Look at the Big Picture

Friday, December 13, 2019 5:19 PM

Note Written by: Zulfadli Zainal

Objective: Build a model of housing prices in California

Data contains:
1. Population
2. Median Income
3. Median Housing Prices
4. Etc

Model should learn from this data and be able to predict the median housing price in any district!

Frame the Problem

What is the main objective: Building a model is probably not the end goal.

The end goal should be how can we benefit from this model.

This is important because it will determine:

1. How you frame the problem
2. What algorithm you choose
3. What performance will you choose to evaluate your model
4. How much effort spend to tweak the model

Based on this example, end goal is whether it is worth to invest on housing at this area.

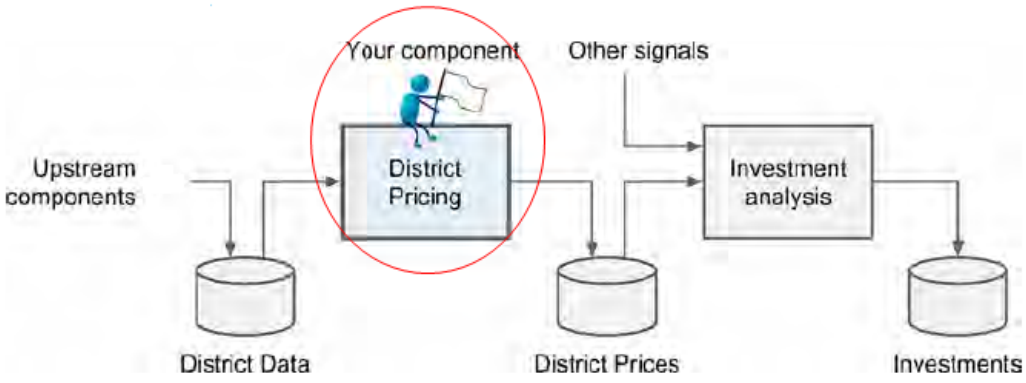


Figure 2-2. A Machine Learning pipeline for real estate investments

What is **Pipelines**?

A sequence of data processing components is called a data pipeline.

Before creating a machine learning model, **it is important to understand what is the current solution looks like? In this exercise:** Normal practice is all real estate experts will use complex calculation to estimate housing price. This is Costly, Time-Consuming, and High Error Rate.

Try to Frame the Problem!!

- Is it supervised? Unsupervised? Or Reinforcement Learning?
- Is it Classification Task or Regression Task?
- Should I use Batch Learning or Online Learning?

Based on the data given, seems like:

1. It is a supervised learning problem: Because the data is labelled
2. Also, it is a regression task: Because we need to predict a value
3. Since the data is not continuous and data can fit in the memory: Train by batch learning is enough.

Select Performance Measure - How to Evaluate your Model Performances

A typical performance measure for regression problem: RMSE!

RMSE = Root Mean Square Error

What RMSE do? *It measures the standard deviation of the errors the system makes in its predictions.*

Equation 2-1. Root Mean Square Error (RMSE)

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2}$$

Eg: If RMSE = $50,000 ->

1. Means 68% of system prediction fall within $50,000 of the actual value.
2. It also means 95% of system prediction fall within $100,000 of the actual value.

Explanation on RMSE:

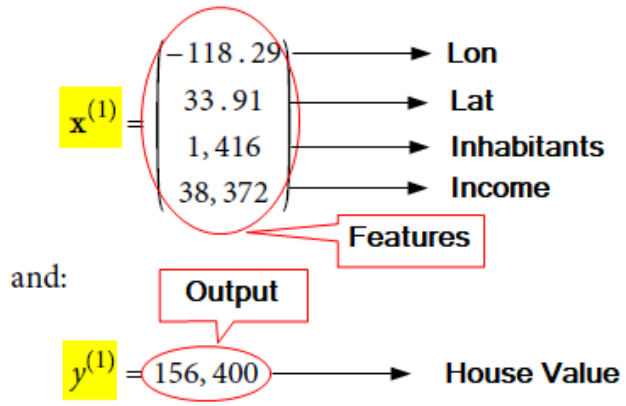
This equation introduces several very common Machine Learning notations that we will use throughout this book:

• m is the number of instances in the dataset you are measuring the RMSE on.

For example, if you are evaluating the RMSE on a validation set of 2,000 districts, then m = 2,000.

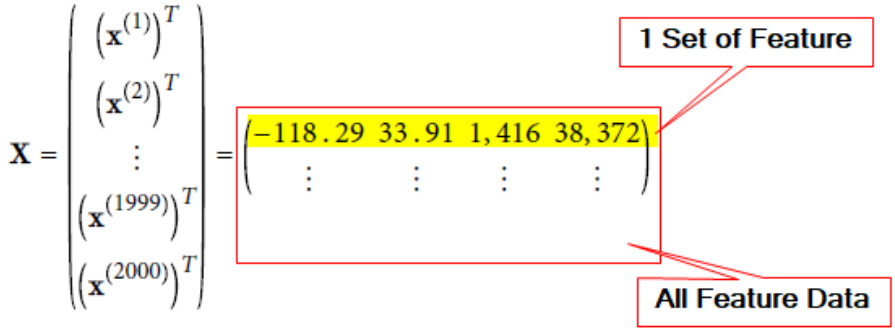
• x(i) is a vector of all the feature values (excluding the label) of the ith instance in the dataset, and y(i) is its label (the desired output value for that instance).

For example, if the first district in the dataset is located at longitude –118.29°, latitude 33.91°, and it has 1,416 inhabitants with a median income of $38,372, and the median house value is $156,400 (ignoring the other features for now), then:



• X is a matrix containing all the feature values (excluding labels) of all instances in the dataset. There is one row per instance and the ith row is equal to the transpose of x(i), noted (x(i))T.

For example, if the first district is as just described, then the matrix X looks like this:



• h is your system's prediction function, also called a hypothesis. When your system is given an instance's feature vector x(i), it outputs a predicted value ŷ(i) = h(x(i)) for that instance (ŷ is pronounced "y-hat").

For example, if your system predicts that the median housing price in the first district is $158,400, then ŷ(1) = h(x(1)) = 158,400. The prediction error for this district is ŷ(1) – y(1) = 2,000.

• RMSE(X,h) is the cost function measured on the set of examples using your hypothesis h.

Although RMSE is the most preferred performance indicator to measure regression - In case of having many Outliers districts, you may consider to use Mean Absolute Error.

Means Absolute Error - For Outliers Case

Equation 2-2. Mean Absolute Error

$$\text{MAE}(\mathbf{X}, h) = \frac{1}{m}\sum_{i=1}^{m}\left|h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right|$$

Both method is ok - **RMSE is more sensitive to outliers compared to MAE**

Check the Assumptions

Good practice to list and verify assumptions.

Purpose: To catch serious early on.

Eg: (Good Analogy!) -Read this example!!

It is important to know how the data being used. For example, the district prices that your system outputs are going to be fed into a downstream Machine Learning system, and we assume that these prices are going to be used as such. But what if the downstream system actually converts the prices into categories (e.g., “cheap,” “medium,” or “expensive”) and then uses those categories instead of the prices themselves? In this case, getting the price perfectly right is not important at all; your system just needs to get the category right. If that’s so, then the problem should have been framed as a classification task, not a regression task. You don’t want to find this out after working on a regression system for months.

For our case in the chapter, we assume that the downstream need actual prices in their system.