

ECS 171: Machine Learning

Summer 2023

Edwin Solares easolares@ucdavis.edu

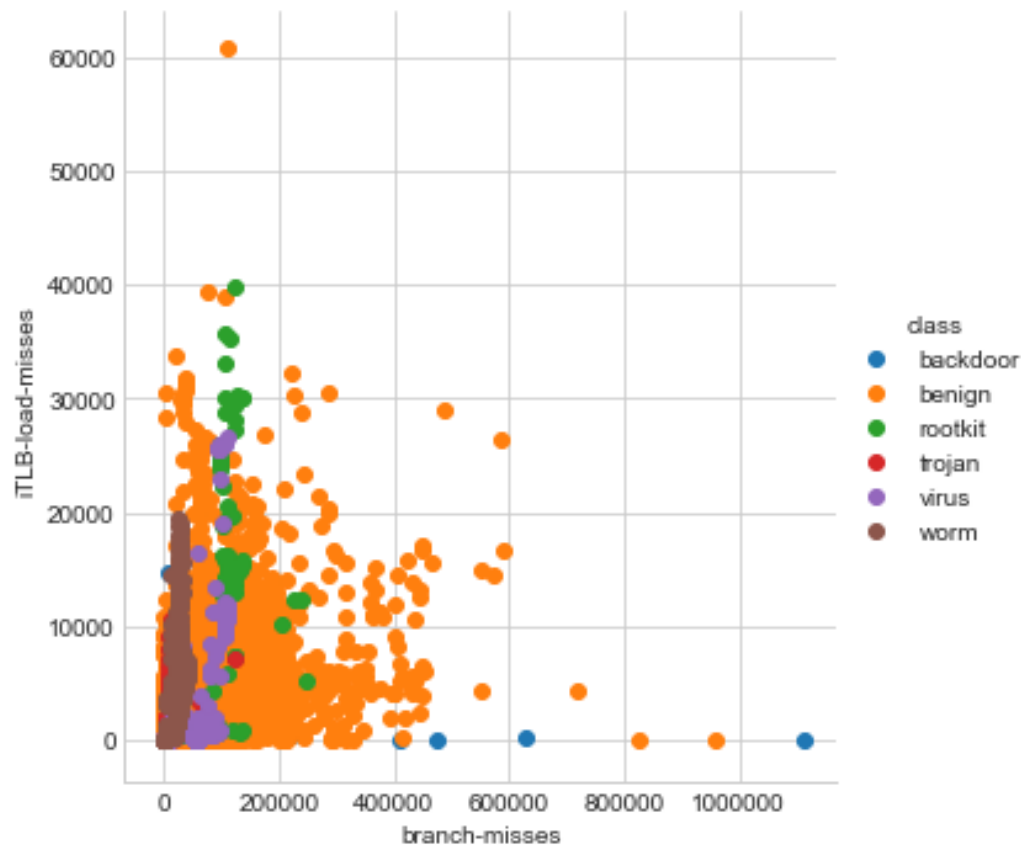
Intro to ML

What is Machine Learning

Subfield of AI

Enables systems to derive meaning from huge volume of data

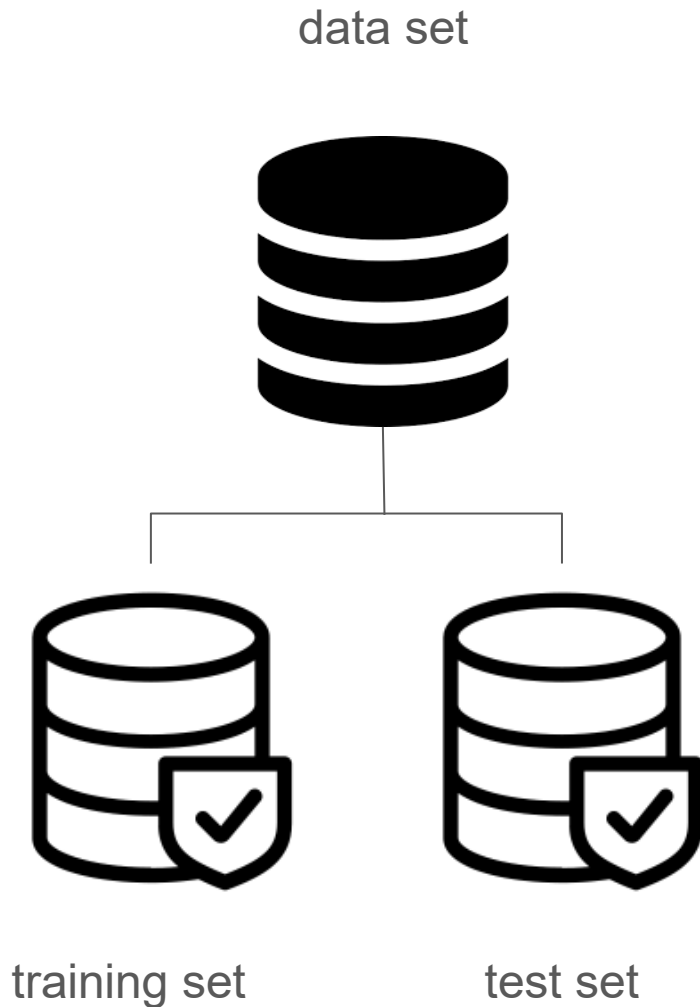
Learning from the data



Machine Learning Definition

Given a collection of observations (training set)

- Each **observation** contains a set of **attributes**, where **one** of the attributes is the **class**.
- Task: find a **model** for class attribute as some **function** of the values of other attributes.
- The **Goal** is to assign a **class** to previously **unseen records**, as **accurately** as possible.
- For this, we need a **test** set to evaluate the **accuracy** and **robustness** of the model.



