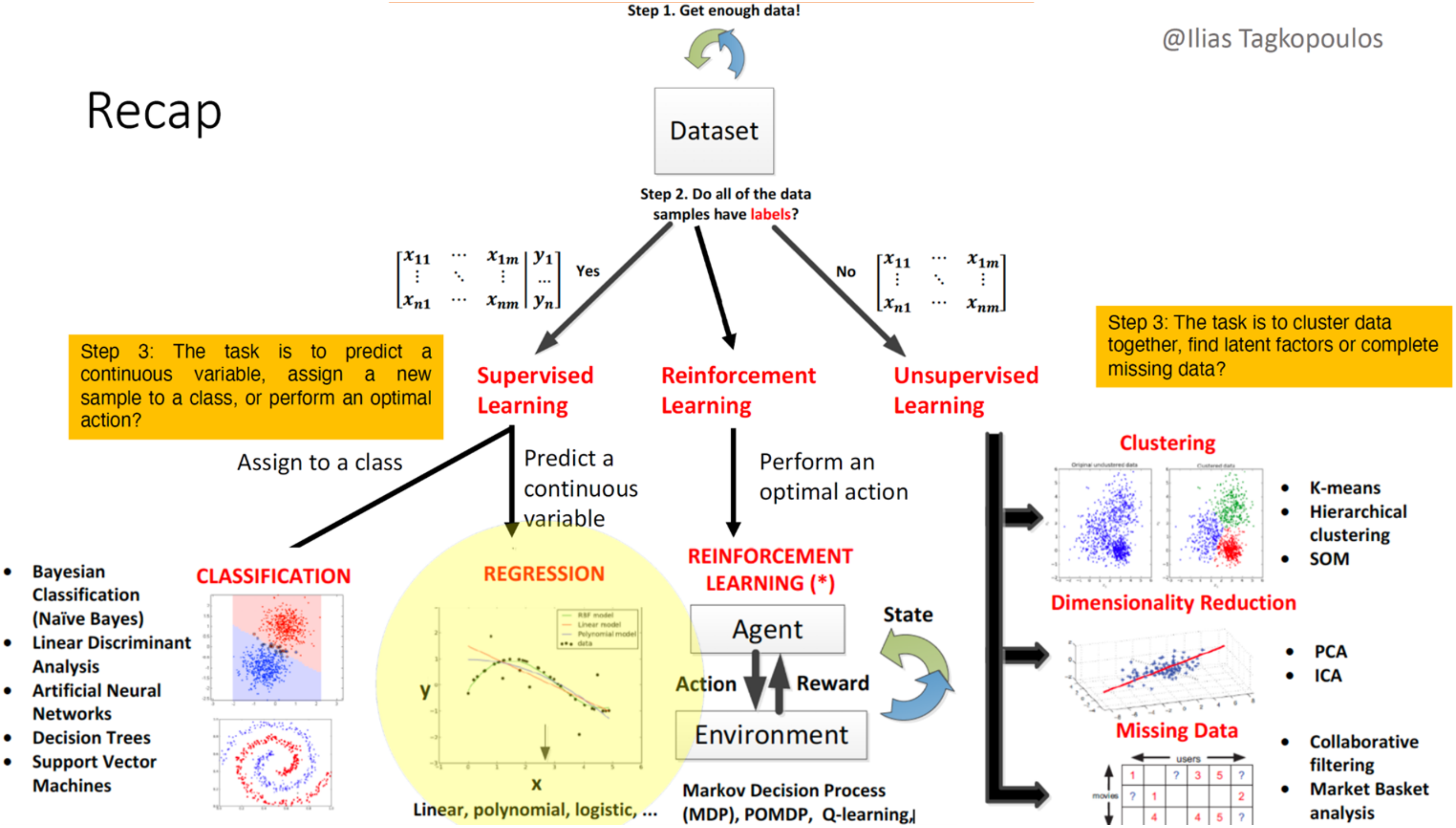
ECS 171: Machine Learning

Summer 2023

Edwin Solares
easolares@ucdavis.edu

Linear Regression

What is Machine Learning: Recap



@Ilias Tagkopoulos

Recap

Step 1. Get enough data!

Dataset

Step 2. Do all of the data samples have labels?

