Foundations of Data Science Project Proposal

Jae Hong Lee and Wyatt Napier
Boston University
1 Silber Way
jhonglee@bu.edu wnapier@bu.edu

1 Problem

As college students in Boston, housing prices are often on our minds, especially as we look for off-campus housing. Housing price is based on a variety of factors, but we want to ensure that we are getting the most value for our money. By creating a model to estimate housing prices based off of the attributes of a specific house, we will more objectively be able to determine if the housing we're looking at is truly worth its price. Then, once we can estimate housing prices based off of their attributes we can use this model to compare the housing prices in different cities to determine which city is more affordable.

2 Methods

To start, we need to determine what type of regression we would need to do. In the case that we only had two variables, an x which would be some attribute of the house, and a y would would be housing price, we could simply plot the training data on various individual graphs. From these graphs we can determine the relation between the attribute and the housing price. However, this doesn't account for all of the different attributes at the same time, and by isolating them it would likely distort the outcomes. On the contrary, at any dimension above the 3rd, we'd no longer be able to visualize the data though and which would make plotting is relatively useless.

Alternatively, we could attempt to compress the data. We are aware of some models such as CNN that compress image data in order to make it more manageable and decipherable, but this data is raw as it comes directly from a table. In this case, we still need to explore other methods of compression to determine worthy alternatives for this application. On first thought though, it could be helpful to remove the less interesting or impactful attributes such as whether or not the driveway is paved. We may also need to reduce the dimension of the data in order to replicate that of the city-specific data we will use later. We will try a few different methods such as SVR, linear regression, and potentially CNNs as a model.

Once we determine which type of regression we want to apply, we can then train our model on the data. Then, once the model is trained we can run it on the city-specific data. From here, we can draw conclusions about the inflation of prices in various cities and more accurately compare the value of homes across the US and even the world.

2.1 Support Vector Regression (SVR)

Support Vector Machine (SVM) is a model that finds the most efficient hyperplane or decision boundary to classify categories of each class.

As you can see in Figure 1 below, the red line corresponds to the boundary of the classification decision. SVM sets this boundary based on support vectors, which are observations found at the outermost edges of each class.

[1] Source: Wikimedia Commons

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

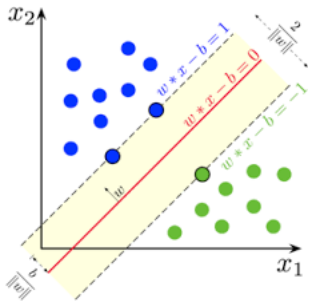

Figure 1: SVM Hyperplane[1]

The distance between the boundary and the position of the support vectors located at the outermost edges is referred to as the margin. An SVM that tolerates errors within the margin is called a soft margin SVM, while an SVM that does not tolerate errors is called a hard margin SVM.

While similar to classification methods for classifying class values, Support Vector Regression (SVR) outputs continuous numerical values instead of class values. In SVR, unlike SVM, a loss function known as an insensitive loss function (epsilion) is used to determine the amount of data included within the margin width, thereby finding the optimal boundary. SVR uses the RBF kernel function, and the parameter gamma can be adjusted to control the number of kernels.

2.2 Radial Basis Function (RBF)

A kernel is a function that helps increase dimensionality during finding the appropriate boundary. The Radial Basis Function (RBF) kernel, also known as the Gaussian kernel, is a method of repositioning data into an infinite-dimensional polynomial space. Its function is represented as follows: $K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$. This kernel function expresses the distance between two data points, x_i and x_j, using a Gaussian kernel. Gamma is a crucial factor that indicates the distance an individual data sample affects. A higher gamma value reduces the influence, while a lower value extends the distance over which it is exerted.

3 Dataset

We're looking to use at least 3 datasets. The first dataset that we're going to use to train our model comes from Kaggle. There are 1460 rows of training data and 2919 rows of testing data in this dataset. It is almost a 1:2 ratio. This dataset is incredibly verbose as it has roughly 80 different columns. We can analyze the differences in results based on the selected attributes and diversify the combinations of attributes chosen to increase accuracy.

From there, we still need to use two more datasets which we can use as input to analyze the variation in prices of homes in different cities. It is possible to change these datasets whether we want to see different cities or need a more accurate train dataset. However, the datasets we want to use are, firstly, the Paris Housing Price Prediction , and the Chicago House Price dataset. Although one dataset has more data sets while the other dataset has fewer data sets than the House Price Dataset, I believe we can test how the number of dataset affects the accuracy. Also, with 17 and 9 columns respectively, they have fewer attributes than the House Price Datas. Hence, we should carefully select the attributes corresponding to the model trained on the House Price Dataset for testing.

Using this House Price Dataset-trained model, we plan to test the two datasets and examine the differences between predicted and actual house prices in Paris and Chicago. Additionally, we aim to analyze how house prices of similar specifications vary depending on location.

4 Work of Each

In order to divide work we were planning on investigating different methods. The basis of our project at this point is SVM, so we would likely explore that together. However, we would explore other possible tools such as linear or polynomial regression individually.