Classifying Architectural Building Typologies Using Unsupervised Learning

MLND Capstone Proposal

Jim Peraino
November 11th, 2017

Proposal

Domain Background

In the domains of architecture and urban planning, building typology is an often referenced idea when analyzing neighbordhoods, buildings, and public spaces. Historically, buildings have been classified in two primary ways: by form (or shape), and by use (what happens inside). In the early 19th century, the architect Jean-Nicolas-Louis Durand proposed the notion of architectural typology in his book Precis des Lecons d'Architecture (1). Since then, many efforts have been made to apply machine learning techniques to Urban Design and Architectural analysis, including with software such as GIS (2) and spatial recognition systems for identifying ideal sites for new development projects (3).

As an architect myself, I'm curious about ways that we can quantify qualitative information about buildings, and vice versa. There are relatively few datasets in the field of architecture, and establishing and automating methods for classifying buildings, in particular into qualitative categories, will be essential in this endeavor. Being able to classify a building as a certain typology enables many other kinds of analysis. This information can be used to assess the characteristics of a neighborhood, predict property values, identify areas for new development, or countless other tasks related to the built environment.

Problem Statement

Databases of building typologies do not exist for most of the world. To exacerbate this problem, buildings often no longer limit themselves to the typical typologies (schools, homes, hospitals, churches, for instance), and instead, are often mixed-use. However, physical charactistics of buildings can be easily observed or deduced from satelite images (3).

The problem, then, is: How can we automate the process of classifying buildings into their respective typologies by analyzing this physical information?

The problem is quantifiable and measurable because we can logically assess any individual building as to the makeup of its use. It is a problem that is replicable anywhere on earth that there are buildings, since google earth shows satelite images from which we can mine physical data.

Datasets and Inputs

New York City's Department of City Planning puts out a dataset called The Primary Land Use Tax Lot Output (PLUTO), which contain information about plots of lands, the characteristics of the buildings on that land, and various administrative districts. This dataset was obtained as a download from Kaggle (4). This dataset contains information for over 65,000 lots.

While the dataset has roughly 80 variables for each lot, those of particular interest are:

Physical Characteristics

- Total Building Floor Area
- Lot Area
- Number of Floors
- Lot Frontage
- Lot Depth
- Building Frontage
- Building Depth
- Floor Area Ration (FAR)

Other Data

- Commercial, Residential, Office, Retail, Factory, Storage, Garage, and Other Floor Area
- Building Class
- Zoning District

Since my goal is to use physical characteristics to classify buildings into typological groups, I will test different combinations of Physical Characteristics to see which best enable clustering of the data into building typologies. I will then compare the percentages of commercial/residential/office etc. floor areas in each cluster to better understand and evaluate the clusters.

Solution Statement

Unsupervised learning can be used on this task. As was used in the Customer Segment project, this solution can follow the process of: Data Exploration, Data Preprocessing, Feature Transformation, Clustering, and drawing conlusions(5). For more detail, see the project design section.

Benchmark Model

Existing methods for classifying building typology enter the use of each building one by one, through analysis of the use. In the case of this dataset, this data is included, and will serve as the benchmark model. This data will not be used in constructing the new model, but will instead be used to compare its performance. The benchmark model should have near 100% accuracy, as each item is entered individually. We expect that the unsupervised learning model will have significantly lower accuracy than the benchmark since it will be using only the information that can be gained from physical analysis.

Evaluation Metrics

The evaluation metric will be the extent to which each cluster is homogenous in terms of its program (residential, retail, etc). This can be determined by calculating the standard deviation of the primary programmatic element in each cluster. A smaller standard deviation would indicate that the cluster is more homogeneous, and a larger standard deviation would indicate that it is heterogeneous.

Project Design

The project design will follow a similar method to the MLND Customer Segment Project (5), and will use the following main steps:

Data Preprocessing

Most of the data is in a float format. For this data, it can be left as is, though it will potentially need to be scaled in feature transformation. When evaluating and considering the evaluation metrics, it will be important to use the Building Class or Zoning District data, which will need to be one-hot encoded. Some of the data will need to go through a process of feature scaling.

Data Exploration

Individual features can be identified, and their relevance to the problem at hand can be determined by calculating the coefficient of determination.

Feature Transformation

I will then use Principal Component Analysis to draw conclusions about variance in the data.

Clustering

I will then use K-Means clustering to create clusters from the data. If necessary, I may also try using a Gaussian Mixture Model. This information can be visualized.

Evaluation Within each cluster, I will determine the main programmatic element from the benchmark model, and will then calculate the standard deviation for that cluster.

Conclusions Any conclusions from the data can then be analyzed and stated.

References:

(1) https://books.google.com/books?hl=en&lr=&id=wilTAAAAcAAJ&oi=fnd&pg=PA1&dq=precis+durand&ots=hY1zElBwh2&sig=clePUvTcCCBGFLN (https://books.google.com/books?hl=en&lr=&id=wilTAAAAcAAJ&oi=fnd&pg=PA1&dq=precis+durand&ots=hY1zElBwh2&sig=clePUvTcCCBGFLN

(2) https://pdfs.semanticscholar.org/24c5/4f200302943cde042aee677956adc0b2498e.pdf (https://pdfs.semanticscholar.org/24c5/4f200302943cde042aee677956adc0b2498e.pdf)

(3) http://certainmeasures.com/spatial_recognition.html (http://certainmeasures.com/spatial_recognition.html)

(4) https://www.kaggle.com/new-york-city/nyc-buildings (https://www.kaggle.com/new-york-city/nyc-buildings)

(5) Udacity MLND Customer Segment Project

In [ ]: