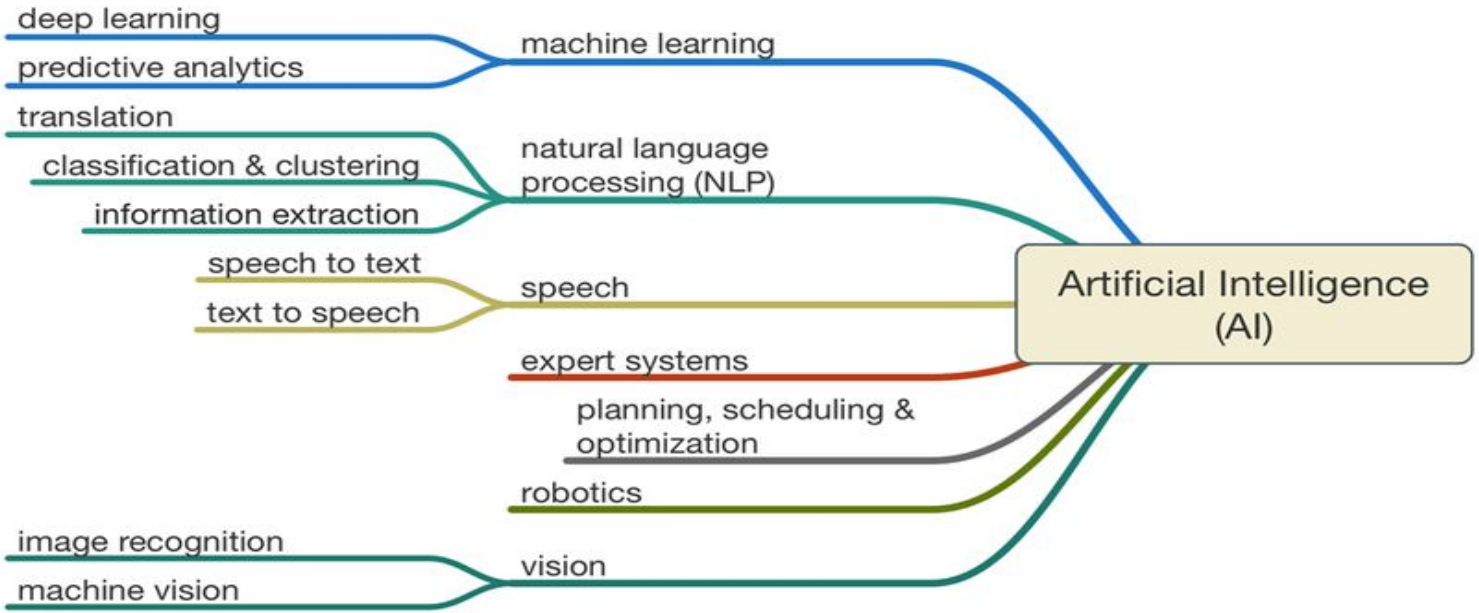
Introduction to AI with practical examples

Machine Learning

AI branches


Artificial Intelligence (AI)
machine learning
deep learning
predictive analytics
natural language processing (NLP)
translation
classification & clustering
information extraction
speech
speech to text
text to speech
expert systems
planning, scheduling & optimization
robotics
vision
image recognition
machine vision

Definitions

AI - Artificial Intelligence

- *is a technology using which we can create intelligent systems that can simulate human intelligence*

