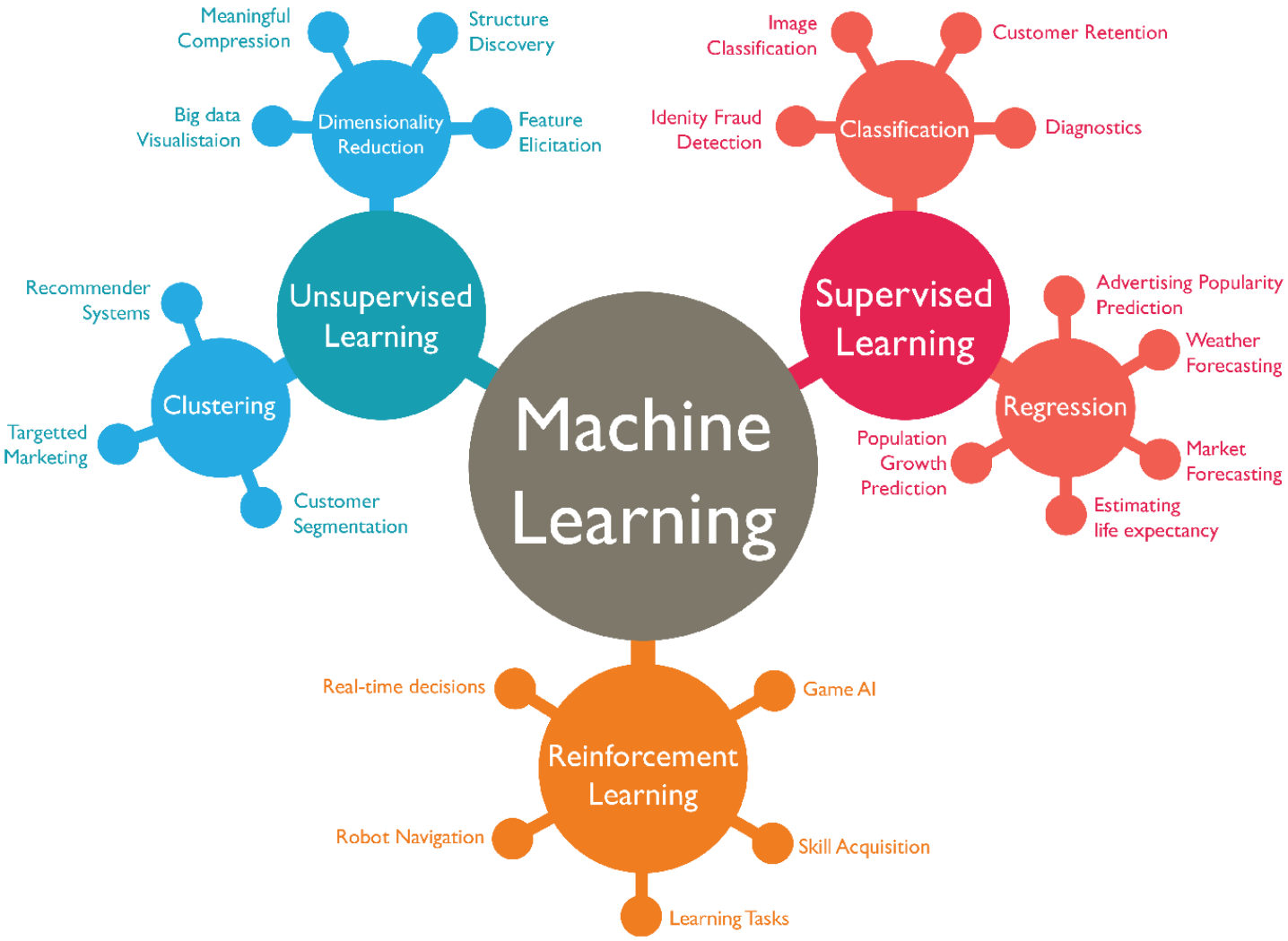
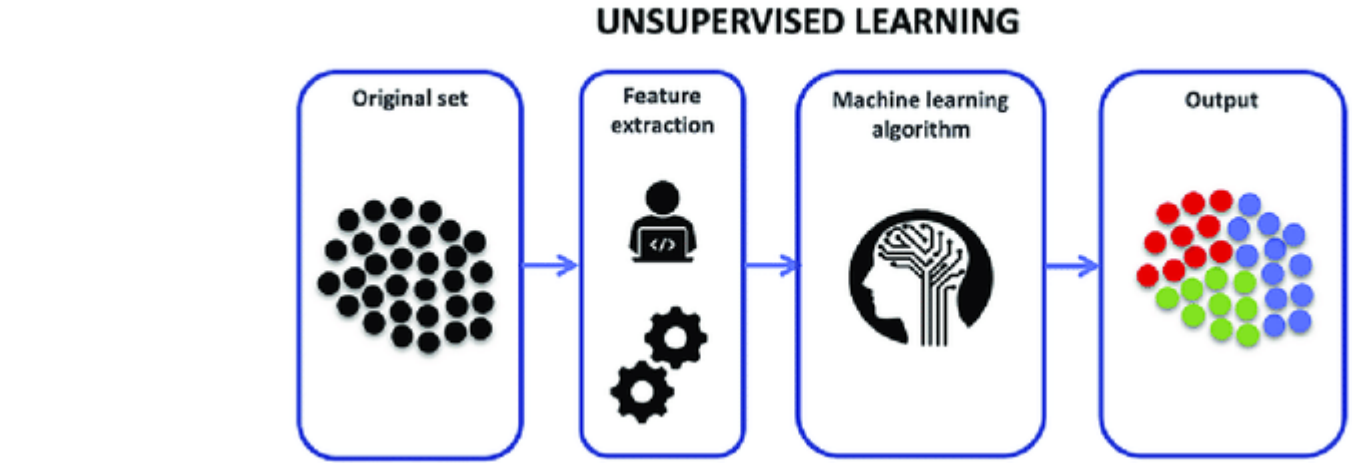
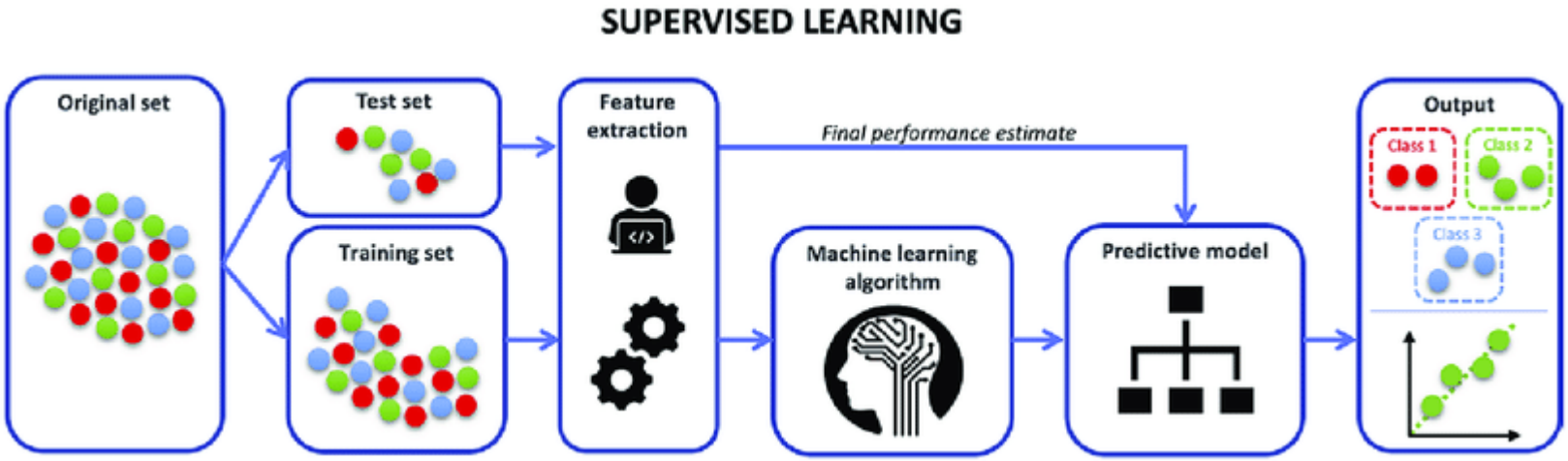
Today Focus



INVESTIC

ML Process

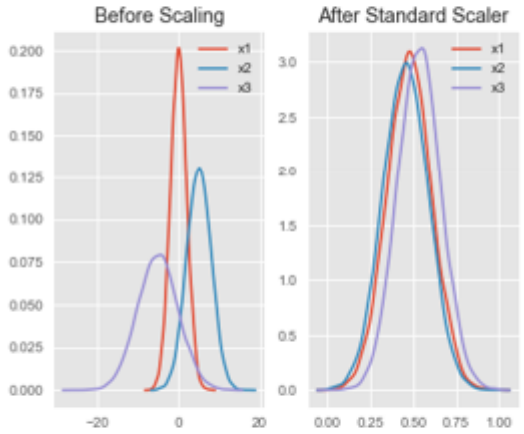




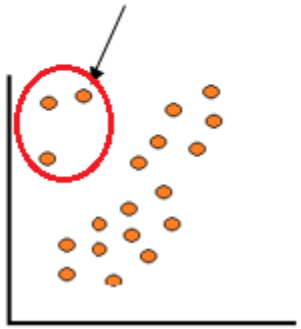
INVESTIC

Today Focus

Normalization



Outliers



Imputation

	First	Second	Third
0	100.0	30.0	NaN
1	90.0	45.0	40.0
2	NaN	56.0	80.0
3	95.0	NaN	98.0