Machine Learning Evaluation Metrics

TP, TN, FP, FN (True +, True -, False +, False -)

Precision and Recall

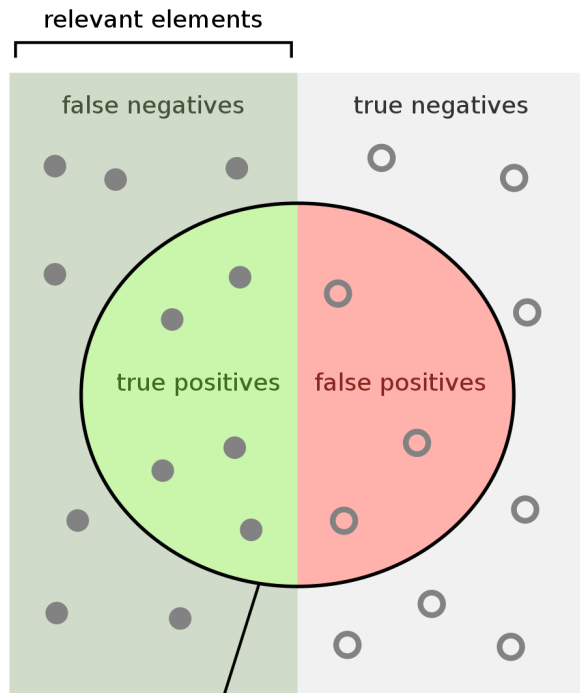
Receiver operating characteristic (ROC) curve and Area under curve (AUC)

Accuracy

F1 Score

- <https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative>
- <https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall>
- <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>
- <https://developers.google.com/machine-learning/crash-course/classification/accuracy>
- <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>

Machine Learning Evaluation Metrics (wiki)



Dogs

retrieved elements

Cats & Donkeys

How many retrieved items are relevant?

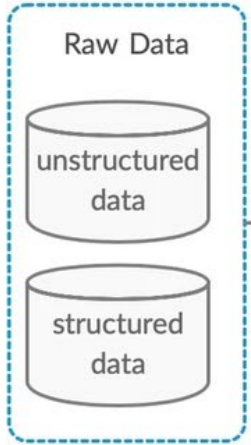
$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are retrieved?