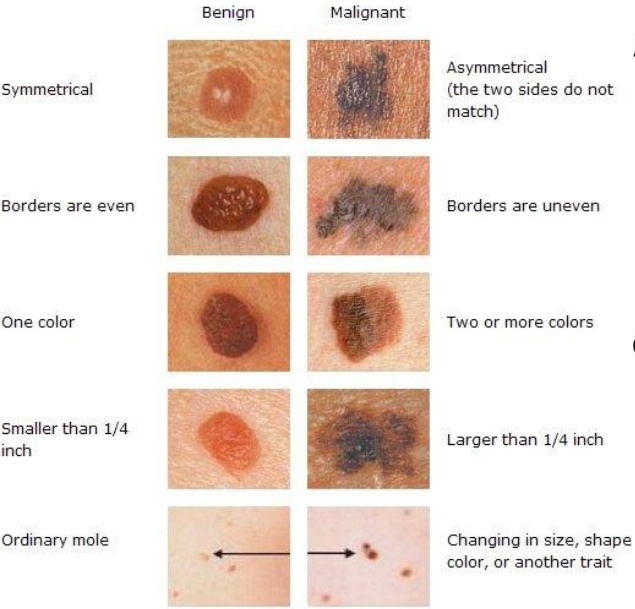
ML - Machine Learning

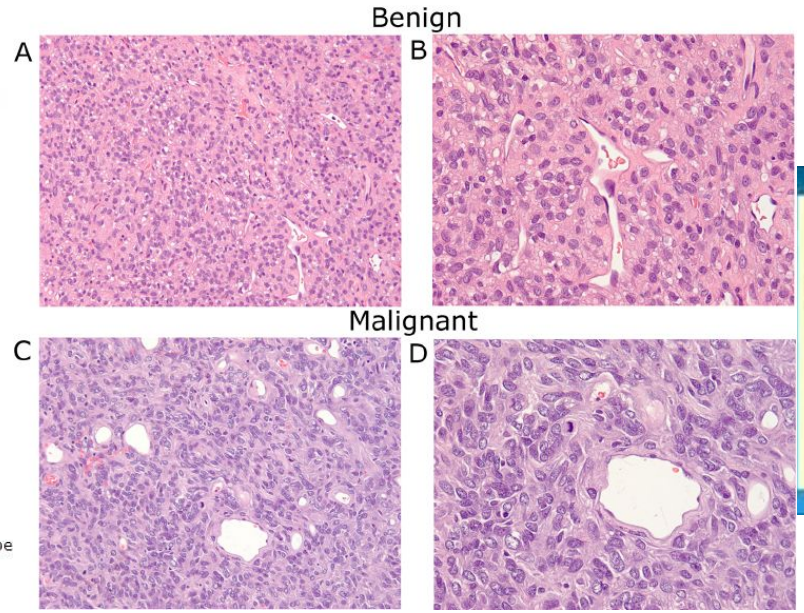
- *is a subfield of artificial intelligence, which enables machines to **learn** from past data or experiences without being explicitly programmed*
- *covers the statistical part of AI*
    - *it teaches the computer to solve problems by looking at hundreds or thousands of examples, learning from them,*
    - *use that experience to solve the same problem in new situations*

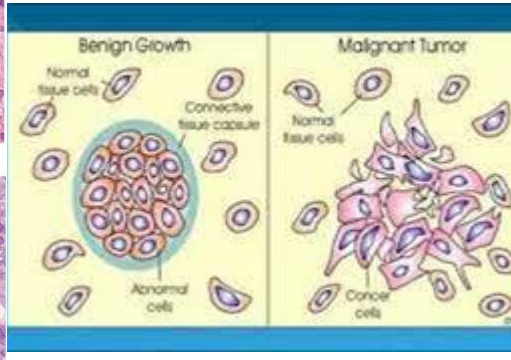
Example of machine learning

- **Recommend** product on websites
  - It is based on your *profile* build from past purchases, or *associations* with a selected product
    - If you frequently buy books, it will recommend new books and related products you didn't even know you wanted
    - If you search for a bike, it will recommend other similar bikes or bike related items (e.g. helmet)
- **Predict** whether a human cell that is believed to be at risk of developing cancer is either benign or malignant
- **Detect** objects in images
- Automated assistant, game playing, search, machine translation, etc

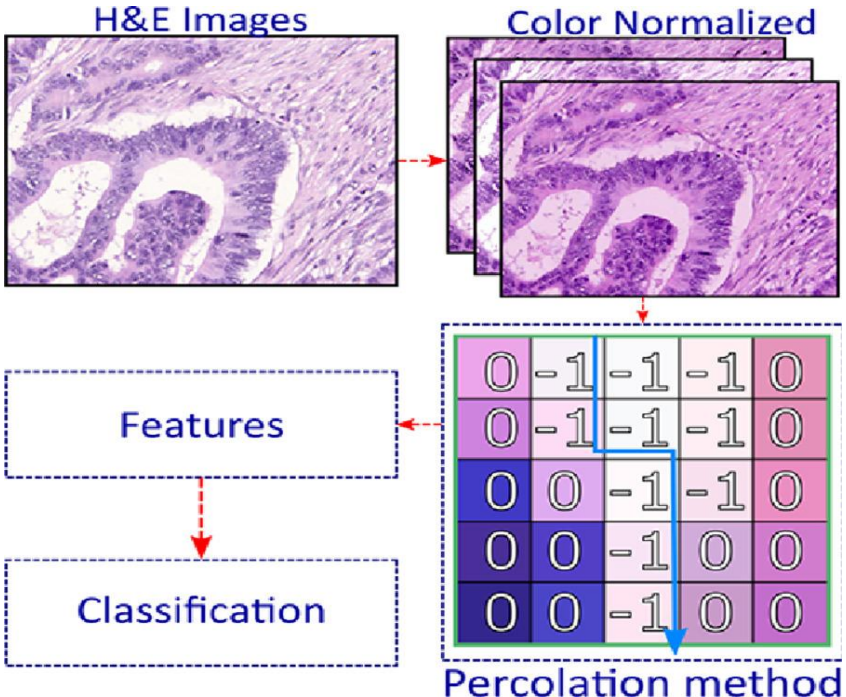
Detect cancer cells: malignant vs benign







Detect cancer cells: malignant vs benign

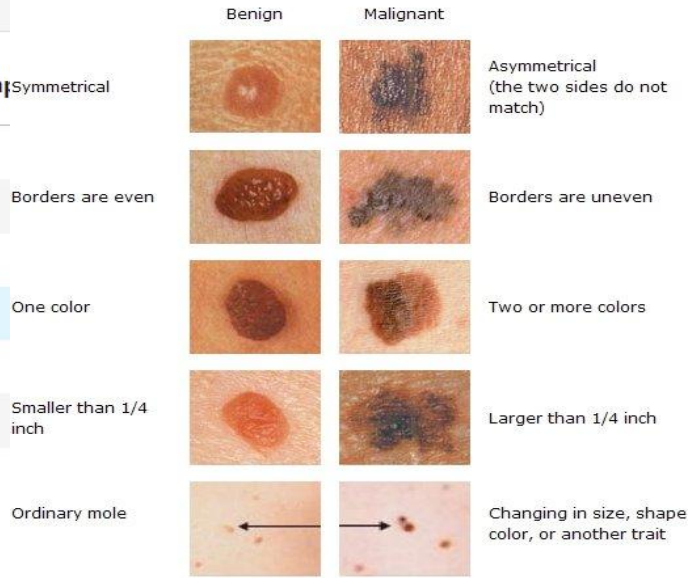


A practical view - detect cancer cells

```
df.tail(6)
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	comp
563	926125	M	20.92	25.09	143.00	1347.0	0.10990	
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	
566	926954	M	16.60	28.08	108.30	858.1	0.08455	
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	
568	92751	B	7.76	24.54	47.92	181.0	0.05263	

6 rows × 33 columns

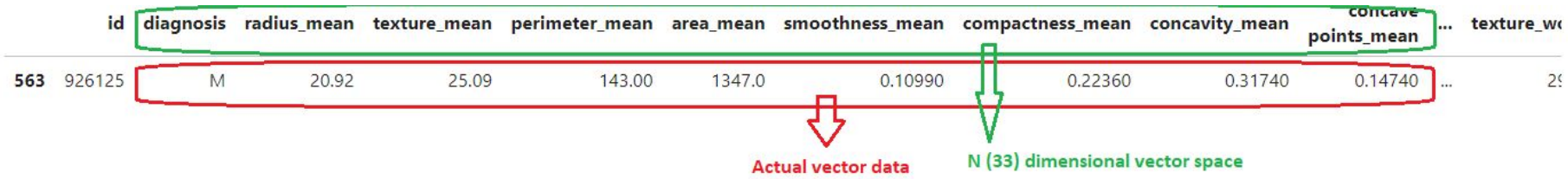


A practical view - cont.

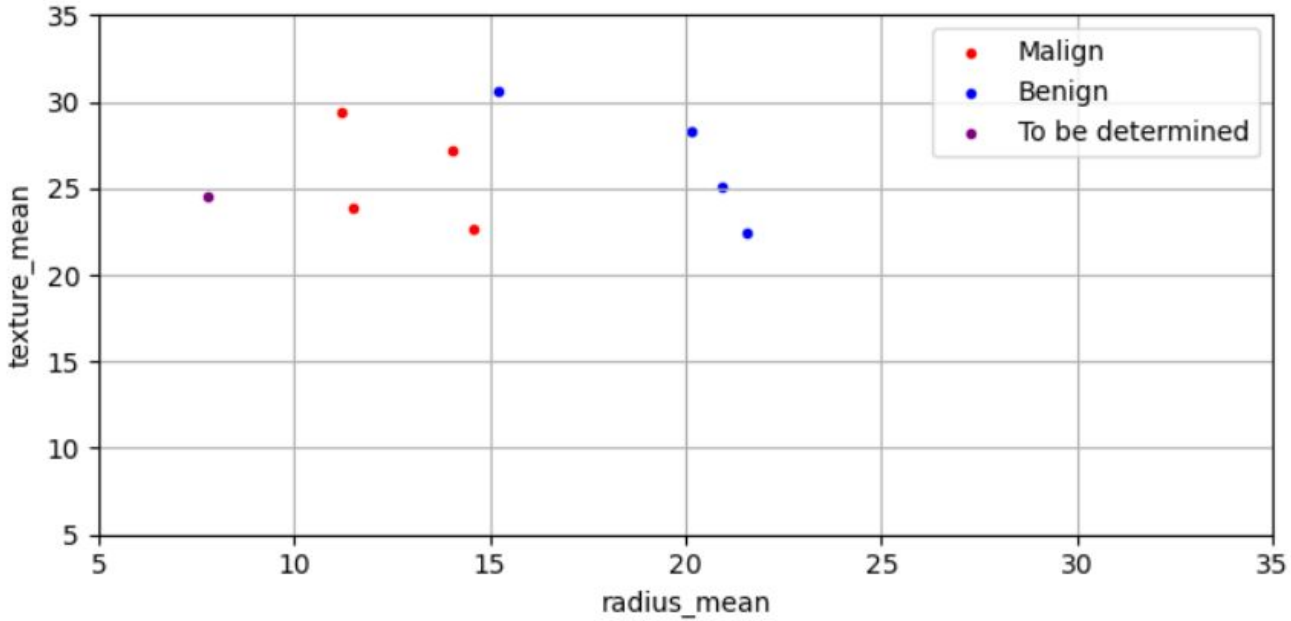
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	texture_w
