House Price Prediction using Locally Weighted Linear Regression

Girish Palya

TL;DR

Does the value of house depend on the neighborhood it is located in? Real estate agents would say yes. To explore the effect of historical selling prices of houses in the neighborhood on the selling price of a house, locally (geo-spatially) weighted linear regression (LWR) model is used to predict selling prices, and results are compared to prices predicted by parametric linear regression model. LWR makes more accurate predictions compared to linear regression.

Methodology

Locally weighted linear regression (LWR) is used to minimize the following cost function.

$$\sum_{i=1}^{m} w^{(i)}(y^{(i)} - \theta^T X^{(i)})^2$$

Where,

- m: number of training observations
- $w^{(i)}$: weight of the i^{th} observation
- $y^{(i)}$: target/output variable (scalar)
- θ: parameter to be estimated (vector)
- $X^{(i)}$: features/input vector

Weights

Weight function is given by

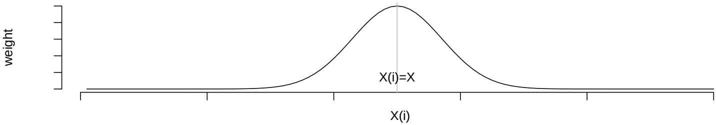
$$w^{(i)} = \exp\left(-\frac{\left(X^{(i)} - X\right)^2}{2\tau^2}\right)$$

$w^{(i)} \approx 1$ if $|X^{(i)} - X|$ is small
$w^{(i)} \approx 0$ if $|X^{(i)} - X|$ is large
X is the feature/input vector of test example
τ is the bandwidth (window size)

Net effect is that cost function sums over only $X^{(i)}$'s that are closer to X, and ignores far away $X^{(i)}$'s.

Shape of $w^{(i)}$

Shape of $w^{(i)}$ is a bell curve



Width of the bell curve is determined by the value of τ. Smaller values τ results in narrower peak of the bell curve. When value of τ is small, $X^{(i)}$'s closet to test example X have higher influence on the prediction function.

Evaluation Metric

Value of τ is approximated through trial-and-error by minimizing mean squared error (MSE) using K-Fold cross validation (k=5). In the regression setting, MSE is the most commonly-used measure.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(Y^{(i)} - h(X^{(i)})\right)^2$$

where $h(X^{(i)})$ is the prediction that h gives for the ith observation. The MSE will be small if the predicted responses are very close to the true responses, and will be large if the predicted and true responses differ substantially.

Dataset

Historical market data of real estate valuation is collected from Xindian district of New Taipei City, Taiwan, during a 10 month period in 2012 and 2013.

There are 6 input variables, and 414 observations.

- Transaction date
- House age
- Distance to the nearest MRT station
- Number of convenience stores in the living circle on foot
- Geographic coordinate, latitude
- Geographic coordinate, longitude

Output variable is house price per unit area.

Data Preparation and EDA

Transaction date (X1) is a time series. All transactions happened in years 2012 and 2013, during which time Taiwan's housing market was in a secular uptrend.

Price appreciation:

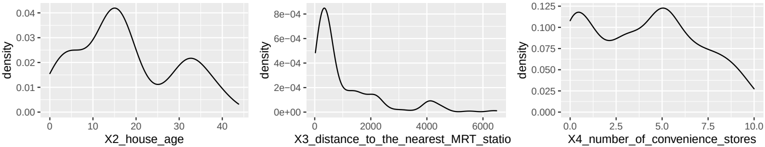
```
##   year    Q1    Q2     Q3    Q4
## 1 2012 2.76% 2.27%  0.18% 2.55%
## 2 2013 5.33% 6.05% -0.42% 3.13%
```

Time series is not stationary, and therefore cannot be included directly in regression. Out variable (house price per unit area) is scaled to adjust for the secular trend.

(contd.)

Geographic coordinates of houses are provided as latitude and longitudes. All houses in the dataset are located within 11 kilometer diameter.

Density plot reveals that variables **House age**, **Distance to the nearest MRT station**, and **Number of convenience stores within foot distance** are more-or-less normally distributed.



Bivariate Correlation

Correlation between input variables is not significant.

```
##                                                 X2
## X2_house_age                            1.00000000
## X3_distance_to_the_nearest_MRT_station 0.02562205
## X4_number_of_convenience_stores         0.04959251
```

Results

LWR can give better prediction results compared to linear regression as measured by lower MSE. Weight window parameter τ is kept at 1000. Since distance is measured in meters, this roughly equates to giving more weight to houses within half a kilometer radius (recall that houses in the training set are located within 11 kilometer diameter). LWR appears to be a better choice for a learning algorithm for predicting real estate prices.

```
##                Method MSE
## 1:               LWR  95
## 2: Linear regression 120
```

The $\sqrt{MSE}$ corresponds to average prediction error of around 24%.