Machine Learning

Concepts

Machine Learning

Supervised

Unsupervised

Reinforcement

Supervised Learning

sepal_length	sepal_width	petal_length	petal_width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
5.5	2.3	4	1.3	Iris-versicolor
6.5	2.8	4.6	1.5	Iris-versicolor
5.8	2.8	5.1	2.4	Iris-virginica
6.4	3.2	5.3	2.3	Iris-virginica

Supervised Learning

What type of Iris plant is this?

attributes

sepal_length	sepal_width	petal_length	petal_width
6.5	3	5.5	1.8

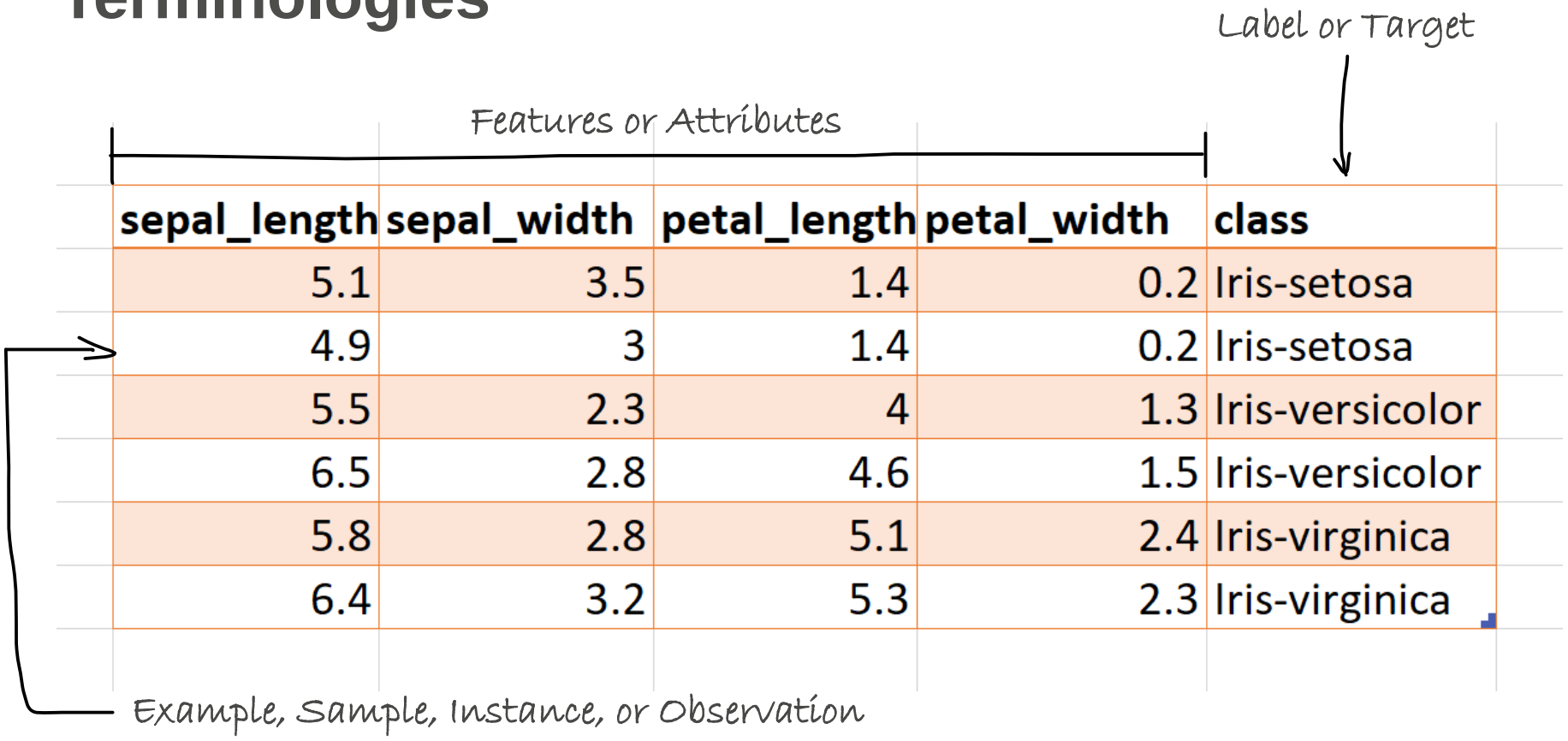
ML Predicted Answer: Iris-virginica

Model Comparison

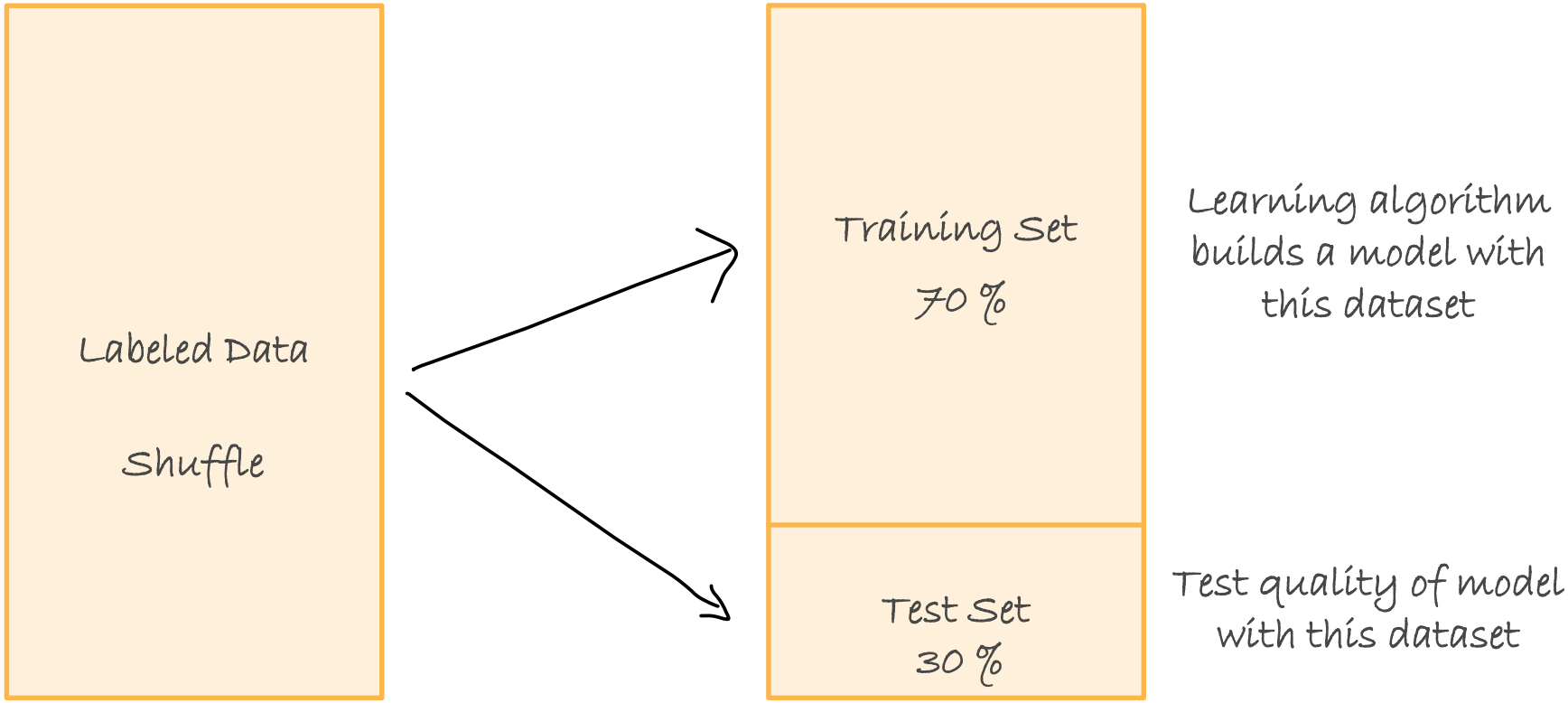
Is this email a spam?

Buy a new home with a low down payment. Our 30 year fixed mortgage can give you the flexibility you need to …

Terminologies



Terminologies



Supervised Algorithm Types

Regression

Binary Classification

Multi-Class Classification

Regression

Used for predicting numeric output

How much is my home worth?

How many passengers are going to travel by air this year?

Binary Classification

Used for predicting a binary output or two classes - We need a YES or NO answer

Is this email a spam?

Does this social media post require a follow-up?

Is this patient showing symptoms of a disease?

Multi-Class Classification

Used for predicting one out of several outcomes

How is the weather in NY tomorrow?
[Sunny, Windy, Cloudy, Rainy, Snow,…]

What ad should be displayed for this search?
[Sport, Real Estate, Home Loan, Auto, …]

Unsupervised Learning

Only data – There is no defined target

Group similar observations

Anomaly Detection

Words used in similar context

Unsupervised Algorithms

Clustering

Dimensionality Reduction

Group words that are used in similar context or have similar meaning

Reinforcement Learning

Decision Making under uncertainty

Autonomous Driving

Games

Reinforcement uses Reward Functions to reward correct decision and punish incorrect decision

Data Types

Data in Real Life

Numeric

Text

Categorical

Categorical

Day of Week	Sales
Sunday	100
Monday	50
Tues	20
Wed	30
Thursday	25
Fri	35
Sat	110

Categorical (Numeric Encoding)

Day of Week	Sales
1	100
2	50
3	20
4	30
5	25
6	35
7	110

Sunday=1, Monday=2, Tuesday=3, …

Categorical – One Hot Encoding

Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sales
1	0	0	0	0	0	0	100
0	1	0	0	0	0	0	50
0	0	1	0	0	0	0	20
0	0	0	1	0	0	0	30
0	0	0	0	1	0	0	25
0	0	0	0	0	1	0	35
0	0	0	0	0	0	1	110

Categorical

Cartesian Transformation - Combine categorical features to form new features

Day of Week	Weather	Sales
Sunday	Sunny	100
Monday	Cloudy	50
Tues	Snow	20
Wed	Snow	30
Thursday	Snow	25
Fri	Rain	35
Sat	Sunny	110

Cartesian Transformation

Day of Week_Weather	Sales
Sunday_Sunny	100
Sunday_Cloudy	75
Sunday_Snow	15
Sunday_Rain	25

Text Data

Movie	Genre
Star Wars	Adventure
Notting Hill	Romance
Star Trek	Adventure

Star	Wars	Notting	Hill	Trek	Genre
1	1	0	0	0	Adventure
0	0	1	1	0	Romance
1	0	0	0	1	Adventure

Text Data

NGRAM Transformation

Orthogonal Sparse Bigram (OSB) Transformation

Lowercase Transformation

Remove Punctuation Transformation

Cartesian Transformation

Words Alter Meaning

“this is working. not disappointed”

“this is not working. disappointed”

After tokenization:

['disappointed’, 'is’, 'not’, 'this’, 'working']

NGRAM

“this is working. not disappointed”

“this is not working. disappointed”

After NGRAM Transformation (window = 2):

['this is', 'is working', 'working not', 'not disappointed’]
['this is', 'is not', 'not working', 'working disappointed']

Stemming

All these words are treated differently:

['working', 'worked', 'works']

After stemming - words have same root

['work', 'work', 'work']

Lower Case

How is the request rate LIMIT Determined?

HOW is the request rate limit determined?

After Lower Case Transformation:

how is the request rate limit determined?

Numeric Data

Numeric value as-is (for linear relationship)

Normalization Transformation (for linear relationship)

Binning Transformation - convert to categorical (for non-linear relationship)

Handling Missing Values

Impute - If there is a reasonable way to fill-in the missing-values, you should do it!

Drop missing features or observations

Knowledge about data and context is important!

References: Missing Values

Handling Missing Values in Time Series - https://www.kaggle.com/juejuewang/handle-missing-values-in-time-series-for-beginners

Working with missing data - https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html

Substitute attribute – Boolean attribute that is set to 1 to indicate if another attribute has missing value: Treatment of missing values