Yes

$$\left[\begin{array}{ccc|c} x_{11} & \cdots & x_{1m} & y_1 \\ \vdots & \ddots & \vdots & \cdots \\ x_{n1} & \cdots & x_{nm} & y_n \end{array}\right]$$

No

$$\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

Step 3: The task is to predict a continuous variable, assign a new sample to a class, or perform an optimal action?

Step 3: The task is to cluster data together, find latent factors or complete missing data?

Supervised Learning

Reinforcement Learning

Unsupervised Learning

Assign to a class

Predict a continuous variable

Perform an optimal action

CLASSIFICATION

- Bayesian Classification (Naïve Bayes)
- Linear Discriminant Analysis
- Artificial Neural Networks
- Decision Trees
- Support Vector Machines

REGRESSION

Linear, polynomial, logistic, ...

REINFORCEMENT LEARNING (*)

Agent

State

Action

Reward

Environment

Markov Decision Process (MDP), POMDP, Q-learning,

Clustering

- K-means
- Hierarchical clustering
- SOM

Dimensionality Reduction

- PCA
- ICA

Missing Data

- Collaborative filtering
- Market Basket analysis

Not enough data? Cross Validation!

Cross validation is **method** to **avoid** producing **biased models**

- A **resampling** procedure to help the model to **generalize** well
- Has a single parameter called **k** for the number of **partitions**
- k-fold cross validation

Procedure for k-fold cross validation:

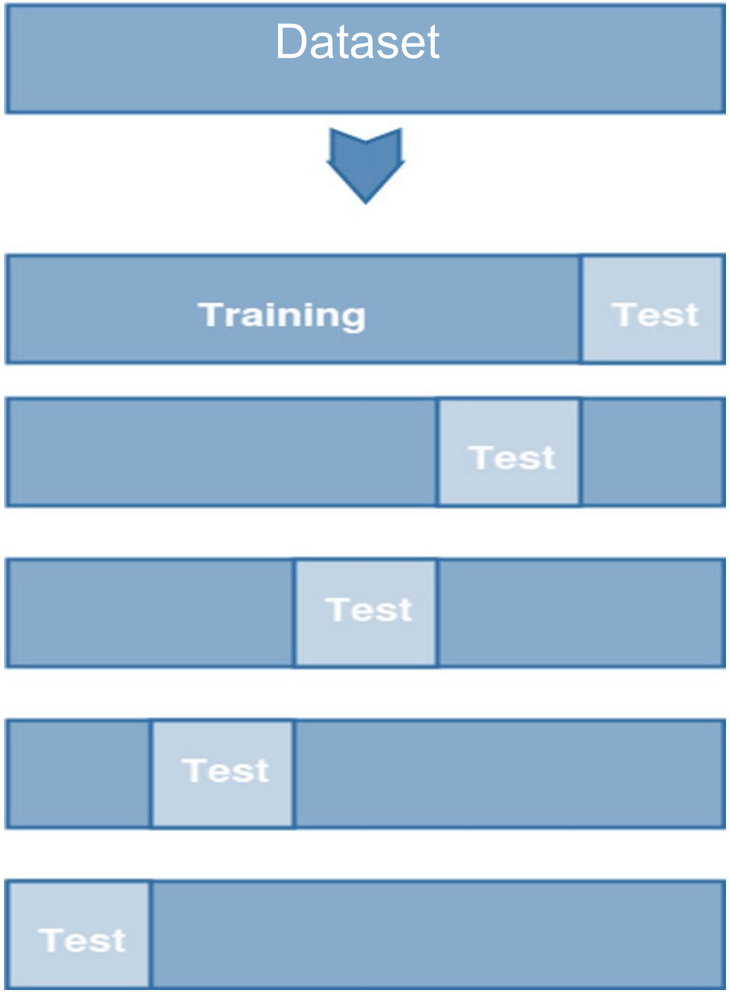
1. **Randomize** the dataset and create k **equal size partitions**
2. Use k-**1 partition**s for **training** the model
3. Use the **kth partition** for **testing** and **evaluating** the model
4. iterate **k times** with a different **subset** reserved for testing purpose each time.

Some commonly used variations on cross-validation are stratified and repeated are available in scikit-learn.