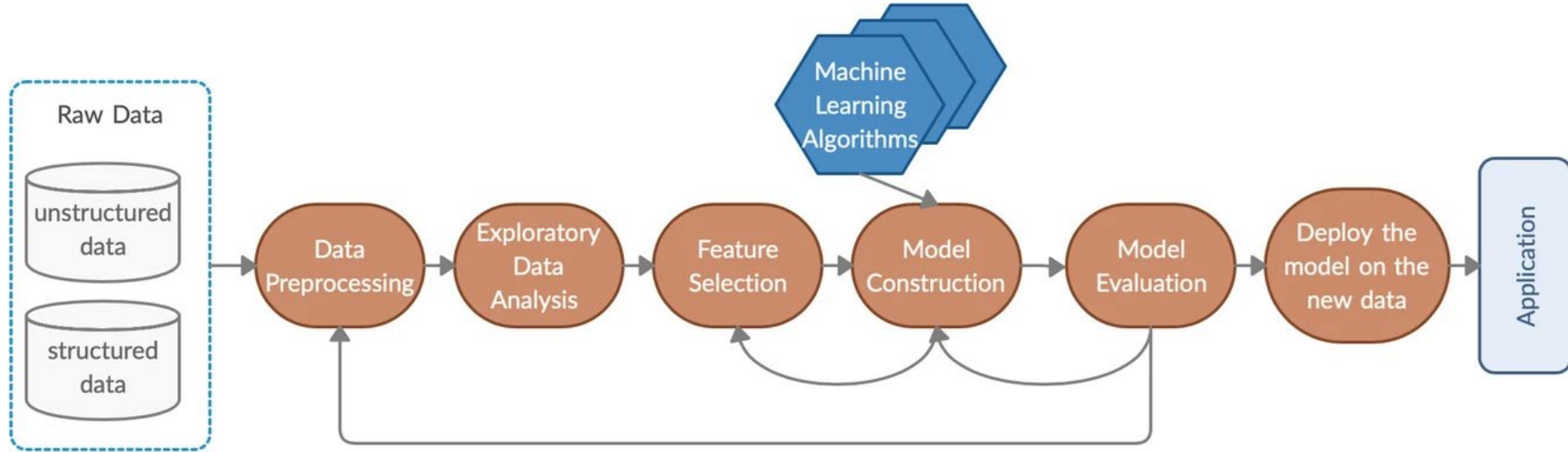
$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Machine Learning Pipeline



Machine Learning Pipeline



Example Tasks

Classification



Human Learning:

We learn through



Examples

Long Ear Black nose
↓ ↓
dog

Diagrams

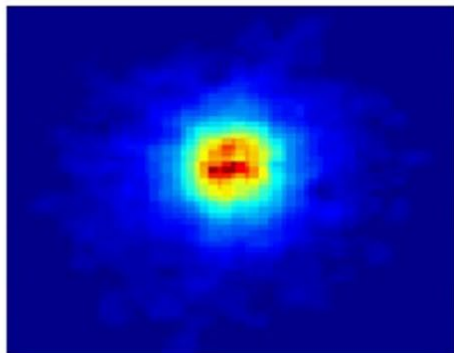


Comparisons

@CMU ML Blog

Classification Example

Early



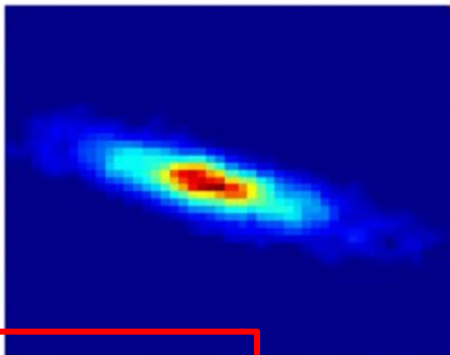
Class:

- Stages of Formation

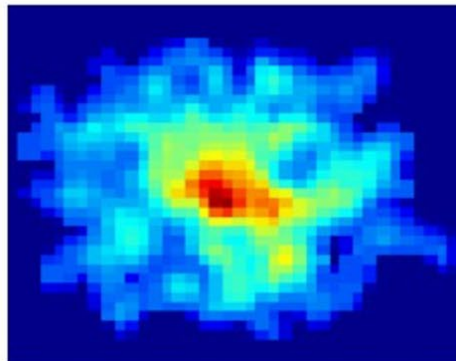
Attributes:

- Image features,
- Characteristics of light waves received, etc.