Encoding

Food Name
Apple
Chicken
Broccoli

→

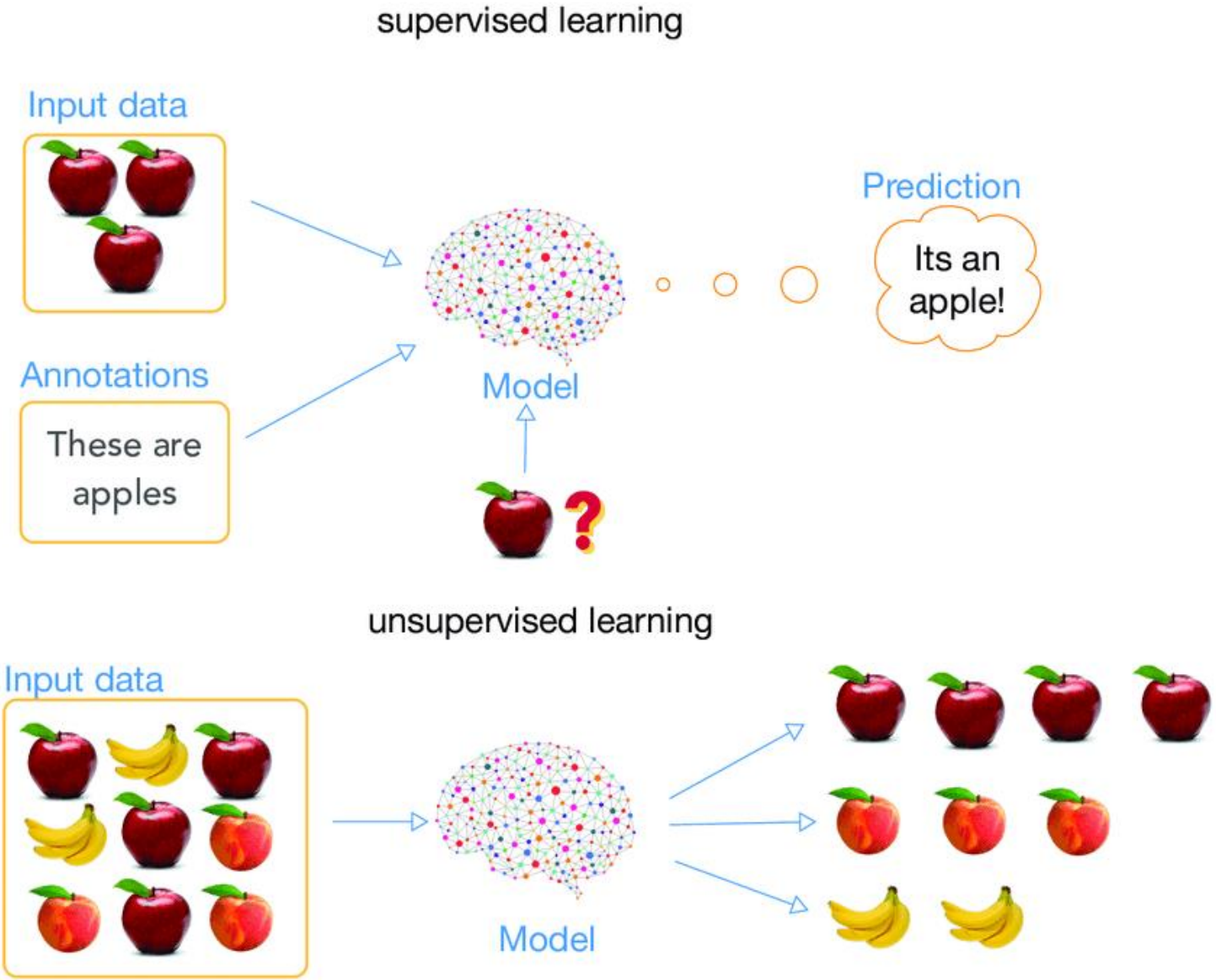
Apple	Chicken	Broccoli
1	0	0
0	1	0
0	0	1

Scaling

Data Preprocessing

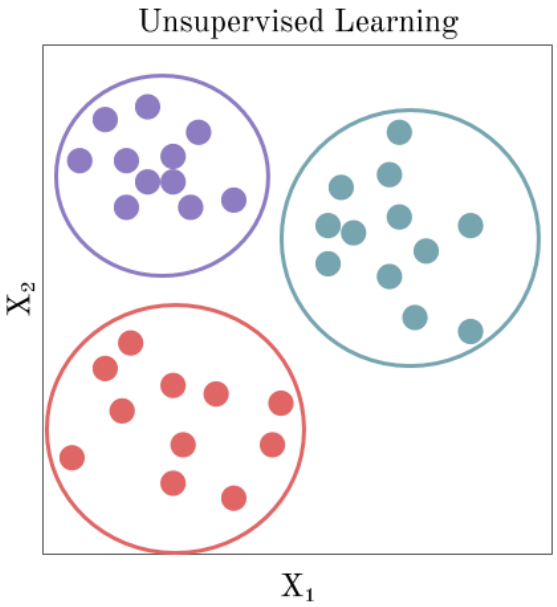
INVESTIC

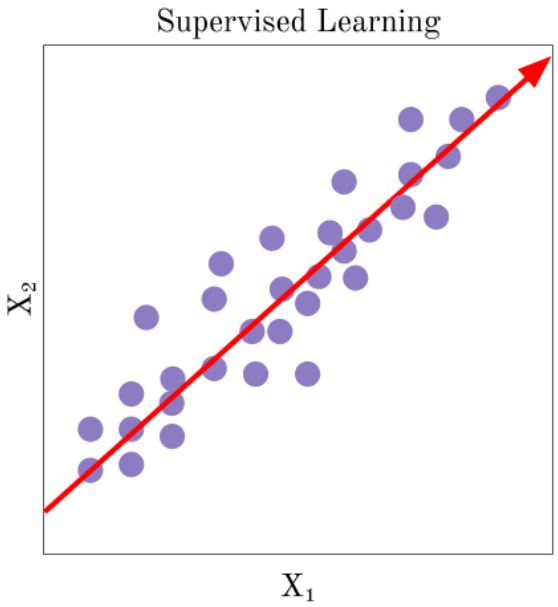
Classified by Type

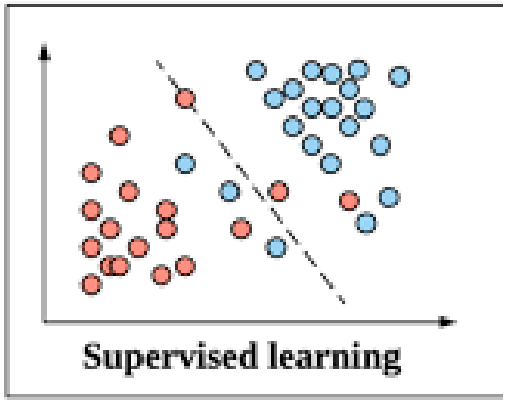


INVESTIC

Which is which

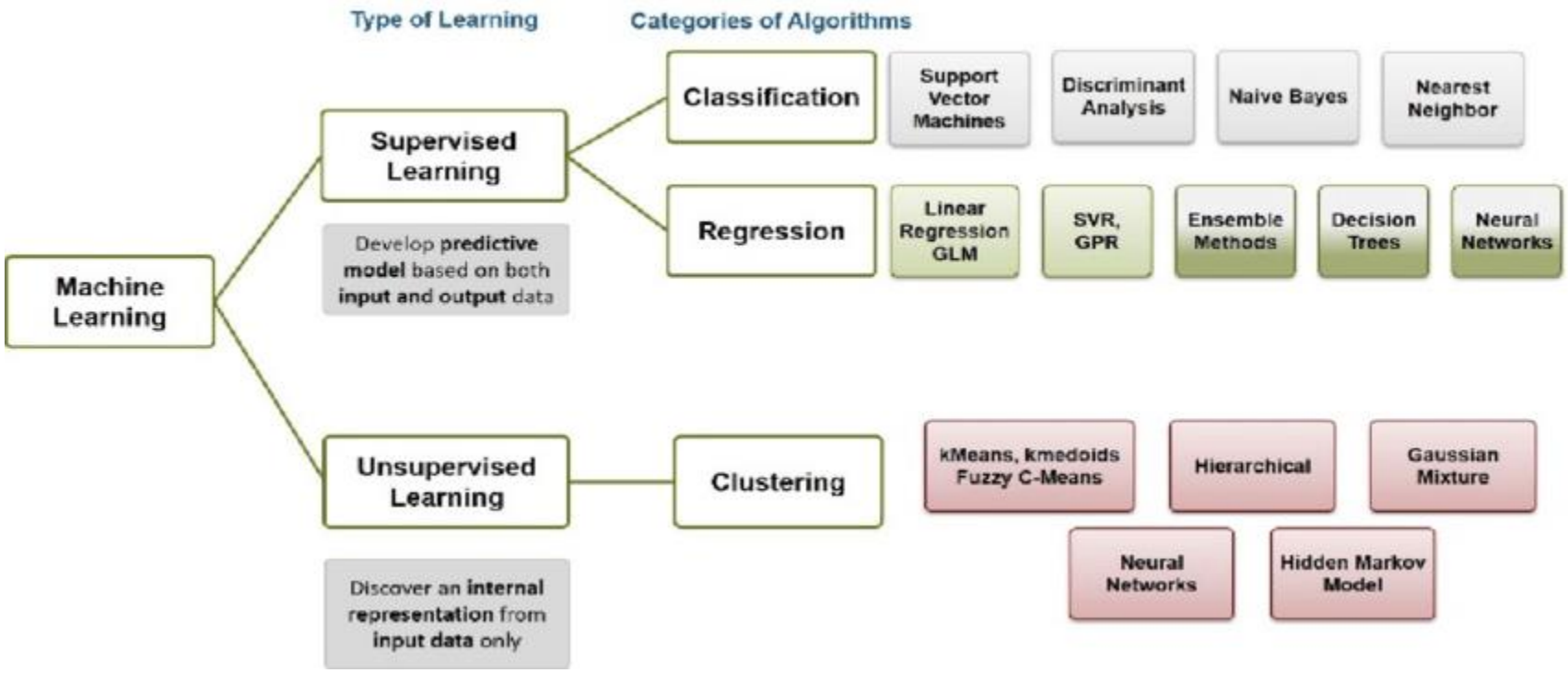






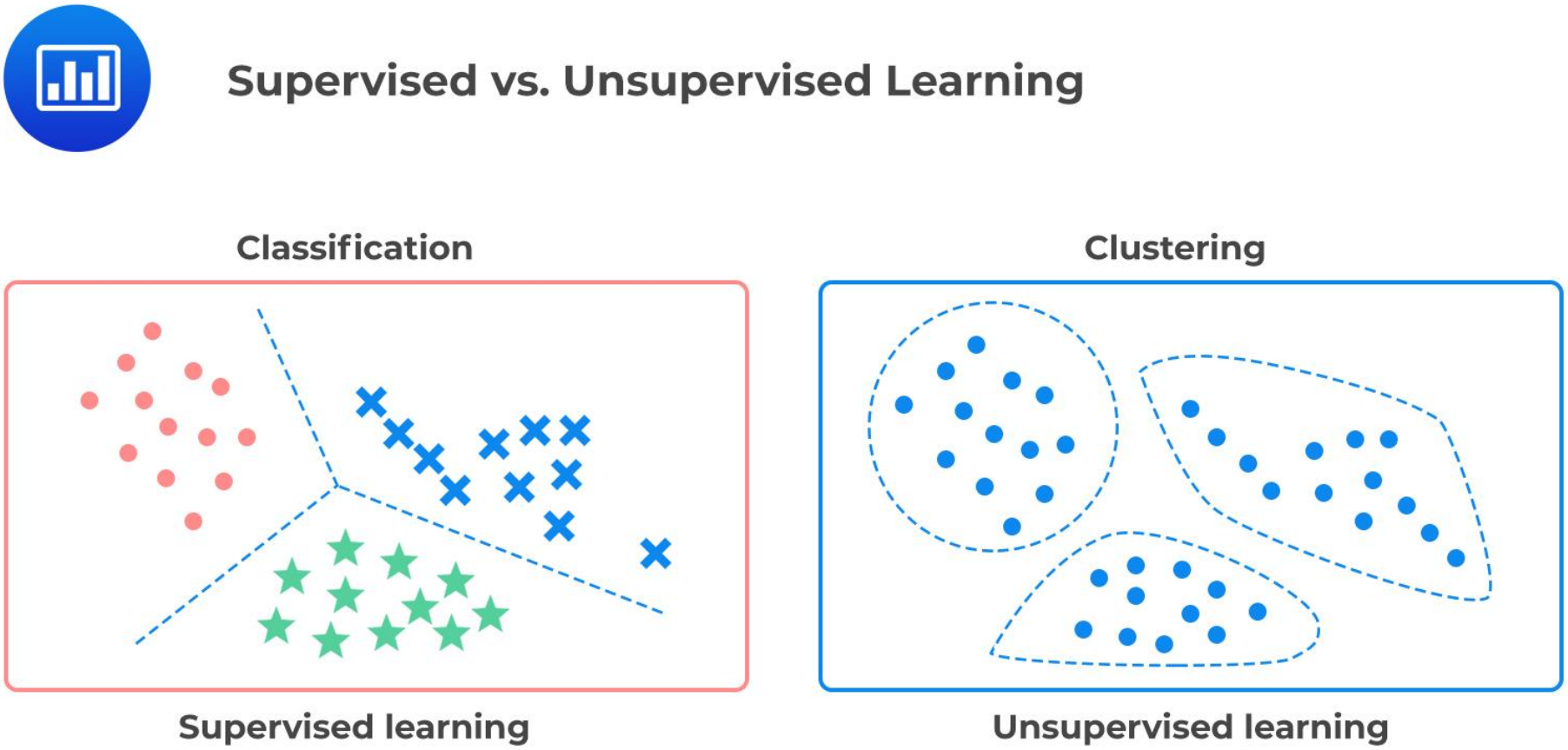
INVESTIC

Regression & Classification



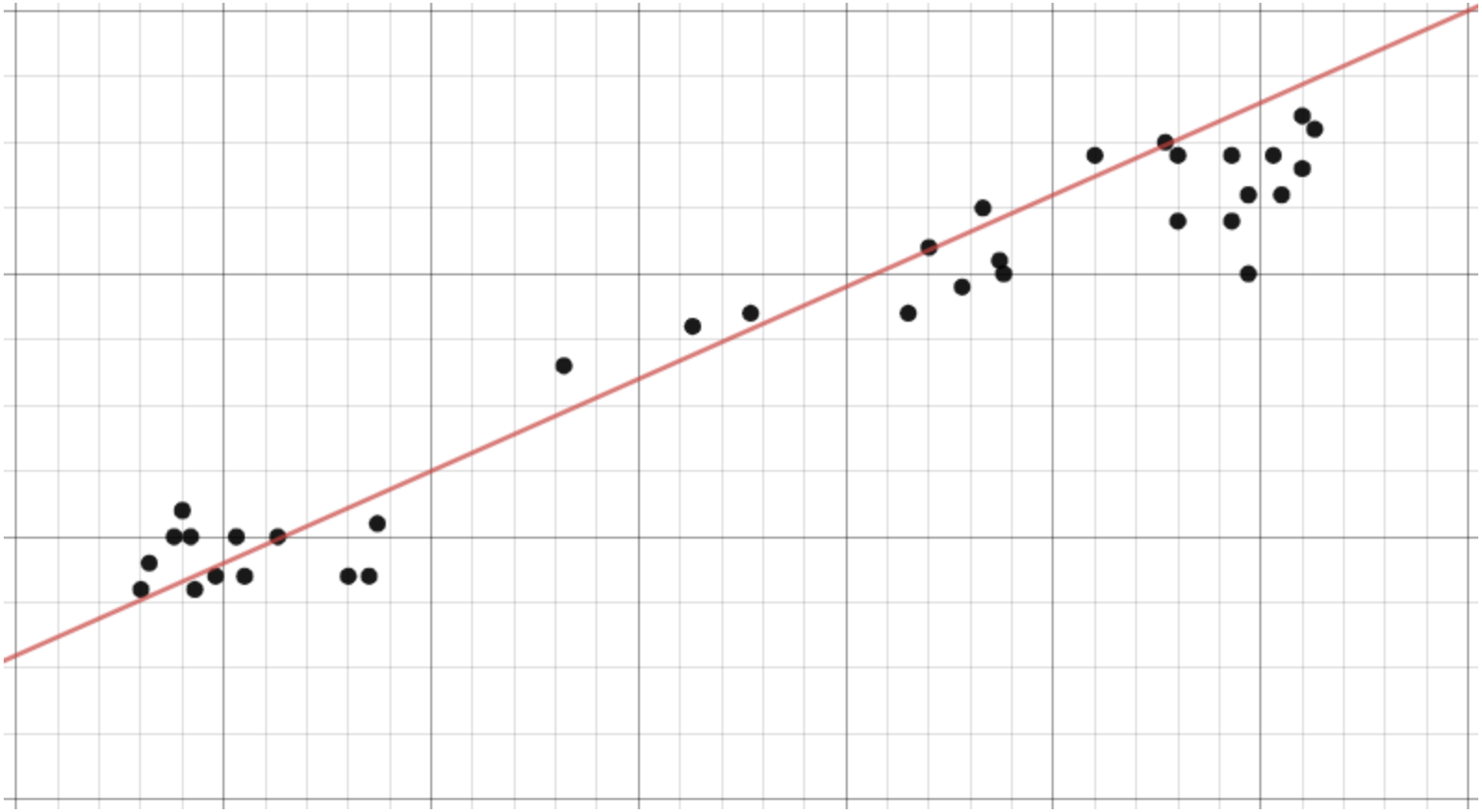
INVESTIC

Classification vs Clustering

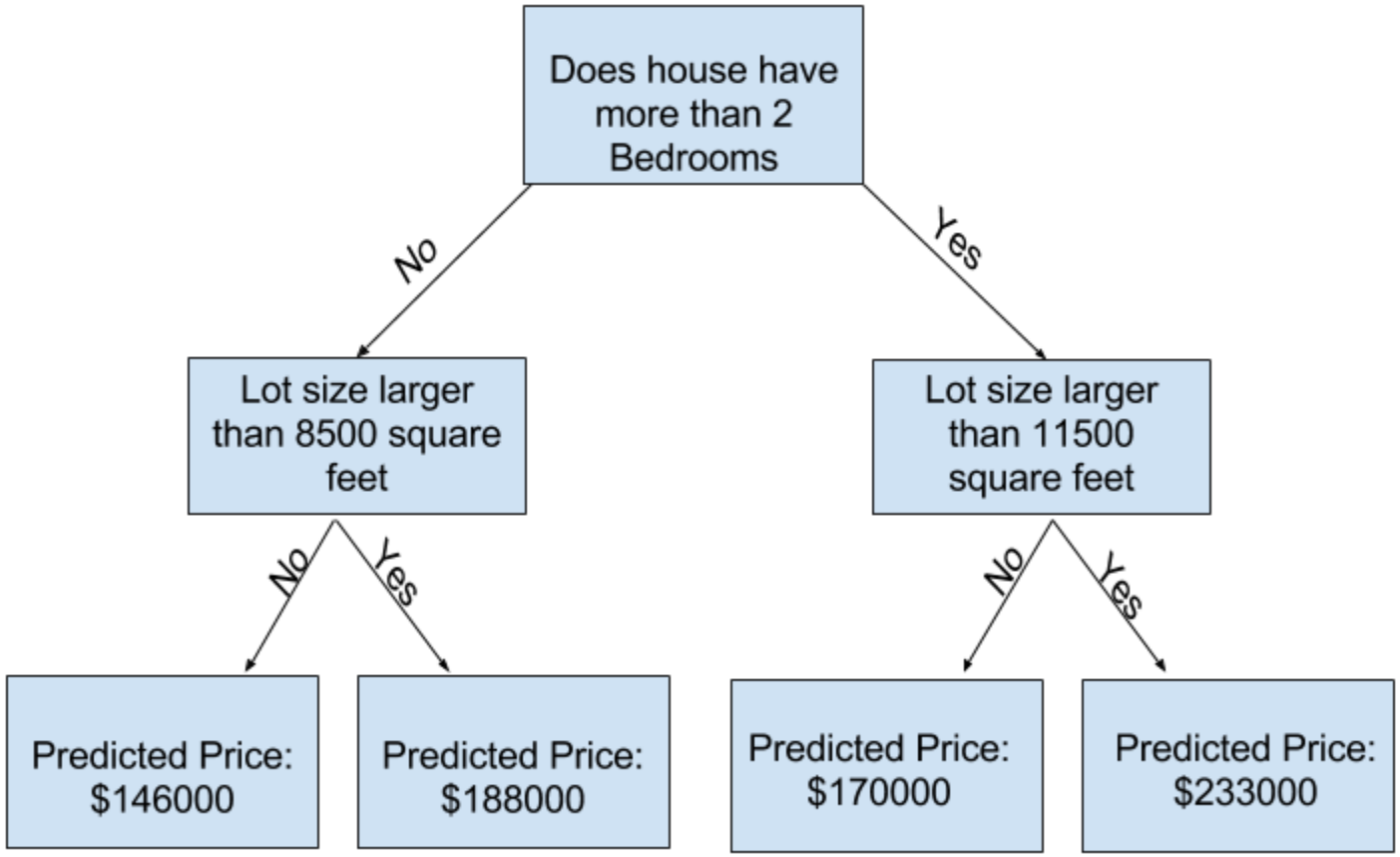


INVESTIC

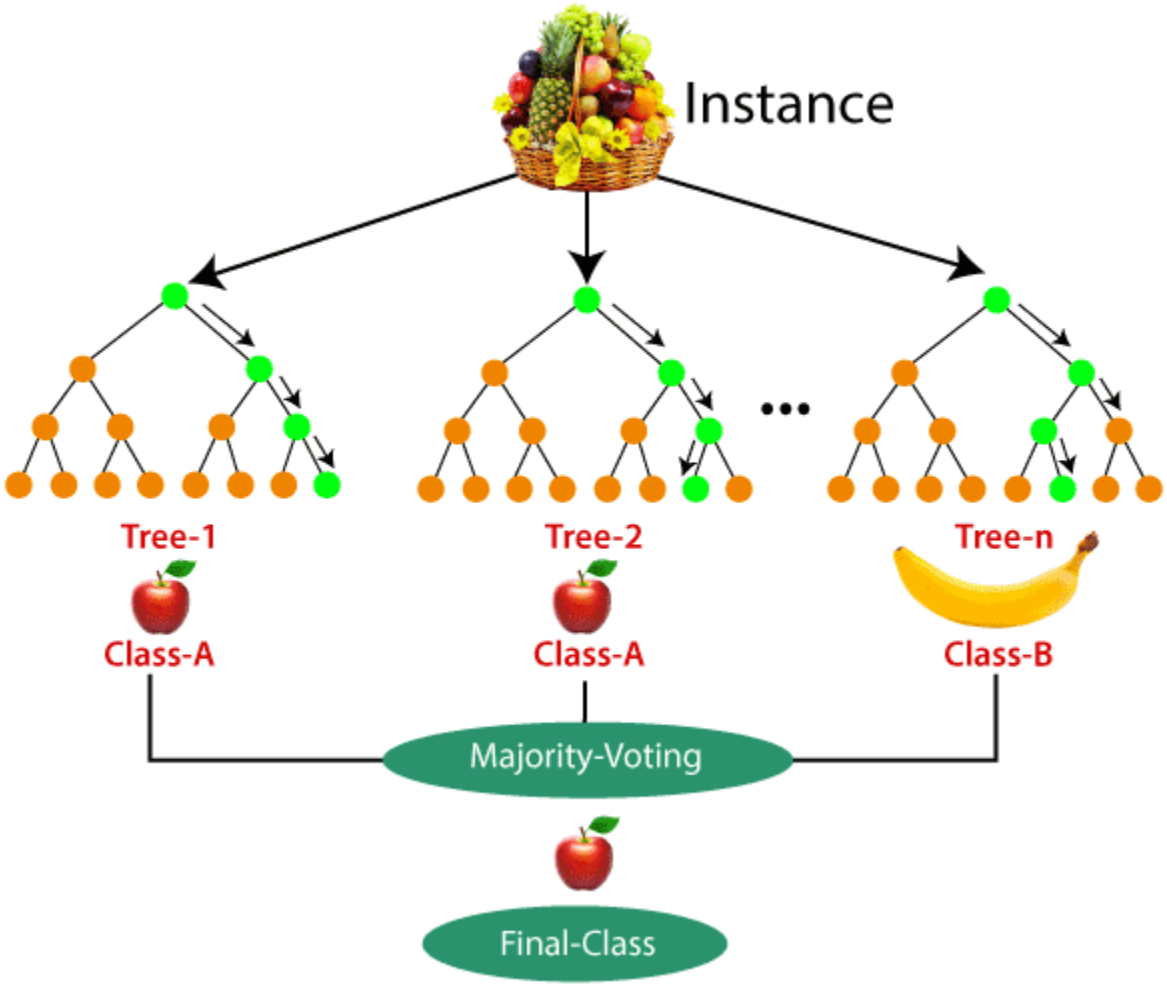
Regression Learning Model



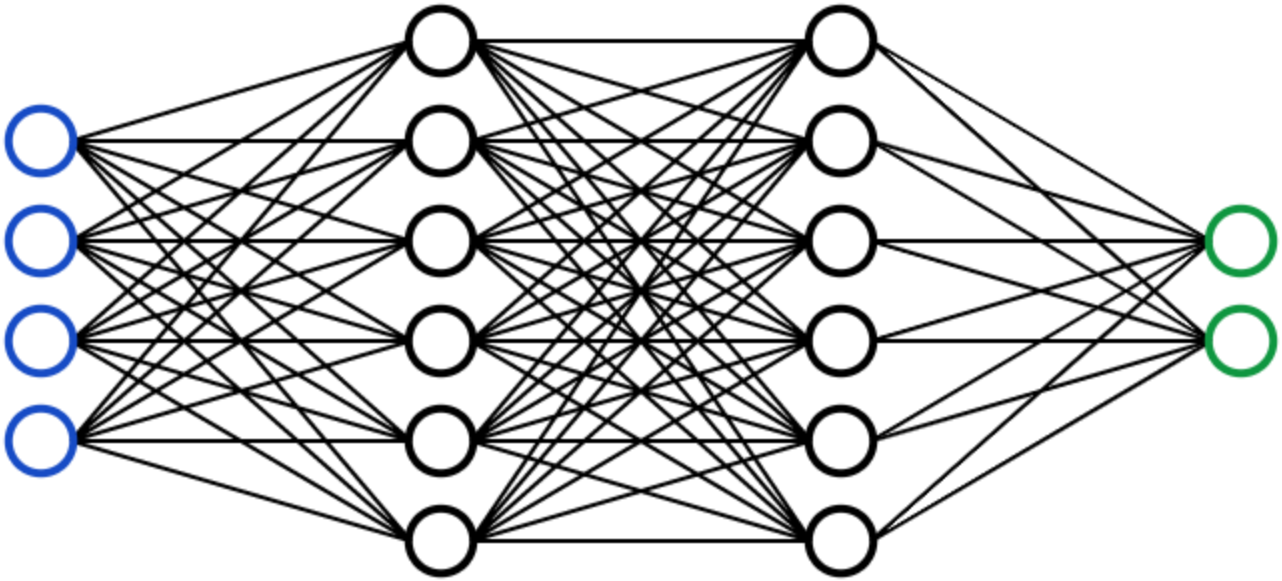
Regression Learning Model



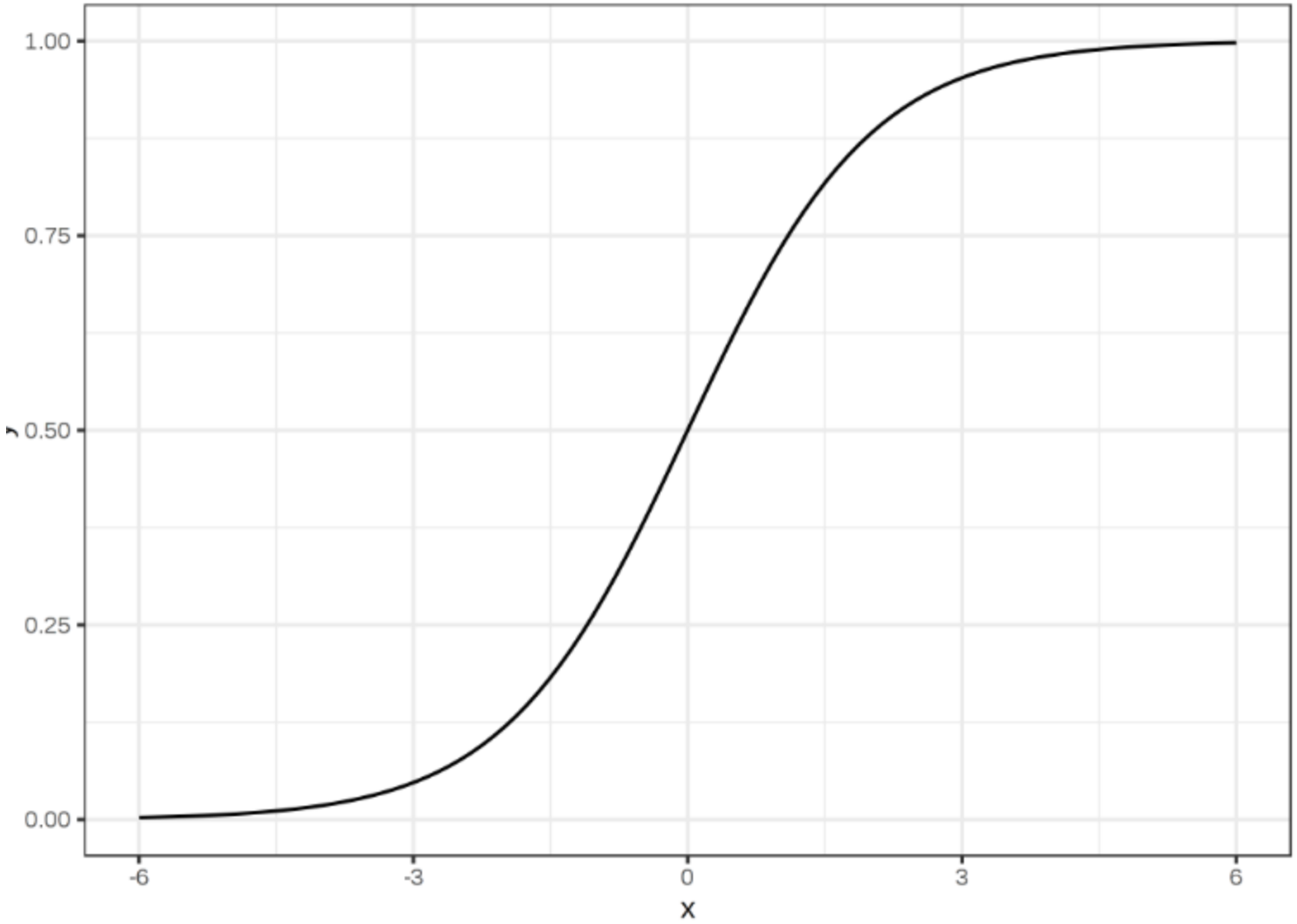
Regression Learning Model



Regression Learning Model

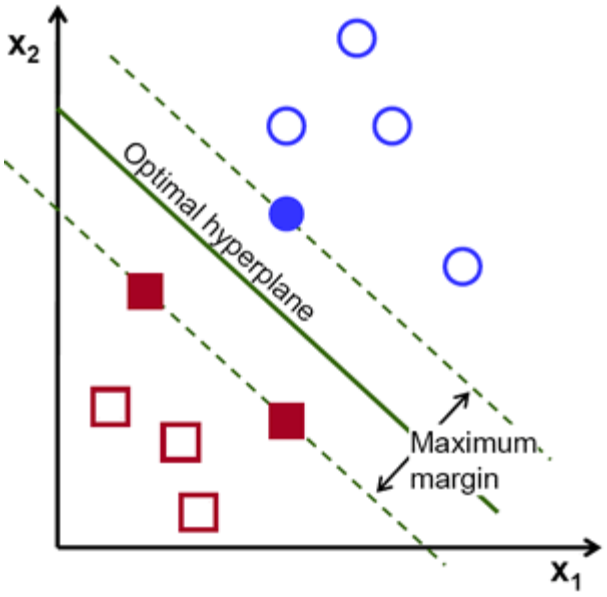


Classification Learning Model

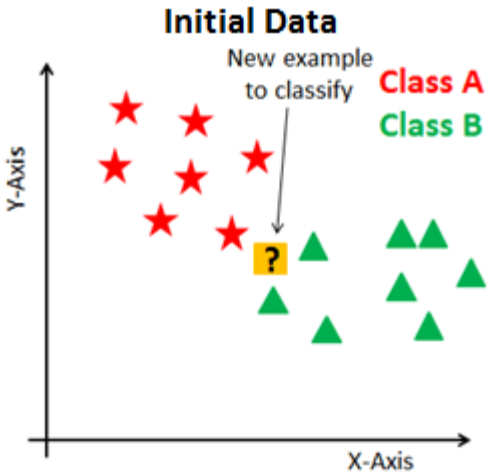


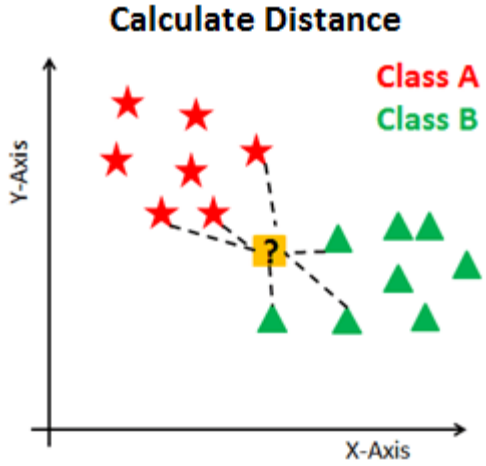
INVESTIC

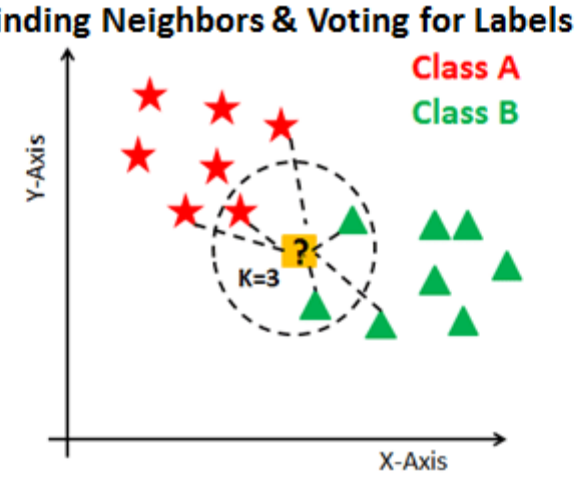
Classification Learning Model



Classification Learning Model







Classification Learning Model

Naive Bayes

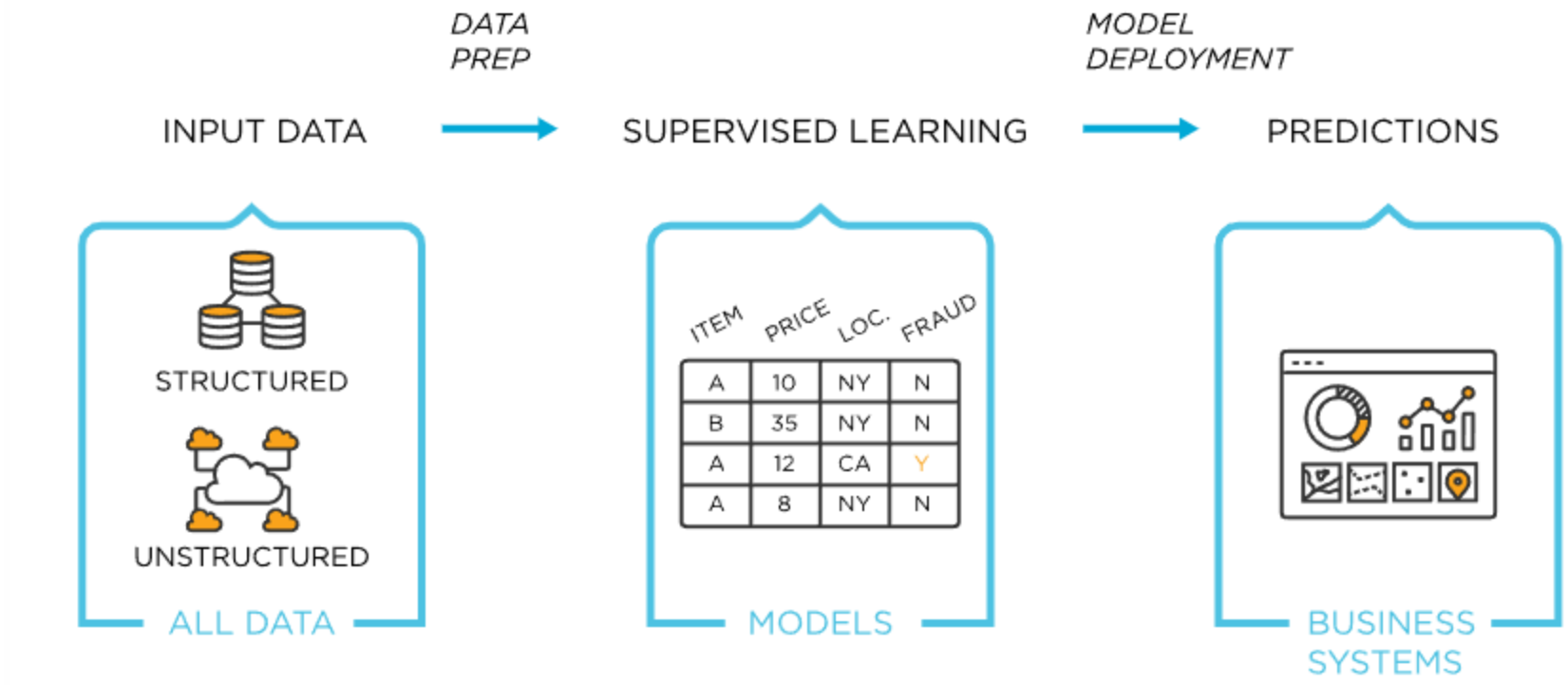
$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

“What is the probability of y given X?”

$$P(y|X) \propto P(X|y) * P(y)$$

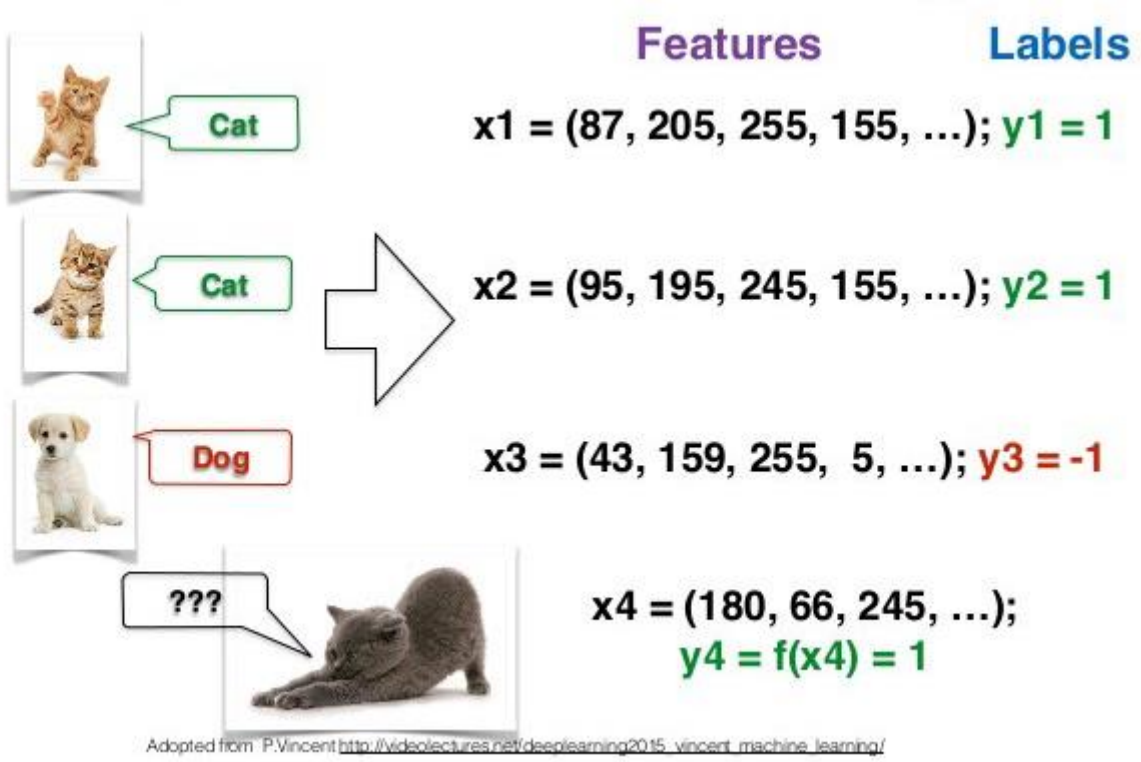
The goal is to find the class y with the maximum proportional probability.

Supervised Learning

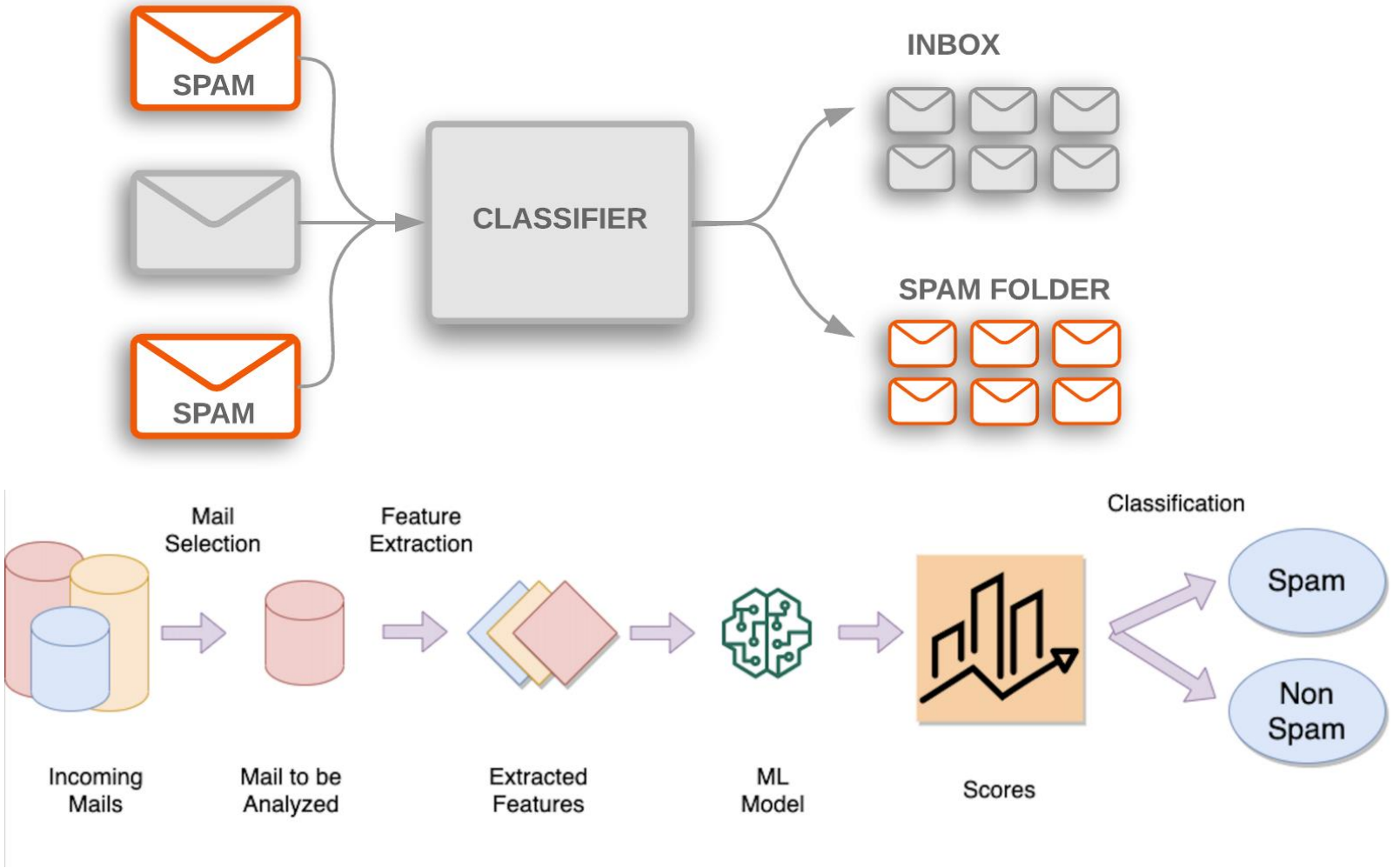


INVESTIC

Supervised Learning



Complete picture



INVESTIC

ML Process

1. Prepare & Transform Data
    1. Normalizing/standardizing
    2. Label Encoding
2. Modeling
3. Evaluating

INVESTIC

ML Process

1. Prepare & Transform Data
    1. Outlier
    2. Feature Scaling
        1. Normalization
        2. Standardization
    3. Label encoding/ One hot encoding

INVESTIC

ML Process Simple

1. Prepare & Transform Data
2. Model
3. Rumble !!
4. Backtest

INVESTIC

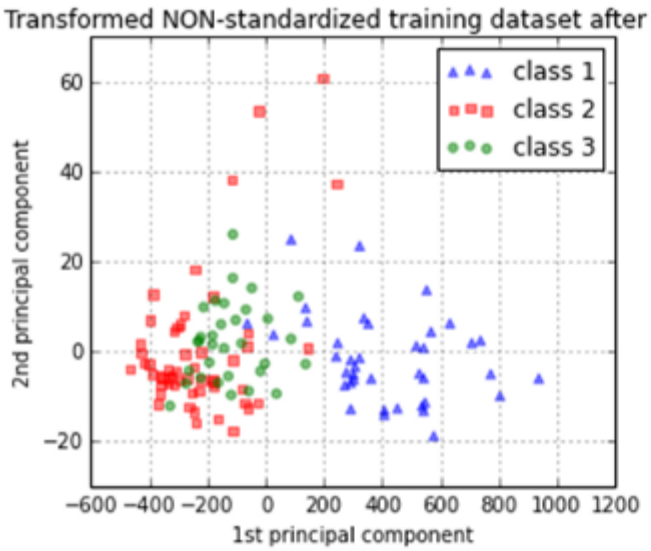
Example

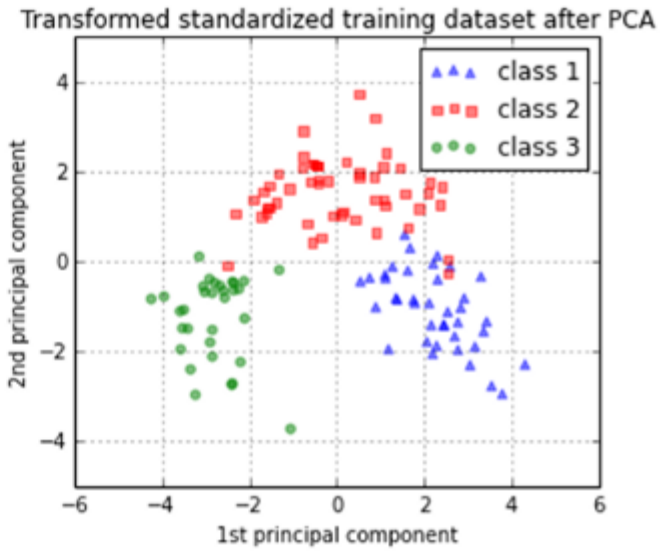
ML Process

1. Prepare & Transform Data
    1. Outlier
    2. Feature Scaling
        1. Normalization
        2. Standardization
    3. Label encoding/ One hot encoding
2. Split Train/ Test Data

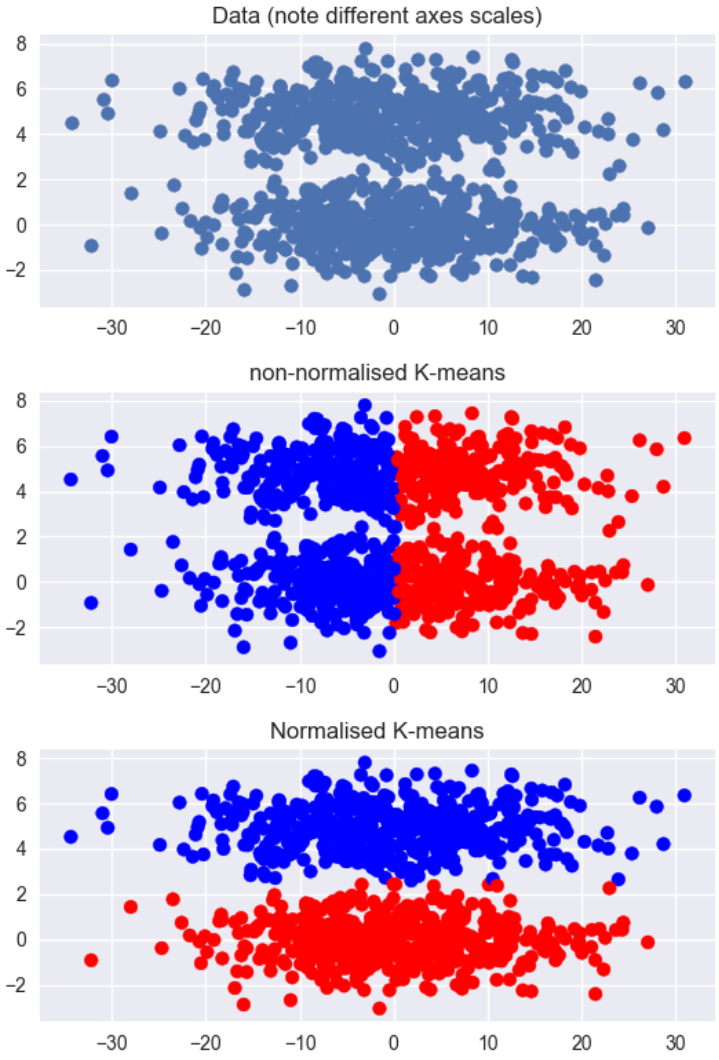
INVESTIC

Is feature scaling that matter??





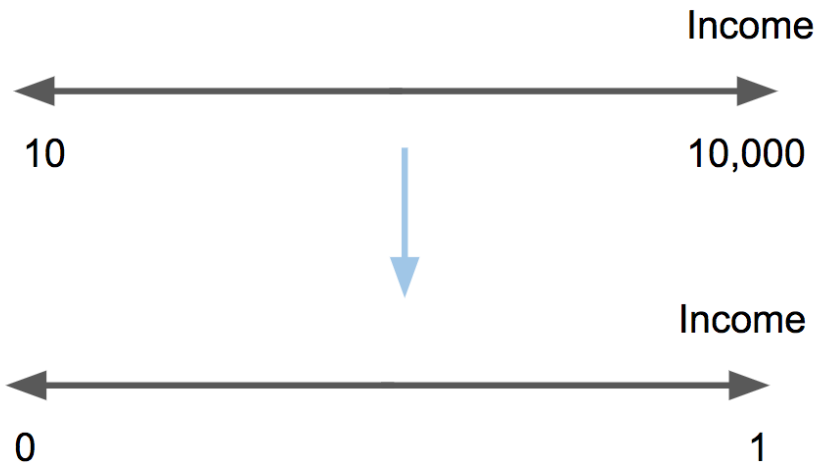
Is feature scaling that matter??