Cross Validation

```
from sklearn import cross_validation

# value of K is 5.
data =
cross_validation.KFold(len(train_set)
, n_folds=5, indices=False)
```



Outline

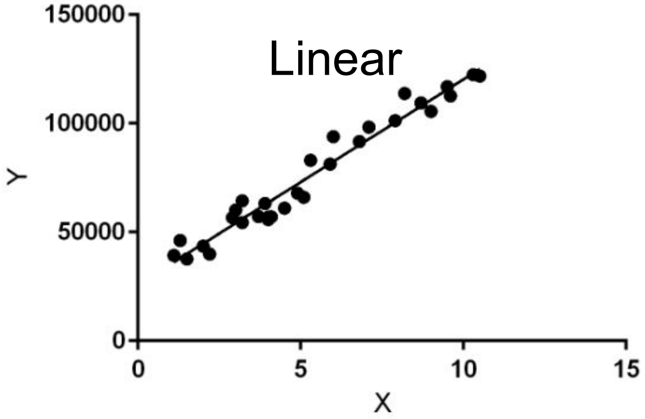
Regression Problem Setting

Linear Regression

- Linear Regression Categories
- Curve Fitting
- Ordinary Least Squares (OLS)

Gradient Descent (GD)

Identifying a Regression Problem

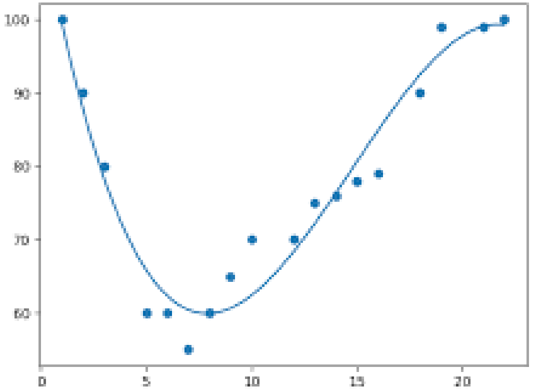


Do we want to **predict values**/targets?

Target data **continuous**?

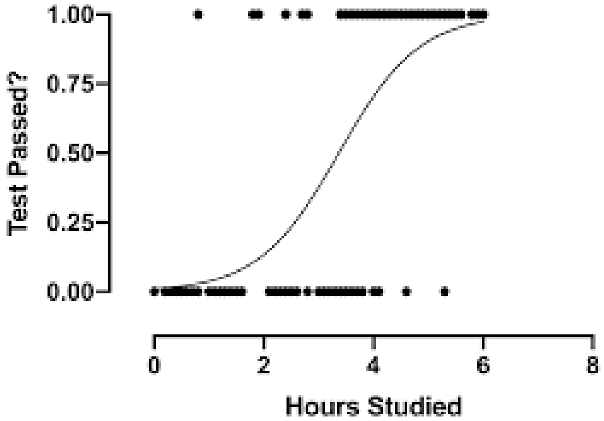
Does it **plot well** in a **scatter plot** i.e. ***y* = *mx* + *b*** where x can be any order

- Linear
- Polynomial
- Logistic
- Logarithmic
- Exponential



Polynomial

Logistic

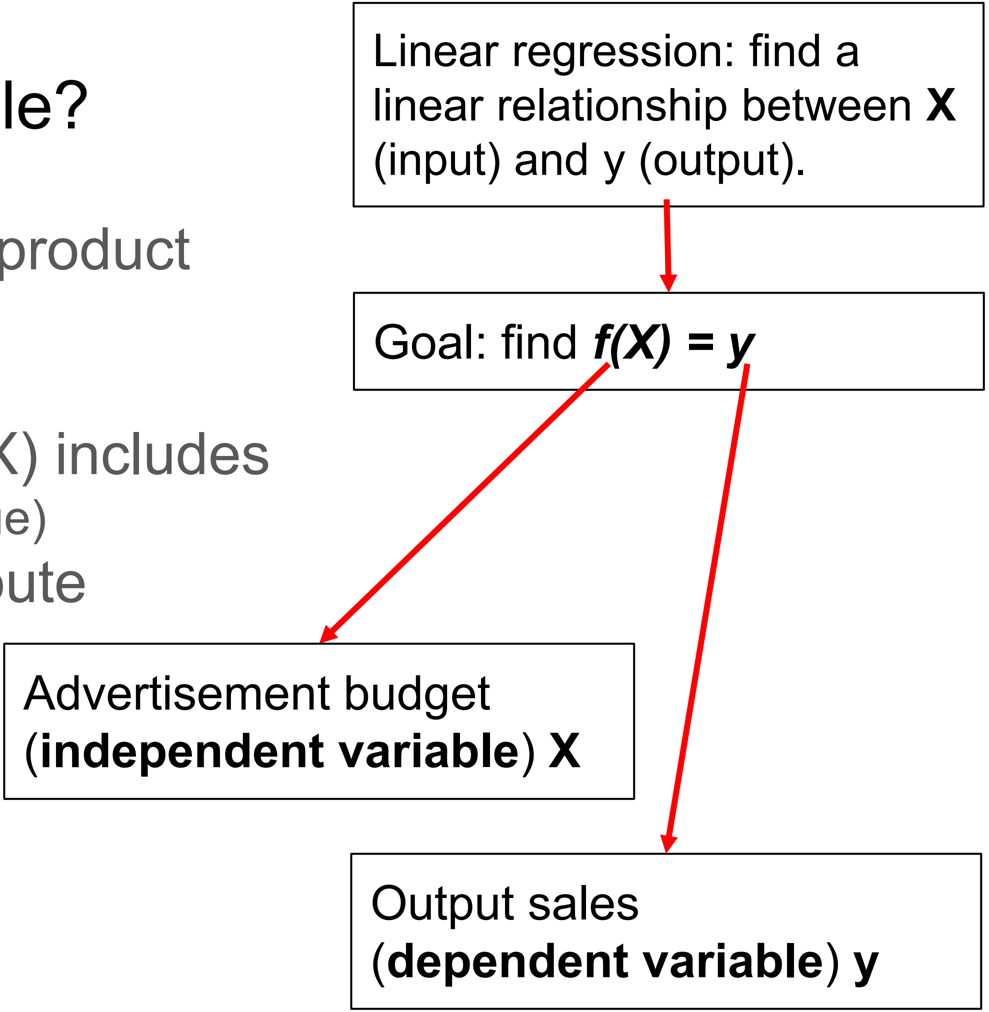


Regression Problem Example?

Predicting sales for a particular product

Data set Description

- Attribute(s) of the data set (X) includes
  - advertising budget (dollar value)
- Output Y i.e., the class attribute
  - sales in thousands of units



Linear regression: find a linear relationship between **X** (input) and y (output).

Goal: find ***f(X) = y***

Advertisement budget (**independent variable**) **X**

Output sales (**dependent variable**) **y**

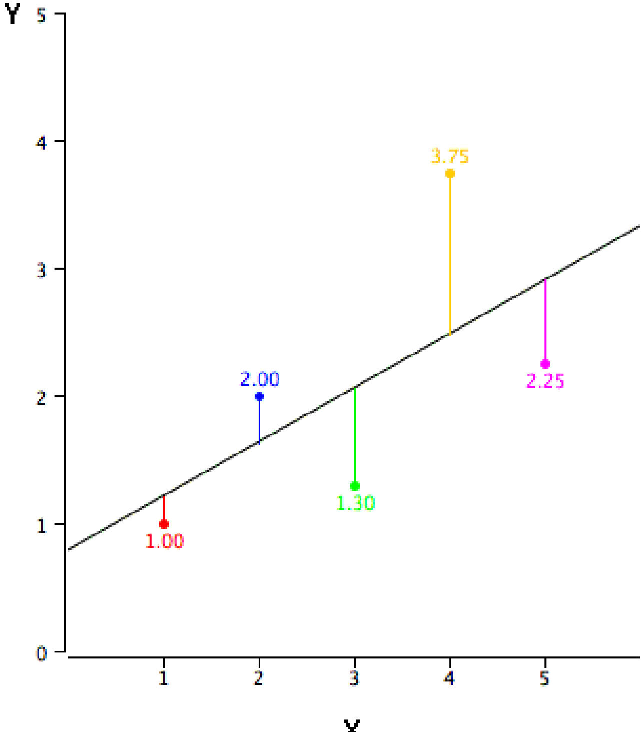
Linear Regression Model

Supervised learning

Popular statistical learning method

Predicts a quantitative response **y** from predictive attribute **X**

Linear relationship between **X** and **y**



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

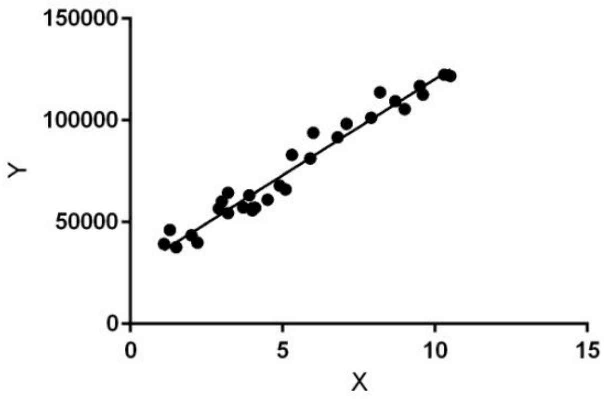
Output

intercept

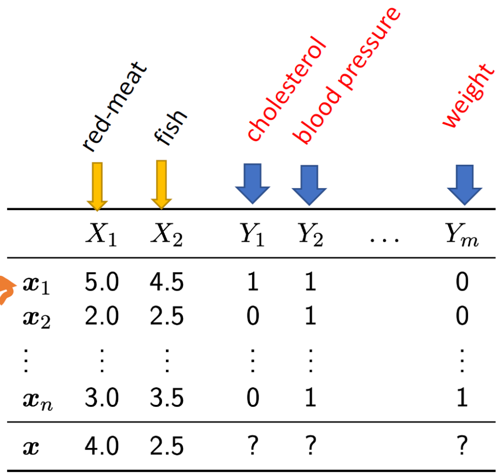
model coefficients (model parameters)

Linear Regression Categories

Simple Linear Regression



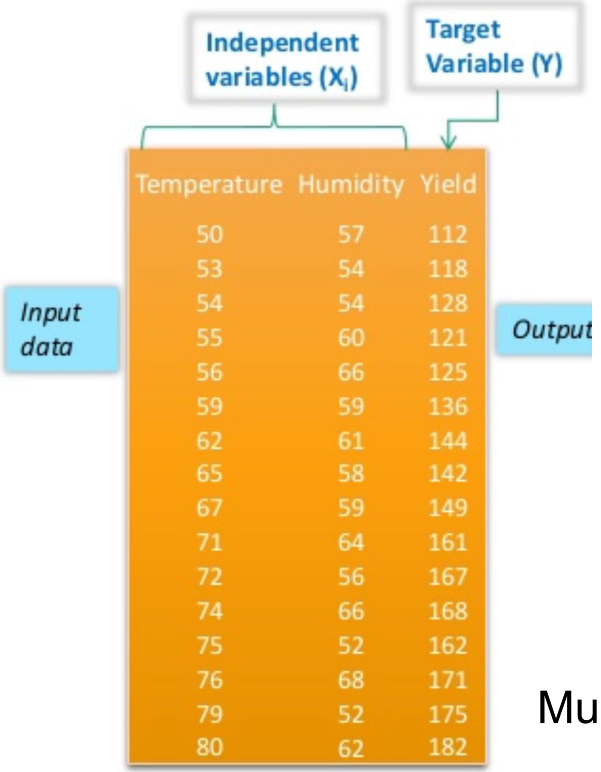
Multivariate LR/General LR


red-meat
fish
cholesterol
blood pressure
weight

	X_1	X_2	Y_1	Y_2	$\dots$	Y_m
$\boldsymbol{x}_1$	5.0	4.5	1	1		0
$\boldsymbol{x}_2$	2.0	2.5	0	1		0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\boldsymbol{x}_n$	3.0	3.5	0	1		1
$\boldsymbol{x}$	4.0	2.5	?	?		?