Intermediate



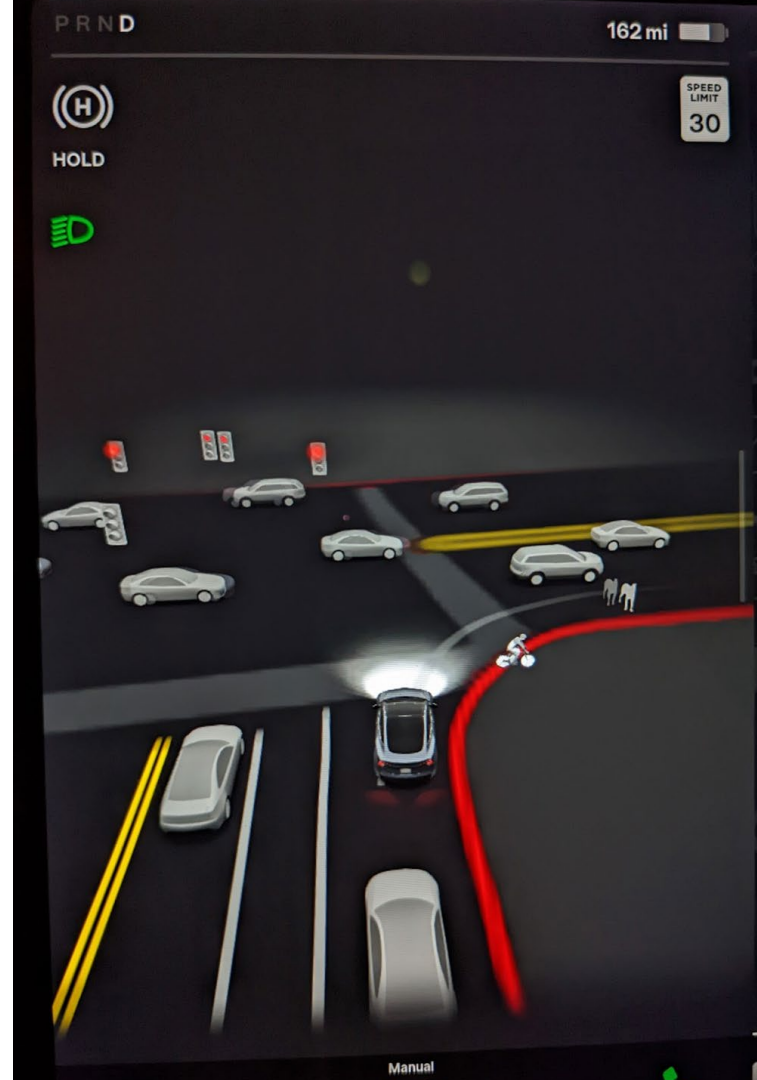
Late



Data Size:

- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Classification Example



Machine Learning Definition

Supervised Machine Learning

- Labeled data set
- Good for prediction
- Example: Classification

Unsupervised Machine Learning

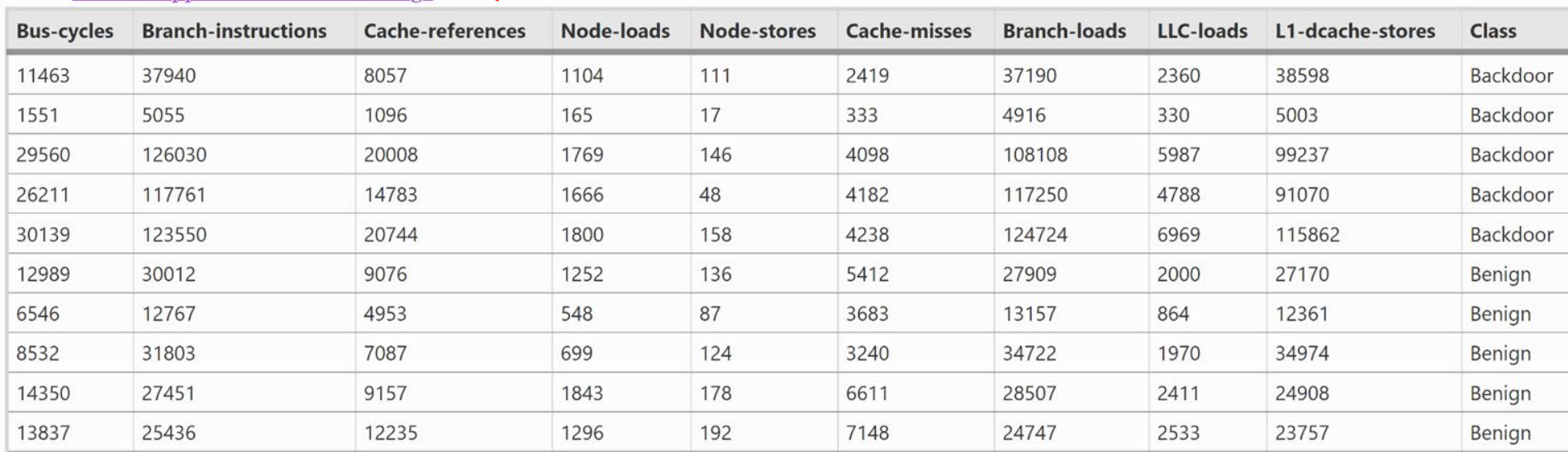
- Unlabeled data set
- Good for data exploration and association rule discovery
- Example: Clustering

Reinforcement Learning

- Interacts with its environment producing actions and discovers errors or rewards through trial and error search.
- Example algorithm: Q-Learning

Tabular Data: Malware

Features



The diagram illustrates a tabular dataset with 10 columns and 10 rows. A purple horizontal line with a double-headed arrow spans the first five columns (Bus-cycles to Node-stores). A red arrow points from the word 'Features' to this purple line. Another red arrow points from 'Features' to the last three columns (Branch-loads to L1-dcache-stores). A third red arrow points from 'Samples' to the first column (Bus-cycles).

Bus-cycles	Branch-instructions	Cache-references	Node-loads	Node-stores	Cache-misses	Branch-loads	LLC-loads	L1-dcache-stores	Class
11463	37940	8057	1104	111	2419	37190	2360	38598	Backdoor
1551	5055	1096	165	17	333	4916	330	5003	Backdoor
29560	126030	20008	1769	146	4098	108108	5987	99237	Backdoor
26211	117761	14783	1666	48	4182	117250	4788	91070	Backdoor
30139	123550	20744	1800	158	4238	124724	6969	115862	Backdoor
12989	30012	9076	1252	136	5412	27909	2000	27170	Benign
6546	12767	4953	548	87	3683	13157	864	12361	Benign
8532	31803	7087	699	124	3240	34722	1970	34974	Benign
14350	27451	9157	1843	178	6611	28507	2411	24908	Benign
13837	25436	12235	1296	192	7148	24747	2533	23757	Benign

Samples

Tabular Data: Airline

Index	Description	Success	Guests	Seat class	Customer ID	Fare	Age	Title	Gender
0	Braund, Mr. Owen Harris; 22	0	1	3	1	7.25	22	Mr	Male
1	Cumings, Mrs. John Bradley ...	1	1	1	2	71.3	38	Mrs	Female
2	Heikkinen, Miss. Laina; 26	1	0	3	3	7.92	26	Miss	Female
3	Futrelle, Mrs. Jacques Heath...	1	1	1	4	53.1	35	Mrs	Female
4	Allen, Mr. William Henry...	0	0	3	5	8.05	35	Mr	Male
5	Moran, Mr. James;	0	0	3	6	8.46	0	Mr	Male
6	McCarthy, Mr. Timothy J; 54	0	0	1	7	51.9	54	Mr	Male

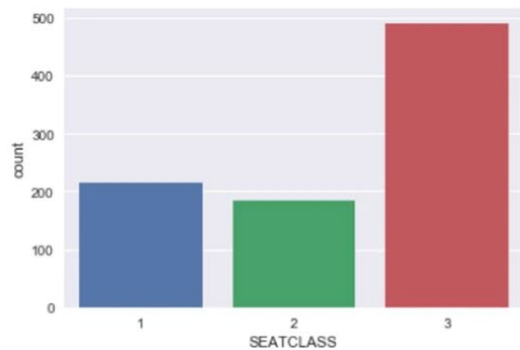
- GUESTS: Number of guests accompanying the customer.
- SUCCESS: Categorical variable that displays whether customer traveled or not.
- SEATCLASS: Categorical variable that displays the seat class of the customer.
- AGE: Numerical variable corresponding to the age of the customer.
- GENDER: Categorical variable describing the gender of the customer.
- FARE: Numeric variable for the total fare paid by the customer.
- SUCCESS: Categorical class variable indicating if the customer flies with the airline.

Correlation Matrix



Data Visualization

TOTAL PASSENGERS PER SEAT CLASS



Visualization tools

- Matplotlib
- **Seaborn**
- **Plotly**
- Pygal
- **Bokeh**
- Missingnoh
- plotnine/ggplot

Resources:

<https://seaborn.pydata.org/>

<https://www.data-to-viz.com/>

The Iris Dataset



Iris Versicolor

Iris Setosa

Iris Virginica

class: { Iris Setosa, Iris Versicolour, Iris Virginica }

Iris Dataset Attributes:

1. sepal length
2. sepal width
3. petal length
4. petal width
5. class



Independent
variables



Dependent
variable

$$f(\mathbf{x}) = y$$

Model type: Classification, Regression, Clustering

Supervised Learning

Classification

Regression

Naive Bayes

Support Vector Machines

Neural Nets

Decision Trees

Classification:

Object Recognition

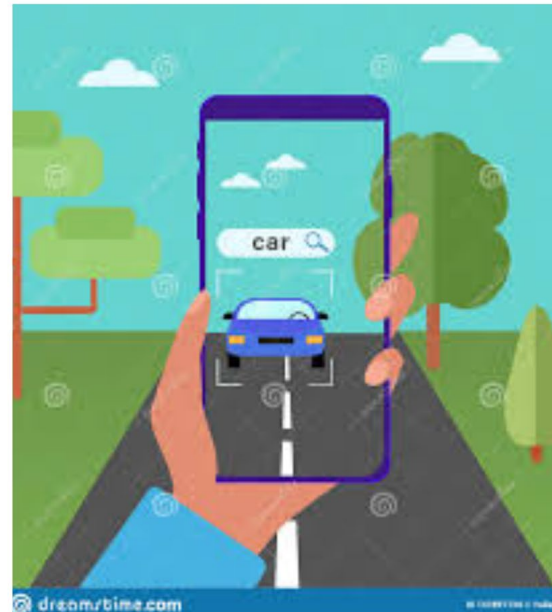
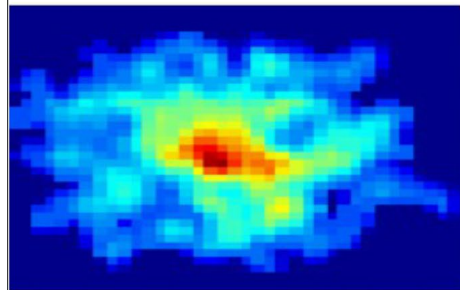
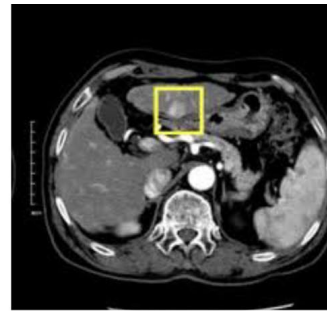
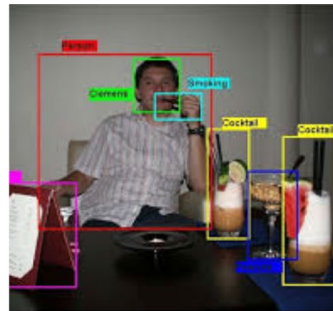
Face Detection

Face Recognition

Scene Recognition

Malware Detection

And more!



Regression:

Engine Performance

Prediction

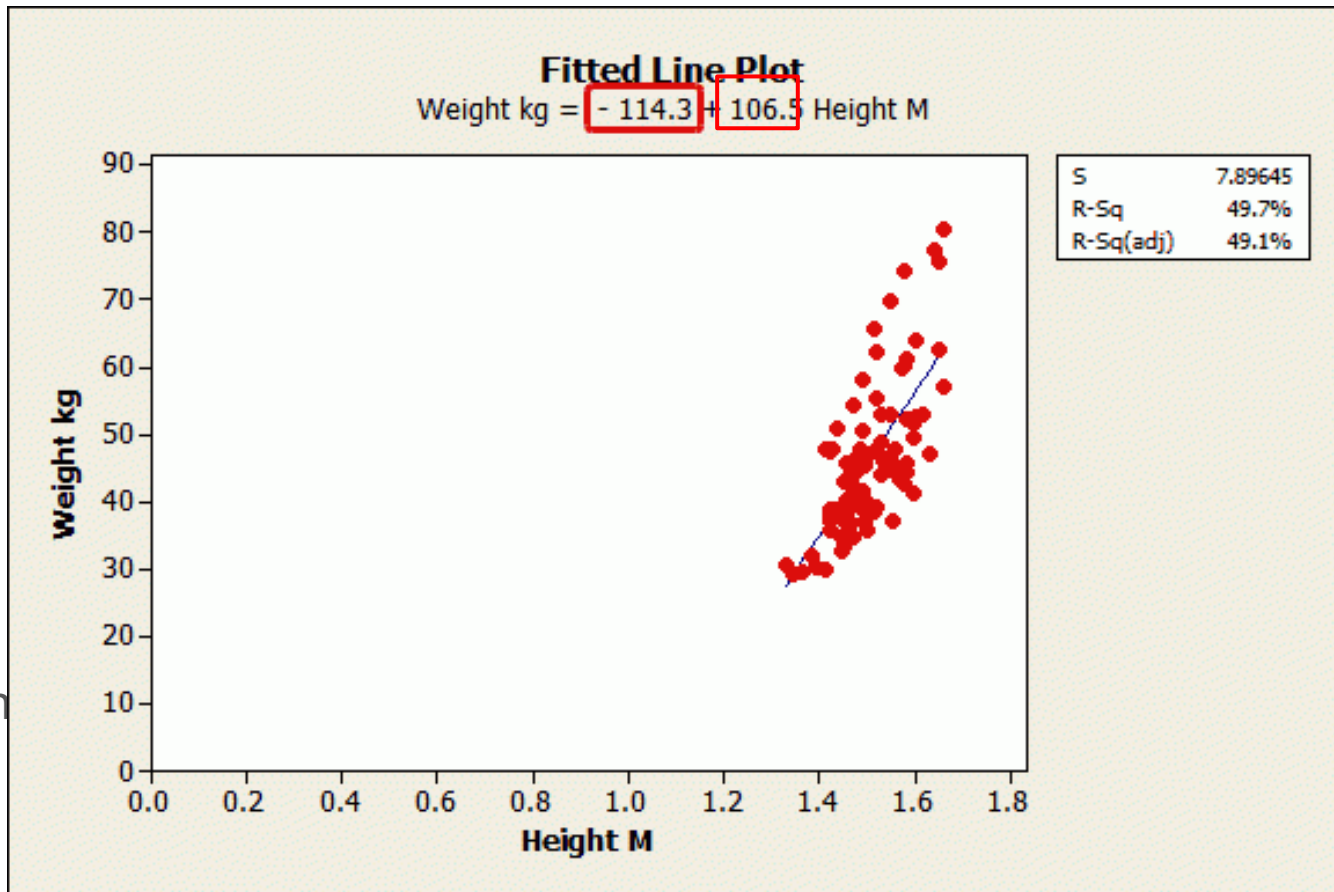
Business

Real-Estate Market

Prediction

Stock Market Prediction

Weather Data Analysis



$$y = mx + \beta$$

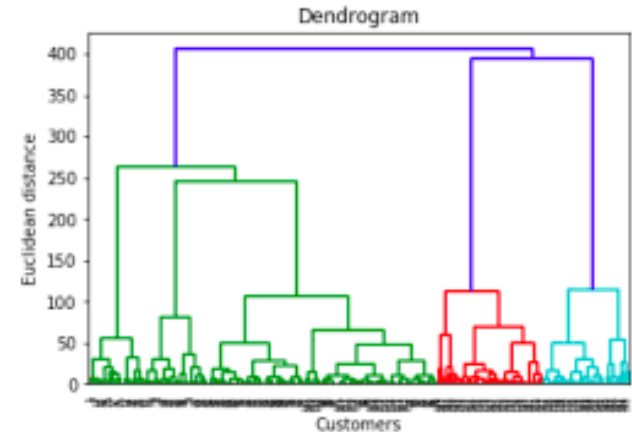
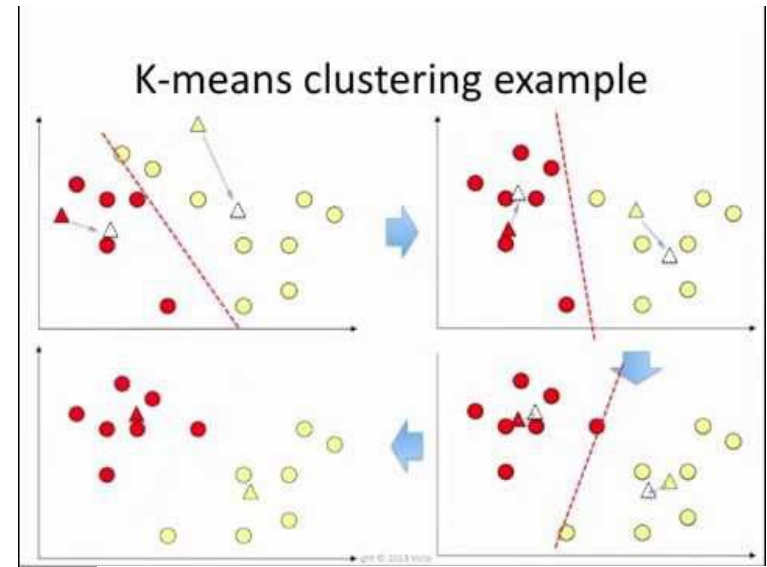
Unsupervised Learning

Knowledge Discovery

- Data Exploration
- Descriptive Task
- Un-labeled Dataset
- Common Algorithms

Clustering

- Example: K-means Clustering,
- Hierarchical Clustering, Self-
- Organizing Map
- Association Rule Discovery



Clustering

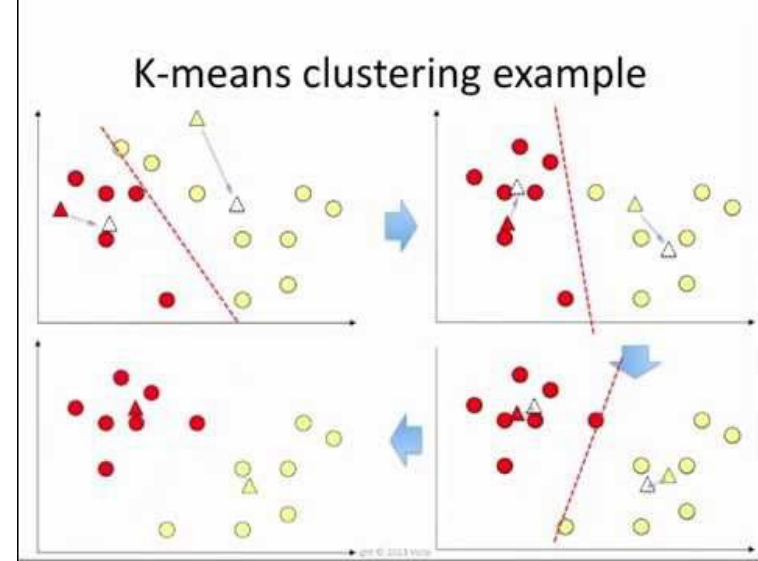
Market Segmentation Example

Goal:subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.

Approach:

Collect different attributes of customers based on their geographical and lifestyle related information.

Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.



Association Rule Discovery

Consumer Shopping Example

Given a set of records each of which contain some number of items from a given collection

Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

Support & Confidence measures

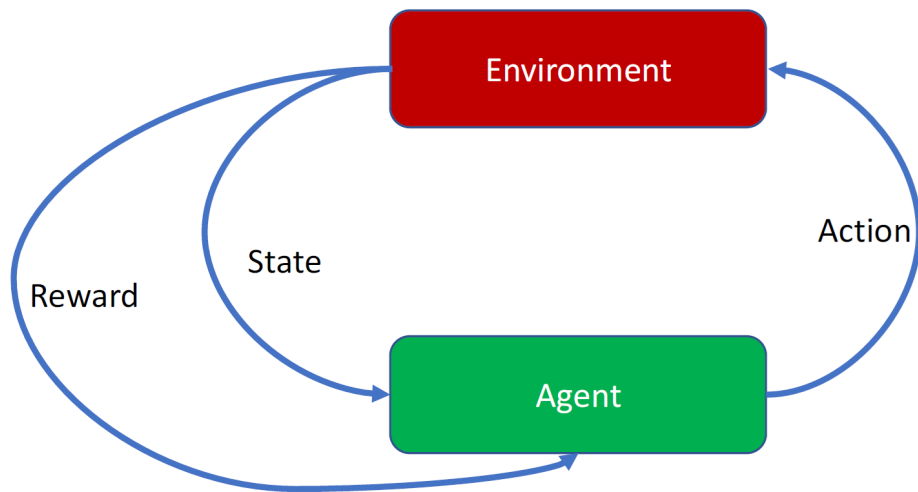
ID	
1	bread, Pepsi, milk
2	beer, bread
3	Pepsi, beer, diaper, milk
4	beer, bread, diaper, milk
5	Pepsi, diaper , milk

Rules Discovered:

$\{\text{milk}\} \rightarrow \{\text{Pepsi}\}$

$\{\text{Diaper, milk}\} \rightarrow \{\text{beer}\}$

Reinforcement Learning



Two important learning models in reinforcement learning:

1. Markov Decision Process
2. Q learning

