**Notations**

The vocabulary of machine learning is precise. This is not surprising, given that the field has its origins in mathematics and computer science. Thus, we need to define key terms. And to do that, we will use the California Housing Prices dataset.

This dataset was created by the U.S. Census Bureau during the 1990 census. As such, it contains the median house price, population, median income, and other metrics for each census block in California. A block typically has a population from 600 to 3000 people.

Stated mathematically, is a **vector** of all the **feature** values of the **ith** instance in the dataset. Below we see four features of the first census block in the dataset.

This block is located at longitude -118.29°, latitude 33.91°, has 1,416 inhabitants with a medium income of $38,372.

And the label for the first block is:

A **label** is the correct answer – the number we hope a trained model will predict. In this example, the label is the median house price.

A **dimension** is just another name for **feature**. The two are the same. So, we could rightly say that pictured above has either four features or four dimensions. The relationship between the two can be expressed as follows:

Dimension = Feature

We can place all our dataset **instances** (**x**) into matrix (). Thus, matrix contains all the instances in the dataset as well as the feature values for each one. There is one row per instance, and the **ith** row is equal to the **transpose** of **x(i)**, noted **(x(i))⊺**.

The features associated with an instance are typically shown as either a column or row of values. The transpose operation converts columns into rows or rows into columns.

A model’s prediction function is usually labelled *h* – its *hypothesis*. When you pass an instance’s feature vector to a prediction function, it returns a predicted value ŷ(i) = h(**x**(i)) for that instance. Predictions are labelled as a y with a hat on it: (pronounced “y-hat”).

* If our model predicts that the median price in the first district is $178,400, then ŷ(1) = h(**x**(1)) = 178,400. The prediction error for this district is: ŷ(1) - = 2,000.



**Statistical Vocabulary & Notations**

Because machine learning and statistics have much in common, the vocabulary of the two fields overlaps to a large degree. We begin with the basics.

Variables

A **variable** is any characteristic, number, or quantity that can be measured or counted. In statistics, there are two kinds of variable. An **independent variable** is any variable predicted to cause changes to another variable. Or stated another way, it *effects* or *influences* another variable. By contrast, a **dependent variable** is effected by the independent variable.

Dimension Reduction

WHAT IT IS: Often, a machine learning training dataset has hundreds, even thousands of features. This slows training and makes it more difficult to find an optimal solution during the process of model convergence. Dimension reduction is the transformation of data from a high-dimensional space (data with many features) into a low-dimensional space. When done well, this process retains the essential properties of the original data.

SEE IT:

WHY IT MATTERS:

GO DEEPER:

Random Projection

Random projection is used to reduce the dimensionality of a set of points which lie in Euclidean space. This algorithm projects data to a lower dimensional space using random linear projection. Random projection methods are known for their power, simplicity, and low error rates. Surprisingly, this type of projection preserves distances well. Two similar instances will remain similar after the projection while two different instances will remain quite different.

Wikipedia & Geron (p. 252)