HUGE

Hello

ML Sessions - Regression

Agenda.

1. Regression

1

Regression

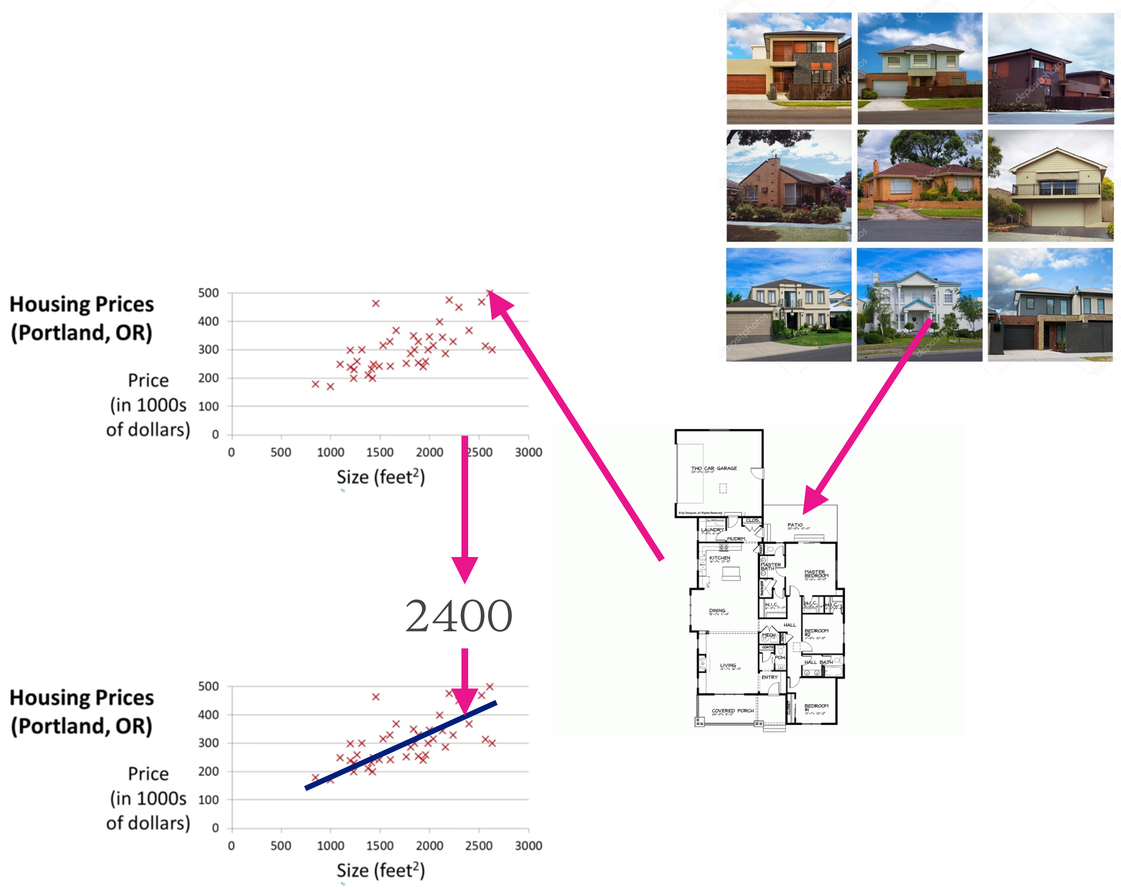
Concept

Regression

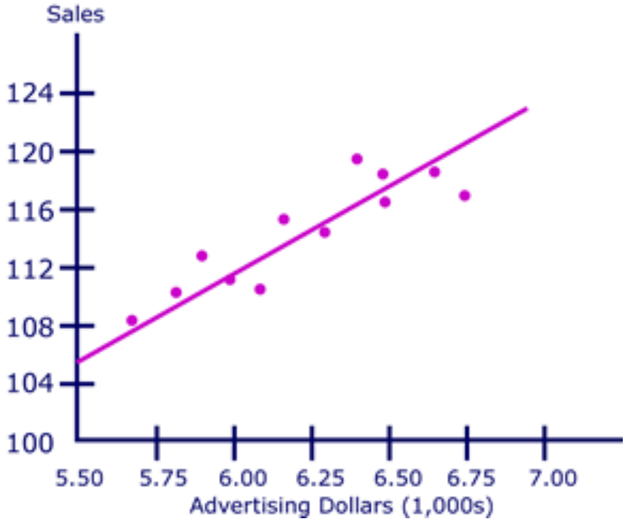
The outcome is a continuous value.

