NEIGHBORHOOD RATING SYSTEM IN GUADALAJARA

By

Alejandro Salvador Orozco Guevara

Introduction

Guadalajara is the second most important city in Mexico, and is famous for tequila and mariachi music. Guadalajara combines colonial and modern architecture with perfect harmony, and a name has been forged as one of the main tech hubs in Latin America. I have lived in Guadalajara for almost thirteen years, and my project is to rate their different neighborhoods based on my criteria, so that a person interested in moving to Guadalajara receives some guidance on the best neighborhoods. This is the methodology: a dataset will be obtained with different venues as features and the number of occurrences as values, and neighborhoods as indexes. Then I will rate some neighborhoods I know with one, two, three, four or five stars, depending on how good the neighborhood considers. A classification model will be trained with that dataset, and will be used to rate the rest of the neighborhoods, which I do not know and cannot qualify a priori.

Data

The names of the neighborhoods of Guadalajara, as well as their coordinates, are obtained from the following source: https://datos.gob.mx/busca/dataset/colonias/resource/7a04858f-da5f-45b8-88bb-ffff7c6ae17e, a repository of official data of the Mexican government. Then, with the coordinates the venues near the neighborhoods are requested using the Foursquare API. In the venues returned by the Foursquare API there are different categories, such as types of restaurants, museums, galleries, etc. The number of occurrences of each category will be counted for further analysis. The numbers of occurrences of the different categories of venues are expected to correlate with the perception of the neighborhood.

Methodology

1. Obtain the names of the Guadalajara neighborhoods and their coordinates.
2. Use neighborhood coordinates and the Foursquare API to find venues near each neighborhood.
3. Create a dataframe with neighborhoods and categories of venues.
4. Count the occurrences of each venue category.
5. Add a classification variable (stars, 1 to 5), where the greater the number of stars the more attractive the neighborhood is.
6. Rate the neighborhoods I know.
7. Divide the dataframe into two: one to train and test the model, and another to predict the neighborhood's rating.
8. Train and evaluate a classifier based on Support-Vector Machine with the training / evaluation dataframe
9. Use the classification model to rate the rest of Guadalajara's neighborhoods.

Support-vector machine

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

When data are unlabeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The support-vector clustering algorithm, created by Hava Siegelmann and Vladimir Vapnik, applies the statistics of support vectors, developed in the support vector machines algorithm, to categorize unlabeled data, and is one of the most widely used clustering algorithms in industrial applications (Wikipedia).

Results

The results are shown on the notebook.

Discussion

A common problem that I faced to carry out this project was to depend on external sources. Sometimes the Foursquare API threw all the requested venues, and sometimes not. So one way to combat this damage was that one occasion, where the API returned all the requested data, I saved it to a local file. That way I gained independence.

Conclusion

Machine learning can be used to decipher patterns in the way we make decisions. The way I rated the neighborhoods was completely intuitive, discretionary; However, I suppose there is a correlation between the number of elements in each category of place and the qualification I granted. A neighborhood full of places that I like will imply a better rating from me. This can be extrapolated to any other user who qualifies a subset of neighborhoods, the algorithm must be able to find the relationship between the subject's decision and the characteristics discussed.