For a given x, predict the vector

$$Y = (Y_1, Y_2, \dots, Y_m)$$


Independent variables (X_i)
Target Variable (Y)
Input data
Output

Temperature	Humidity	Yield
50	57	112
53	54	118
54	54	128
55	60	121
56	66	125
59	59	136
62	61	144
65	58	142
67	59	149
71	64	161
72	56	167
74	66	168
75	52	162
76	68	171
79	52	175
80	62	182

Multiple Linear Regression

What is Curve Fitting

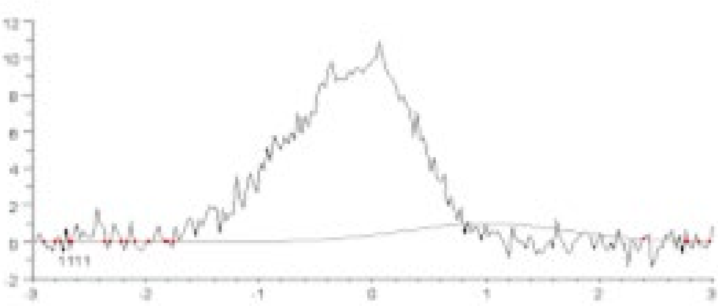
In regression analysis, **curve fitting** is the **process** of **finding a model** that produces the **best fit** with the **lowest error** to the relationships between the variables of a dataset.

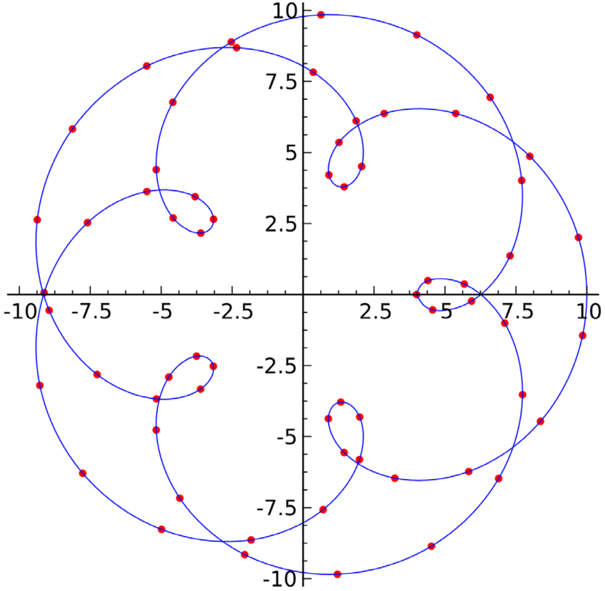
Curve fitting is the process of **constructing a curve**, or **mathematical function**, that has the **best fit** to a **series of data points**.

Interpolation where an exact fit to the data is required

Smoothing in which a "smooth" **function** is constructed that **approximately fits the data**.

Smoothing





Interpolation

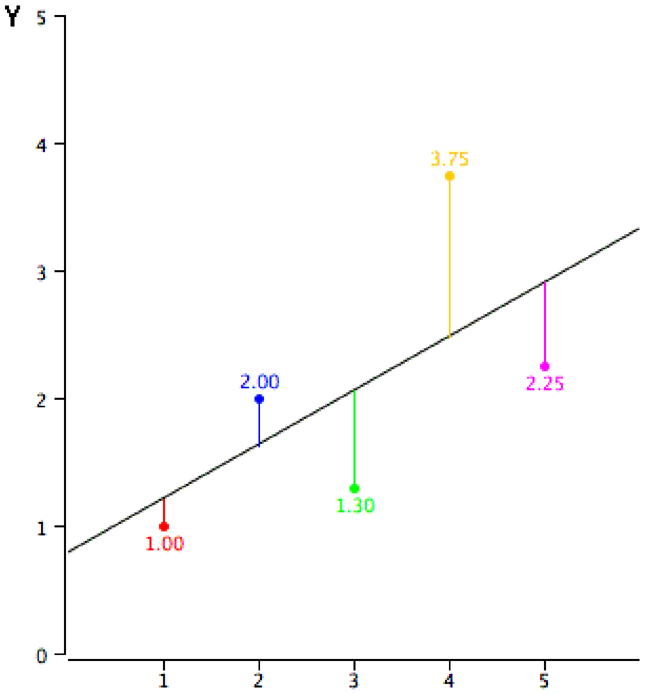
Cost Function

When **training** the model, the goal is to **minimize** the **error** and **update** the model **coefficients** to achieve the **best fit** line.

Error is the **difference between predicted value** (Y) generated by the model and the **class attribute value**.

Cost function L is used to **measure the error**:

$$L = \sum_{i=1}^{n} (\underset{\text{Observed value}}{y_i} - \underset{\text{Predicted value}}{\hat{y}_i})^2$$


Y
5
4
3
2
1
0
1
2
3
4
5
3.75
2.00
2.25
1.30
1.00