INVESTIC

Feature Scaling

Normalize



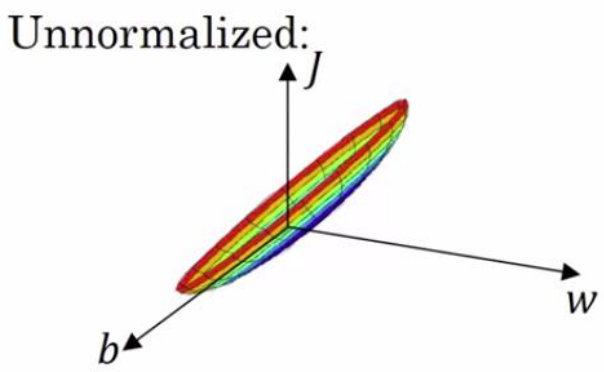
Standardize

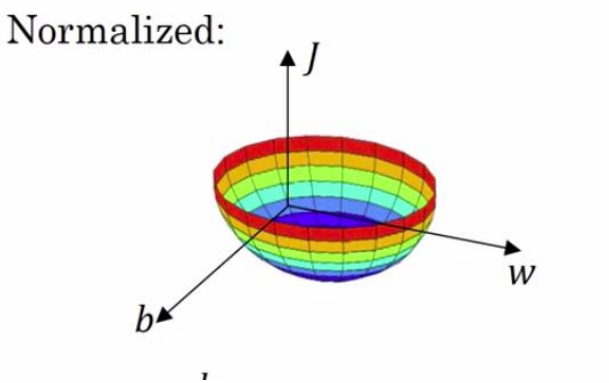
$$z = \frac{x - \mu}{\sigma}$$

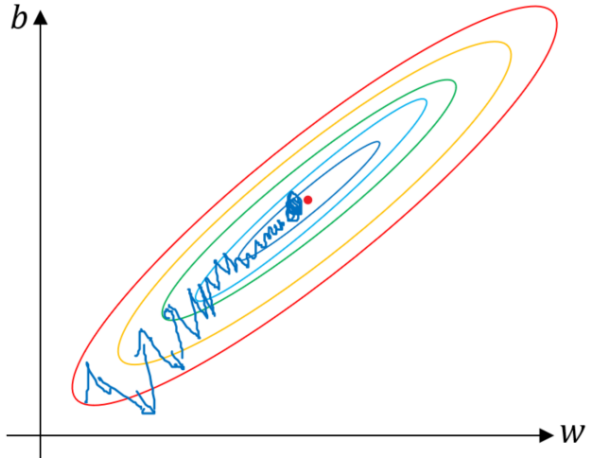
μ = Mean

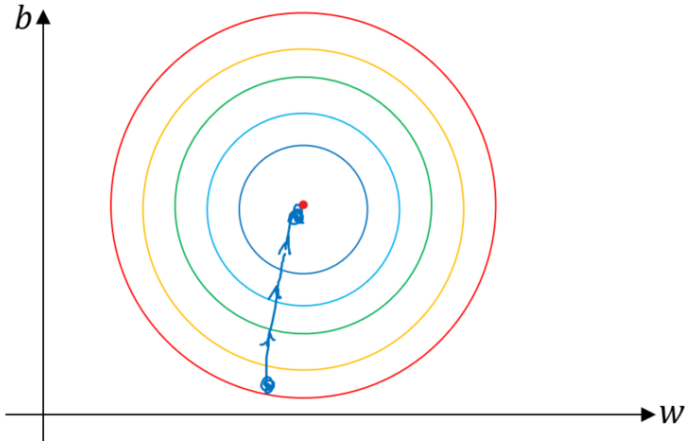
σ = Standard Deviation

Why Normalized



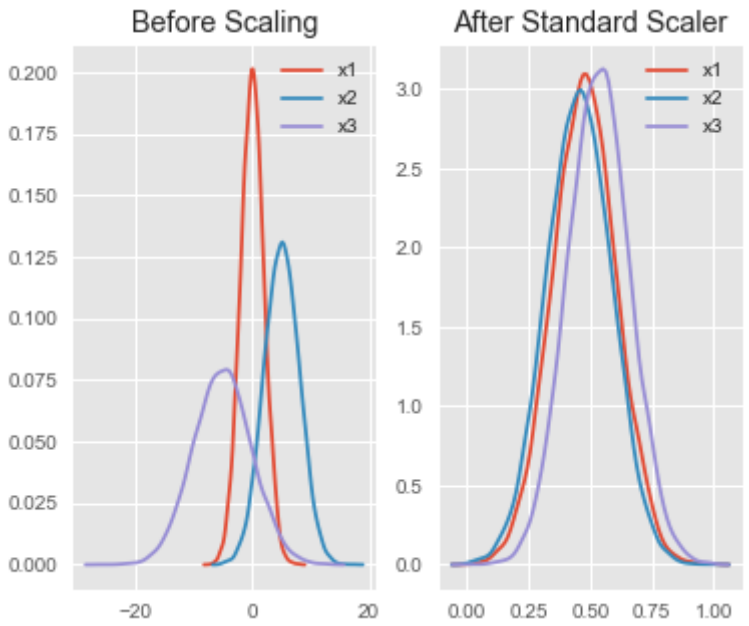






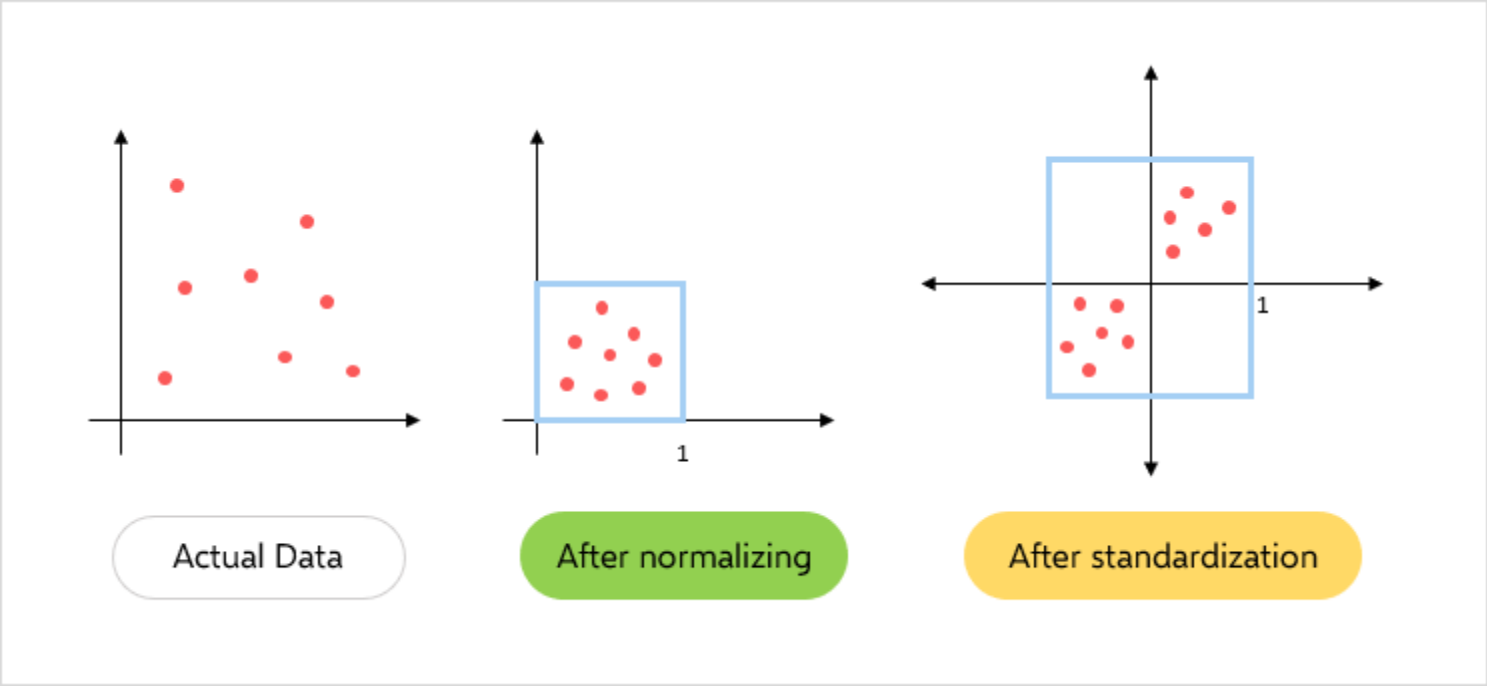
INVESTIC

Why Standardized
–



INVESTIC

Normalized vs Standardized

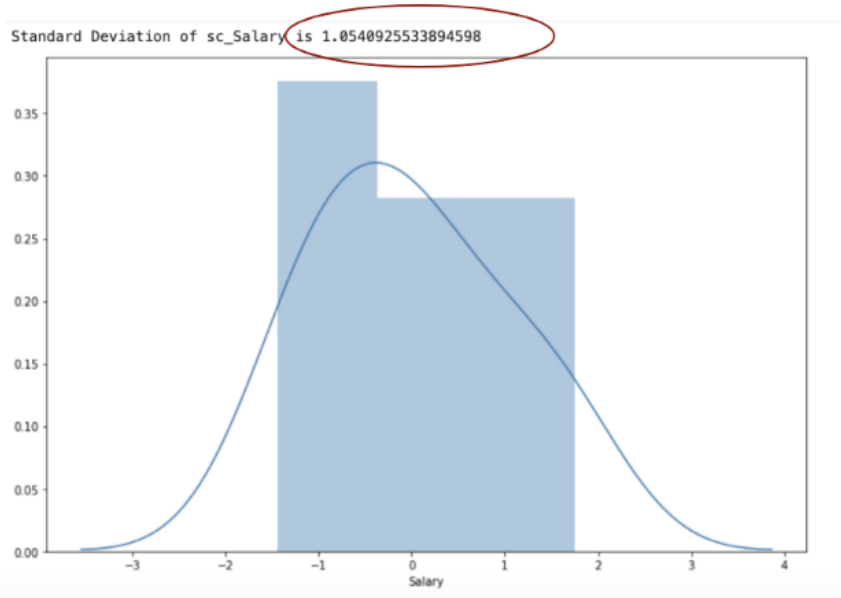


Normalized vs Standardized

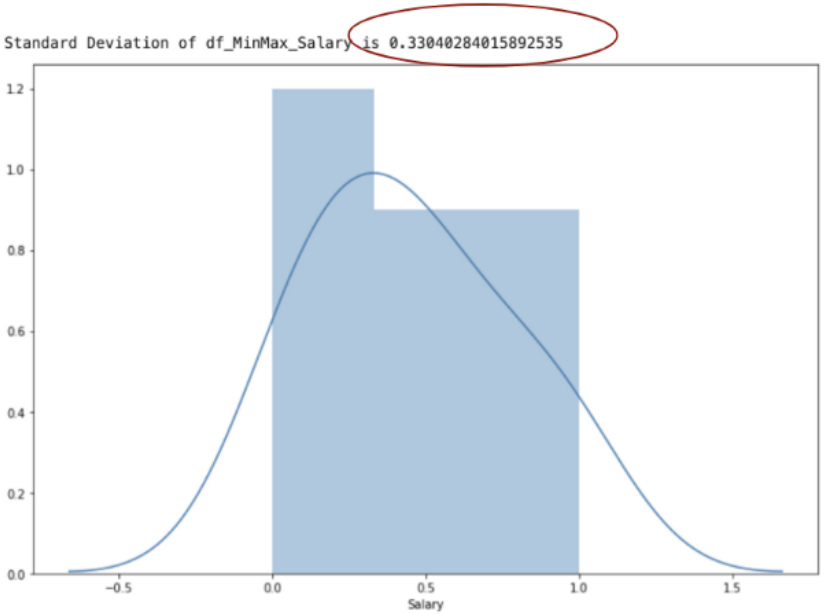
Column: Salary

Standard Deviation (Salary):
Max-Min Normalization (0.33) < Standardisation (1.05)

Standardisation



Max-Min Normalisation



INVESTIC

Sklearn Class

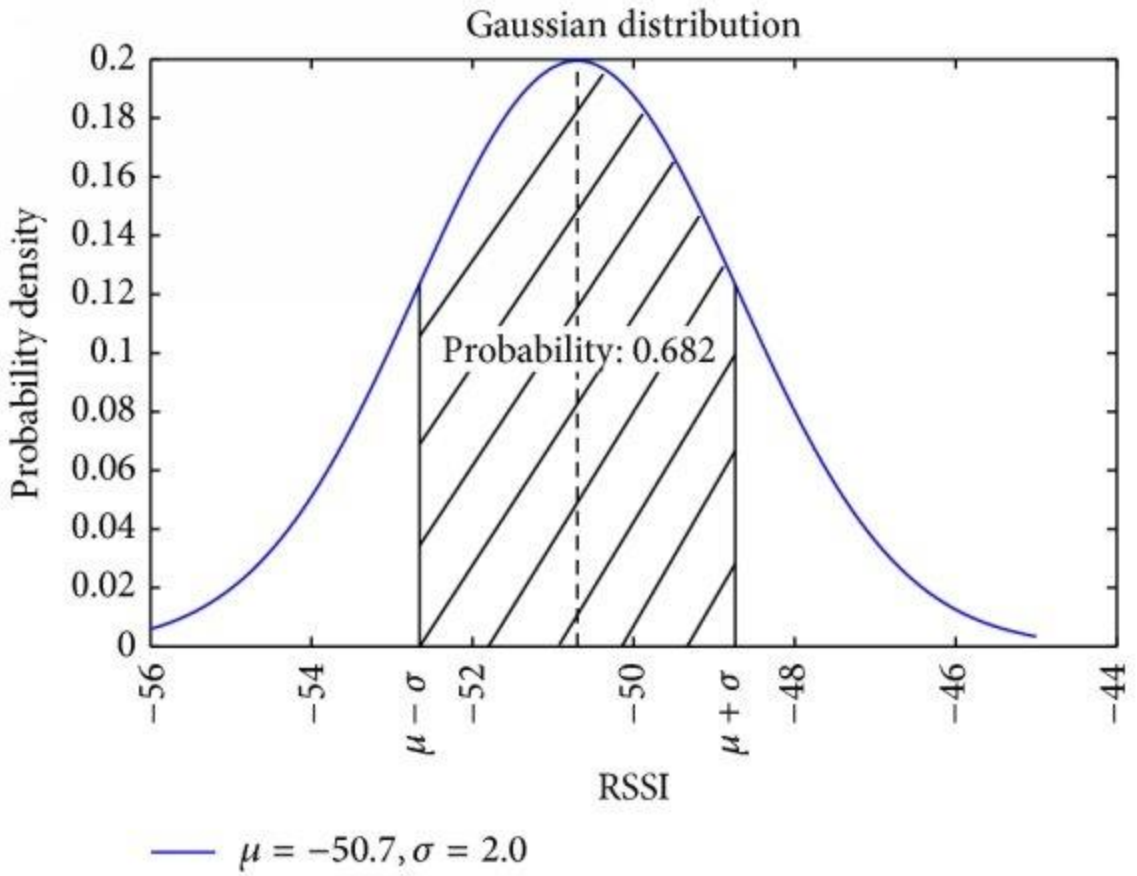
Name	Sklearn_Class
StandardScaler	StandardScaler
MinMaxScaler	MinMaxScaler
MaxAbsScaler	MaxAbsScaler
RobustScaler	RobustScaler
QuantileTransformer-Normal	QuantileTransformer(output_distribution='normal')
QuantileTransformer-Uniform	QuantileTransformer(output_distribution='uniform')
PowerTransformer-Yeo-Johnson	PowerTransformer(method='yeo-johnson')
Normalizer	Normalizer

