The Battle of Neighborhoods

Content table

Introduction

Relocation has never been an easy decision for anyone. It usually takes time for researching with unstructured information on the Internet. Thanks to advancements in data science, many models and libraries are now available for building such systems that can help people shorten and prioritize the list of neighbourhoods of interest, hence reduce the task load and increase the efficiency.

In order to illustrate the problem, we will take a specific case as an example: Raul graduated recently from computer science on Jaen's University and looked for a data scientist position. He finally got an offer, however, the company bases in Madrid. He decided to relocate to Madrid for something new. The question here is that where exactly he could live in that big city.

One of the concerns of Raul is that he also loves his hometown so much and really wants something similar that could help him to be less homesick.

In order to determine the similarity, we will need to somehow describe each neighbourhood as a numerical vector then apply some machine learning technique, e.g., DBSCAN, to cluster them into different groups.

This problem is not restricted to this example but has wider applications in different situations. For example, if a restaurant or shop decides to open another branch somewhere on the other side of the city, they might also need to find a similar neighbourhood, because there are always interactive effects between stores. Finding similar neighbourhood has a wide application to several situations in the real world.

Data

FourSquare offers free APIs for developers to access their database of venues. Each venue in their dataset is usually categorized into a venue category, which is described in their Developers Docs. There are 10 main categories; each includes subcategories which explicitly describe the venue, e.g., Sushi Restaurant or Fishing Store:

Categories	Number of subcategories
Arts & Entertainment	36
College & University	23
Event	12
Food	91
Nightlife Spot	7
Outdoors & Recreation	62
Professional & Other Places	41
Residence	5
Shop & Service	145
Travel & Transport	34
Total	**456**

Methodology

We can apply several machine learning techniques in clustering to achieve the goal. KMeans and DBSCAN are two of the most popular unsupervised algorithms that we can apply to solve the current problem. However, there is no gold technique for all problems.

DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN), one of the best advantages is that it does not require the input of cluster numbers, as we have very little knowledge about the number of cluster for a big city.

DBSCAN is much slower than KMeans in term of time and complexity and not work well over cluster with different densities

KMeans

K-Means requires the number of clusters at first; we can try with different K to see home many clusters might help us answer our question. In this project, we chose K = 8 (Default k described in the docs of sklearn)

Result

I only evaluate KMeans clustering results as DBSCAN falied to cluster this dataset, because DBSCAN always returns only 1 cluster (a half of the neighbourhood as outliers and the other half in a single cluster)

The result shows that actual neihborhood is most similar to the neighbourhoods in cluster 3

Discussion

Though results of unsupervised learning model are sometimes hard to be evaluated, we can always reflect them with some current knowledge. For example, if it is a recommendation system for e-commerce, we can easily compare the revenue before/after application to evaluate the system. Here, I used my geographical-social knowledge about the current neighbourhood to support the validity of the results.

Another point that I would like to discuss at the end of this project is the feasibility of DBSCAN vs KMeans in this problem (or this dataset). While DBSCAN seems to have more advantage (does not need numbers of cluster, able to detect outliers,..), it is not robust to clusters with different densities over high-dimension data space.

Conclusion

Recommending likelihoods in commercial areas is an important problem, as it helps people relocating to new areas can efficiently determine their target neighbourhoods. Not only to assist personal purpose, has it also benefited companies who are considering relocate/expand their branches to other cities/areas. When considering a neighbourhood, people usually want to know what services/venues available around their residence; and whether it is similar to their current living place and/or their habits.

Thanks to the availability of FourSquare APIs, we can easily retrieve the information of interest. FourSquare also provides a nicely structured hierarchy of venue categories including 10 main categories with more subcategories. In this project, we used a histogram vector of number of venues over these 456 subcategories to describe every neighbourhood. With the data extracted of the neighbourhoods in the target city, we can apply some clustering algorithms, such as DBSCAN or KMeans, to train the model, and then use that model to predict the cluster for the current/original hometown.

In this project, to illustrate the model, we used data of neighbourhoods in Madrid as the target city to train the clustering model. Results suggested that only KMeans is robust for this particular problem. A KMeans model clustered Toronto's neighbourhoods into 8 clusters and predicted the actual neihborhood is most likely to belong to cluster 3, which includes other neighbourhoods.

After reflecting the result on geographical map as well as histograms of venue distribution over categories, we believe that the recommendation is fairly appropriate with likely geo-social meaning and similar distribution of venues. This model helped Raul in this example to reduce the number of potential target neighbourhoods by 85%. More experiments might be needed for more solid conclusion; however, here we would conclude that KMeans might be an appropriate algorithm for clustering neighbourhoods' data of venue-category histograms.