We have predictors: (**explanatory**) variables and a continuous response variable(s) (outcome or target).

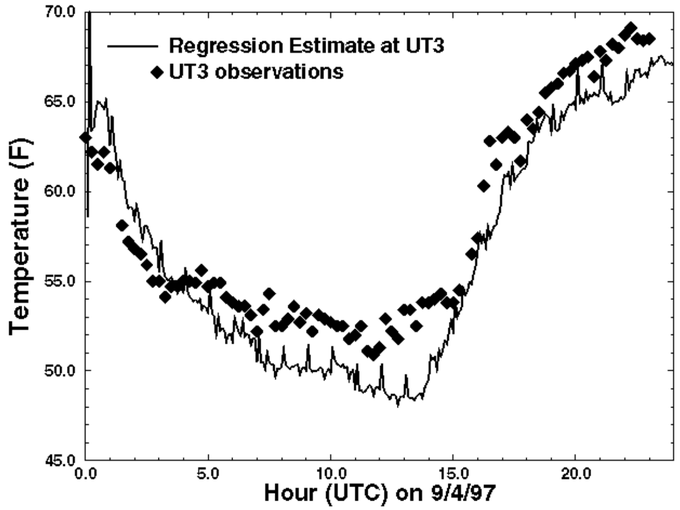
The goal is to find a relationship between those variables that allows us to predict an outcome.



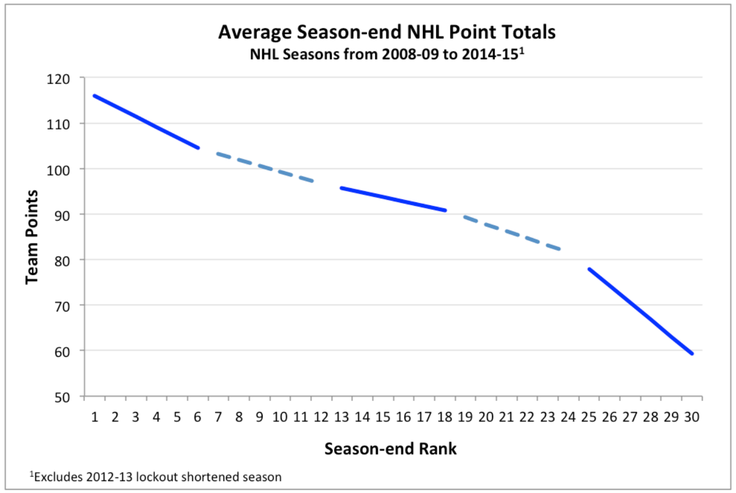
Examples



Sales prediction



Weather prediction



Scores prediction

Some algorithms

1. Linear regression
2. **K-nearest neighbors (KNN)**
3. Parzen Window
4. Random Forest
5. Neural Networks

KNN hypothesis

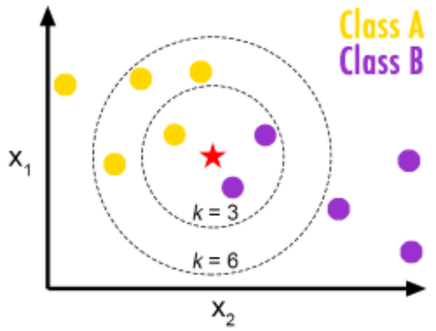
The hypothesis is that the samples with the similar output (continues value, class, etc) must be close to each other.

KNN process

1. Store all training examples
2. Classify new examples based on most similar training examples

Classification idea

For each new point the nearest k samples are found and the point is assigned to the class that repeats the most in its closest samples.