Normalized vs Standardized

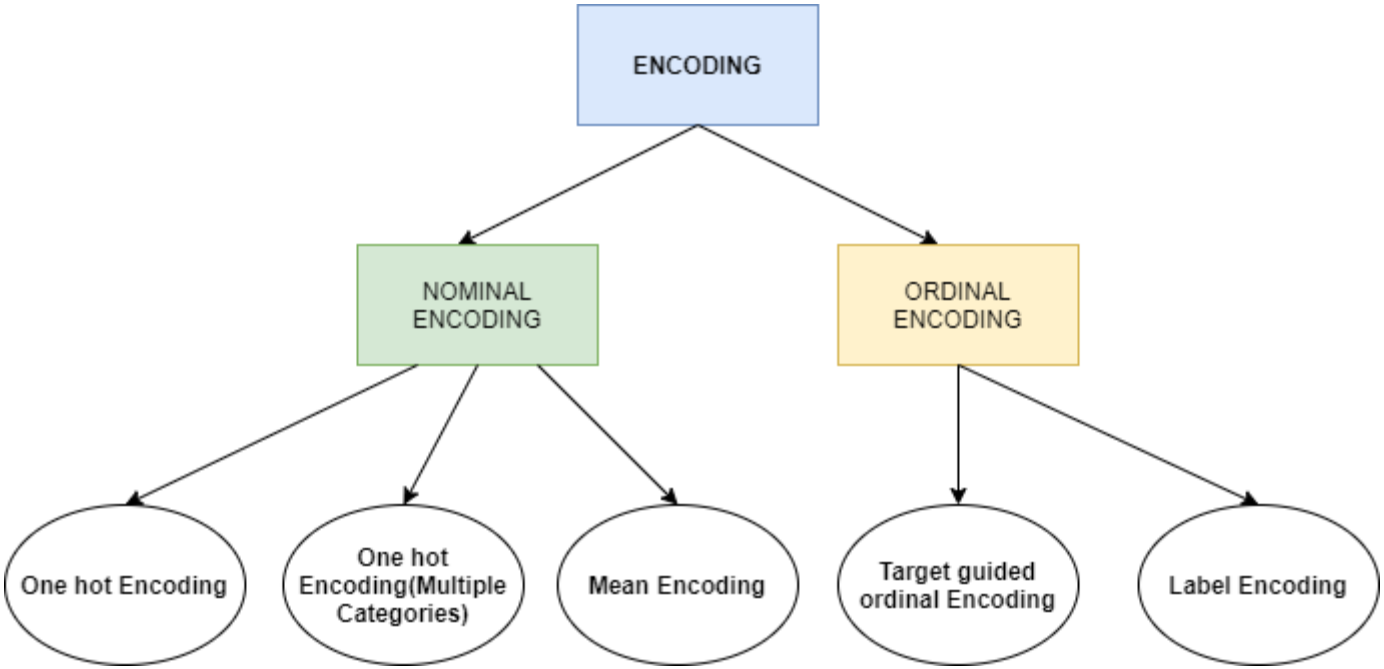
Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution

This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

Gaussian = Normal Distribution



Encoding


ENCODING
NOMINAL ENCODING
ORDINAL ENCODING
One hot Encoding
One hot Encoding(Multiple Categories)
Mean Encoding
Target guided ordinal Encoding
Label Encoding

Encoding

Nominal categorical variables are those for which we do not have to worry about the arrangement of the categories.

Male and Female.

Different states like NY, FL, NV, TX

Ordinal categories are those in which we have to worry about the rank. These categories can be rearranged based on ranks.

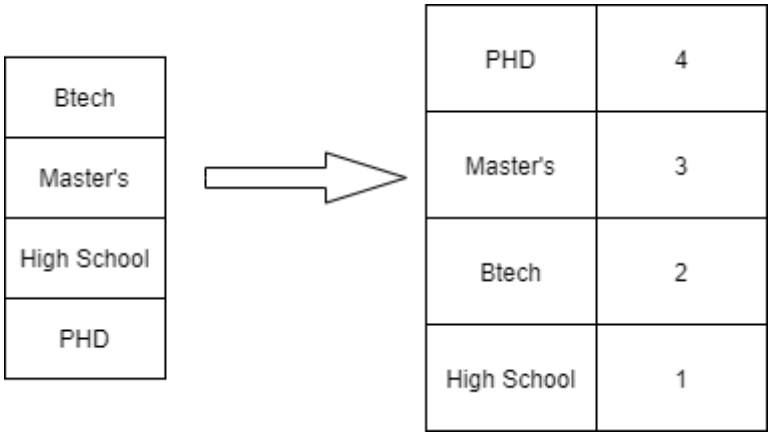
education level (PHD-1, masters-2, bachelors-3).

ML Process

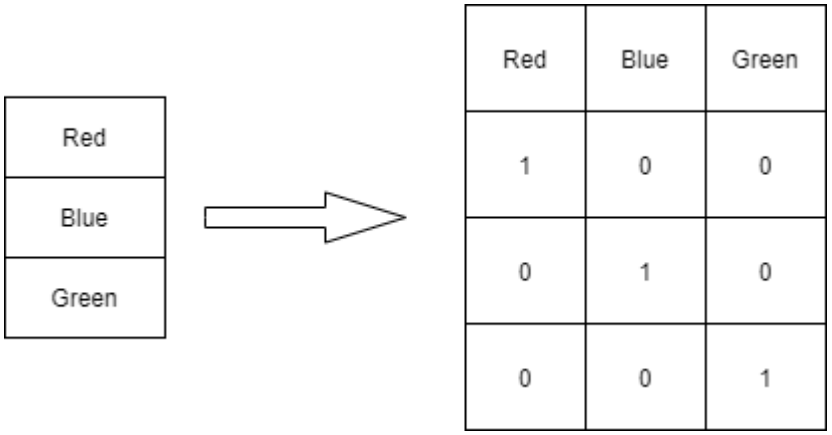
1. Prepare & Transform Data
    1. Outlier
    2. Feature Scaling
        1. Normalization
        2. Standardization
    3. Label encoding/ One hot encoding

INVESTIC

Label Encoding



One Hot Encoding



INVESTIC

Other Encoding

Target guided ordinal categories

Column	O/P
Btech	1
PHD	1
Master's	0
High School	0
PHD	1
Btech	0
Master's	0
High School	0
High School	1

Mean of Btech = 0.5

Mean of PHD = 1

Mean of Master's = 0

Mean of High School = 0.33

Column	Rank
Btech	2
PHD	1
Master's	4
High School	3
PHD	1
Btech	2
Master's	4
High School	3
High School	3

Mean Encoding

Pincode	O/P
753001	1
753002	1
753003	0
753001	0
753004	1
753002	0
753002	1
753001	0
753003	1

Column	Mean
753001	0.33
753002	0.66
753003	0.5
753004	1

INVESTIC

Other Encoding

Target guided ordinal categories

Column	O/P
Btech	1
PHD	1
Master's	0
High School	0
PHD	1
Btech	0
Master's	0
High School	0
High School	1

Mean of Btech = 0.5

Mean of PHD = 1

Mean of Master's = 0

Mean of High School = 0.33

Column	Rank
Btech	2
PHD	1
Master's	4
High School	3
PHD	1
Btech	2
Master's	4
High School	3
High School	3

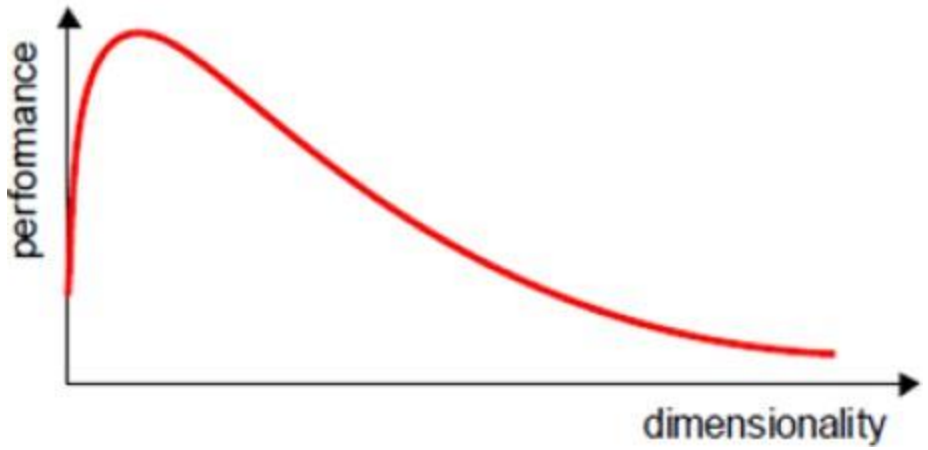
Mean Encoding

Pincode	O/P
753001	1
753002	1
753003	0
753001	0
753004	1
753002	0
753002	1
753001	0
753003	1

Column	Mean
753001	0.33
753002	0.66
753003	0.5
753004	1

INVESTIC

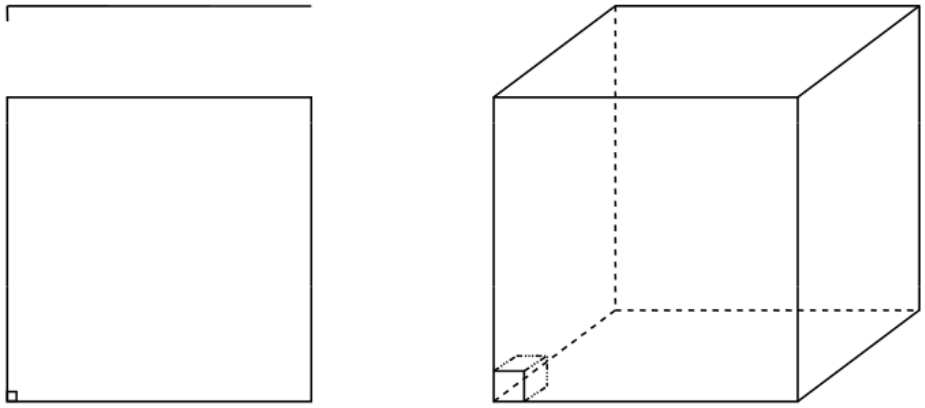
Curse of dimensity



INVESTIC

Curse of dimensity

- *k*NN breaks down in high-dimensional space
  - "Neighborhood" becomes very large.
- Assume 5000 points uniformly distributed in the unit hypercube and we want to apply 5-nn. Suppose our query point is at the origin.
  - In 1-dimension, we must go a distance of 5/5000 = 0.001 on the average to capture 5 nearest neighbors
  - In 2 dimensions, we must go $\sqrt{0.001}$ to get a square that contains 0.001 of the volume.
  - In d dimensions, we must go $(0.001)^{1/d}$



Source: mathematics stack exchange

Curse of dimensity

Curse of Dimensionality

- Low dimension ➔ good performance for nearest neighbor.
- As dataset grows, the nearest neighbors are near and carry similar labels.
- Curse of dimensionality: in high dimensions, almost all points are far away from each other.

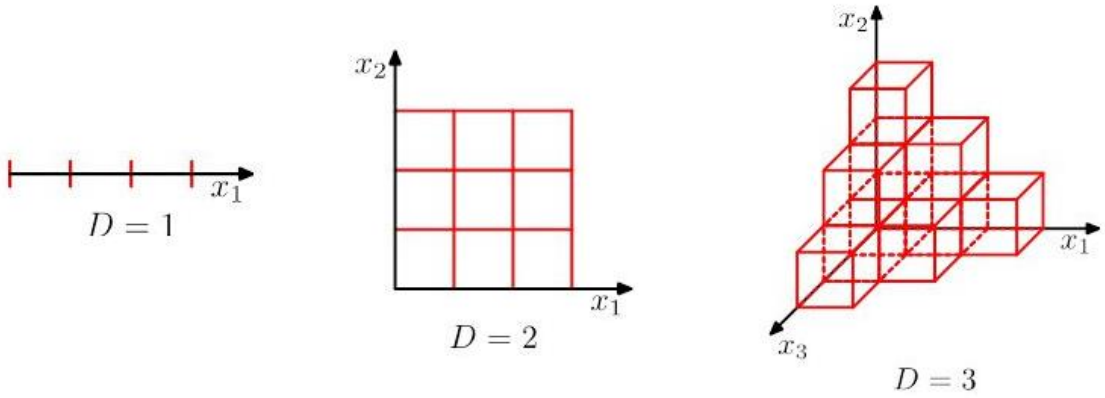


Figure Bishop 1.21