-examples-cluster-plot-agglomerative-dendrogram-py

Affinity Propagation

Q# 16: The greater the preference value, the _____ (more or less) clusters converged in affinity propagation.

Self Organizing Map
https://sklearn-som.readthedocs.io/en/latest/

Back to our research questions:
Do some census tracts share similar demographic compositions? If some do, how tracts of similar demographic compositions distributed across Dallas? Are there prototypical tracts of demographic compositions? Are demographically similar tracts likely to nearby each other? Are there geographic divides of tracts with different demographic compositions?

What do you conclude from the clustering analysis?