$$P(x \in \text{Class A}) = \frac{1}{3}$$

$$P(x \in \text{Class B}) = \frac{2}{3}$$

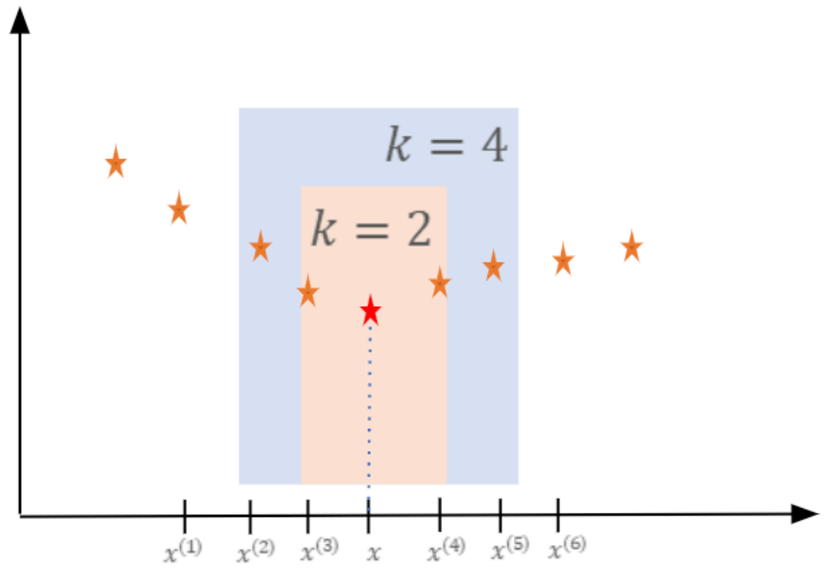
Formula

KNN in regression

Once the nearest k samples are found, we use the labels of these samples and calculate the average.

$$h(x) = \frac{1}{K} \sum_{i=1}^{K} y^V$$

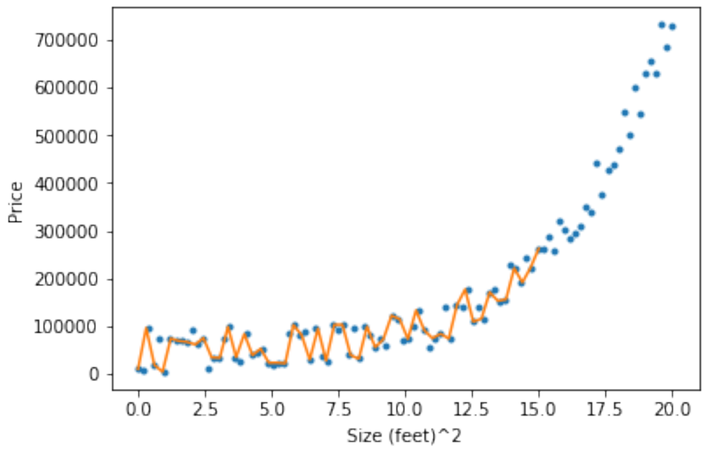
Where y^V corresponds to the value $y^{(i)}$ of that accompanies the vector $x^{(i)}$ considered neighbor of x.



Concept

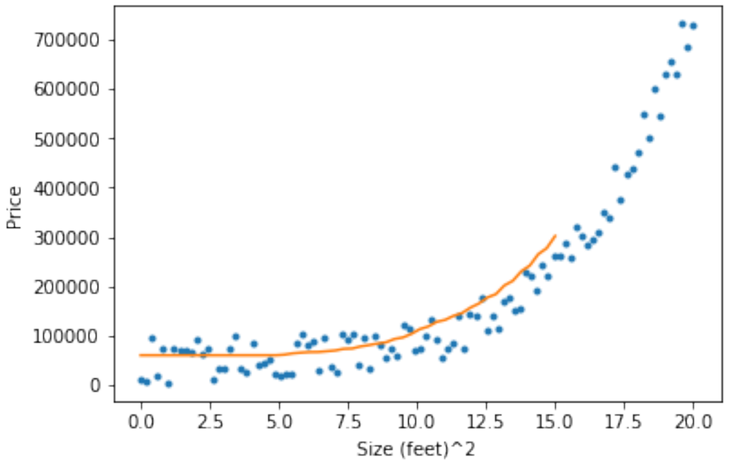
How to choose the number of neighbors K?

K = 1



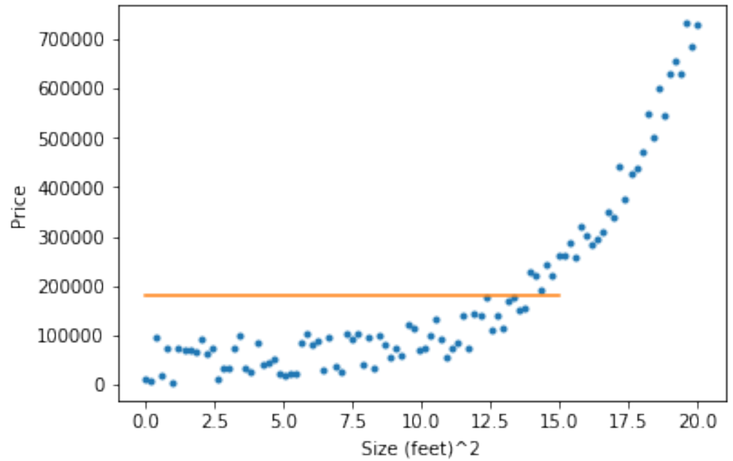
Overfitting

K = 20



Good model

K = 200



Underfitting

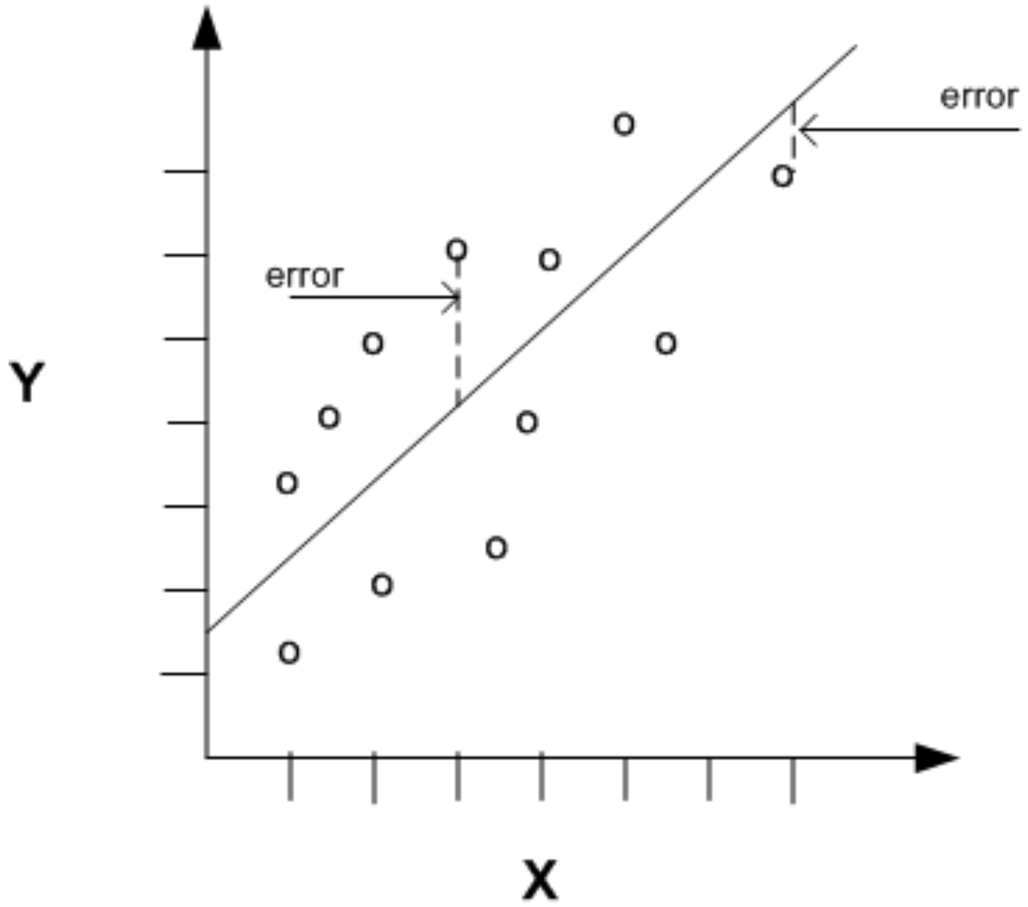
H

2

Metrics

Metrics

1. MSE
2. RMSE
3. R-Squared
4. MAE
5. (R)MSPE, MAPE
6. (R)MSLE



Mean squared error

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

It is used when there is no preference towards the solution method or another metric is not known.

Root mean squared error

$$RSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Similar to the MSE, the advantage is that it allows analyzing the error in the same scale of the labels. It is easier to understand the error.

Root mean squared error

$$RSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Similar to the MSE, the advantage is that it allows analyzing the error in the same scale of the labels. It is easier to understand the error.

R-Squared

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} = 1 - \frac{MSE}{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Metric that qualifies the quality of the regression between 0 and 1.

1. Zero means that the prediction is as bad as predicting for all values a constant equal to $\overline{y}$
2. One means that the term in parentheses is always zero so the prediction is perfect.

Mean absolute error

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

Easy to justify and is used in finance.

Allows to say how many times one model is better than the other.

A mistake of 10 dollars is twice worse than a mistake of 5 dollars.

2

Validation

Two common approaches

Bootstrapping

In this methodology, only the data are divided into a random training set that has 70% (80%) of the data and is tested with the remaining 30% (20%).

This division is made several times and the estimated metric is the average of all iterations.

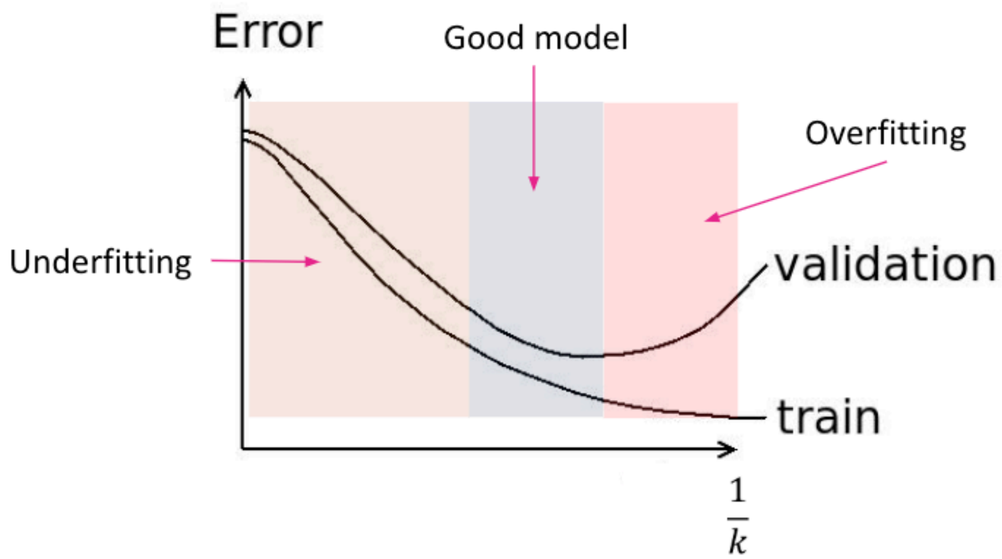
Cross-validation

Also called n fold cross-validation.

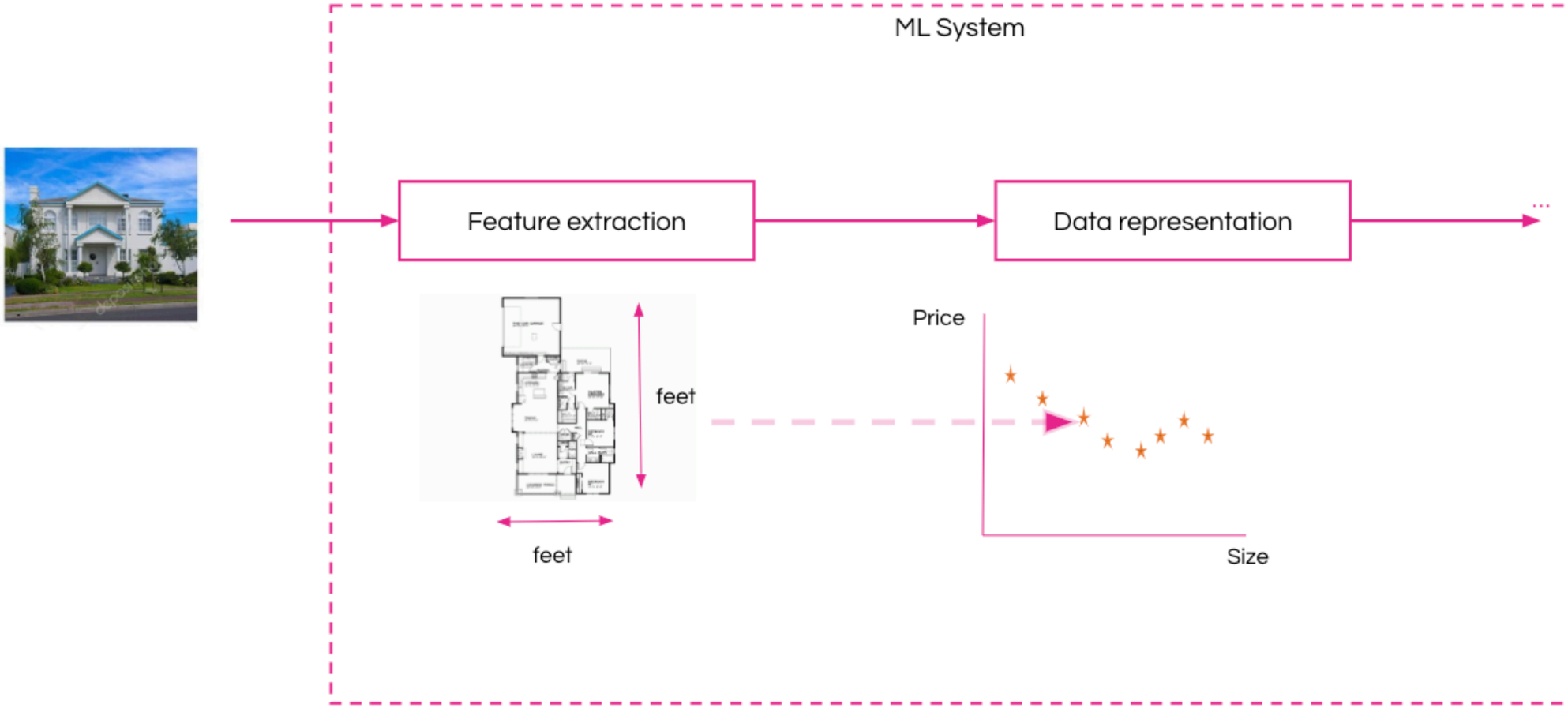
And what is a fold? It's just a subset of the data. These folds are randomly generated.

A powerful model can always memorize ...

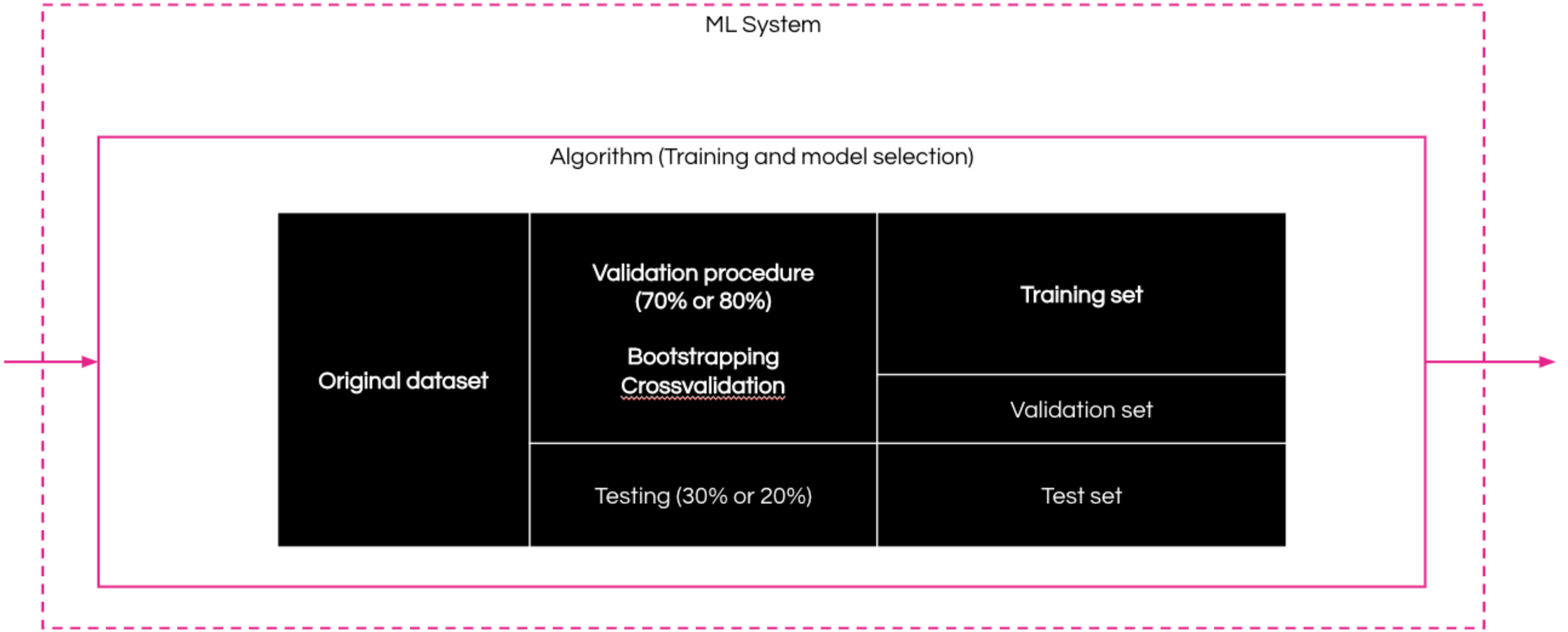
A sufficiently complex model can be adjusted to any type of data, what must be done is to know when to stop training or what parameters guarantee the generality.



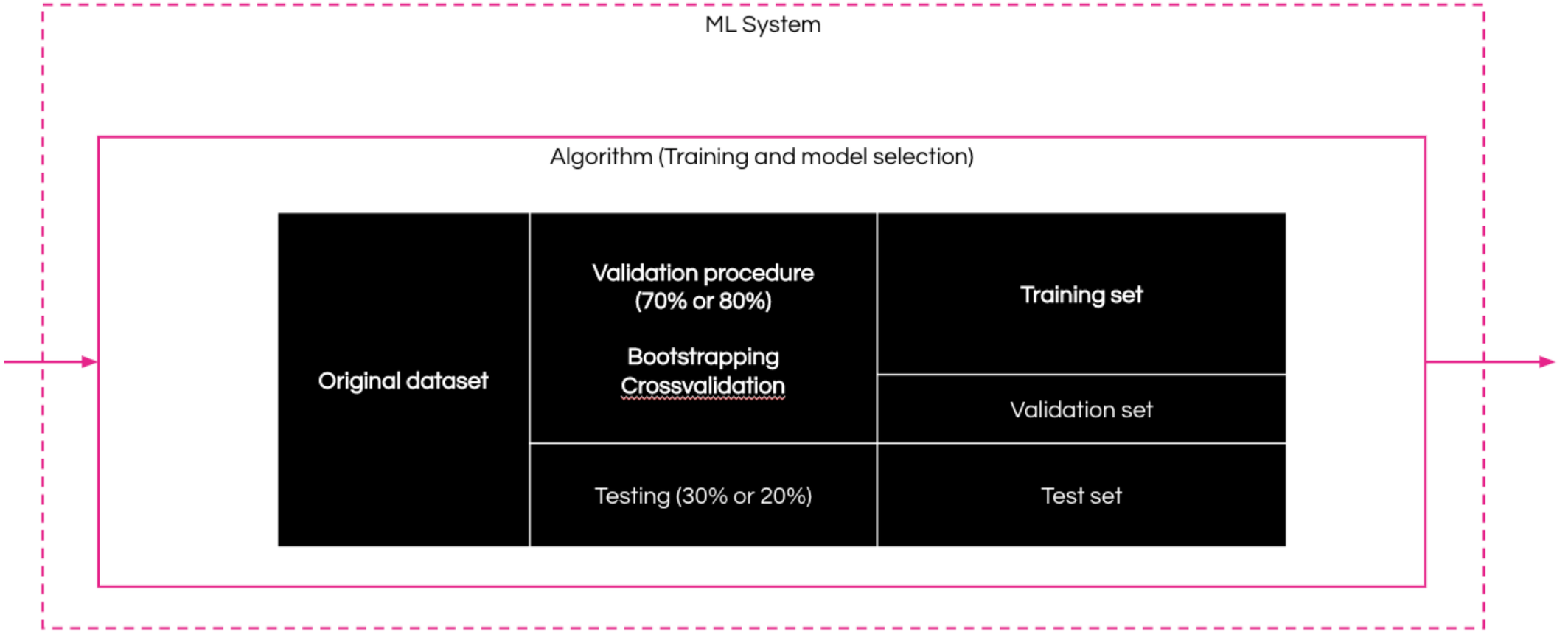
Final system



Final system



Final system



Final system