563	926125	M	20.92	25.09	143.00	1347.0	0.10990	0.22360	0.31740	0.14740	...	2
564	926424	M	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	...	2
565	926682	M	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	...	3
566	926954	M	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	...	3
567	927241	M	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	...	3
568	92751	B	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	...	3
19	8510426	?	13.540	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	...	

Data from vectorial perspective

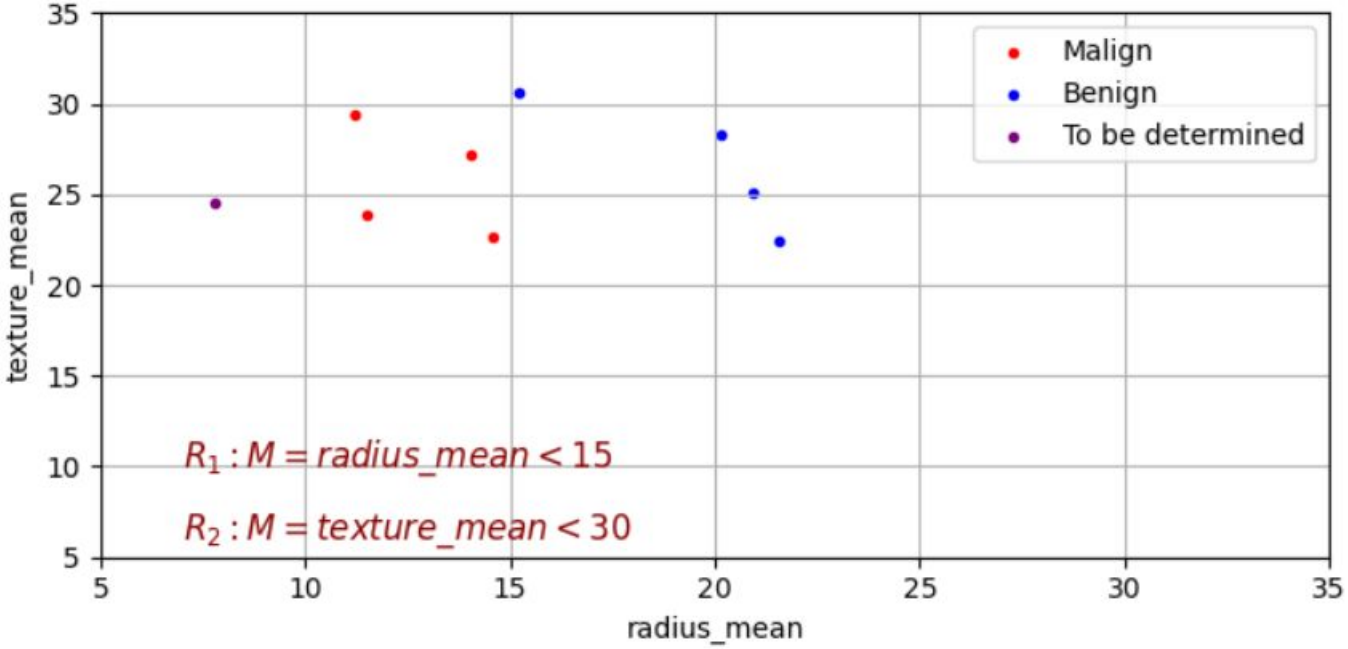
- From data to vector space



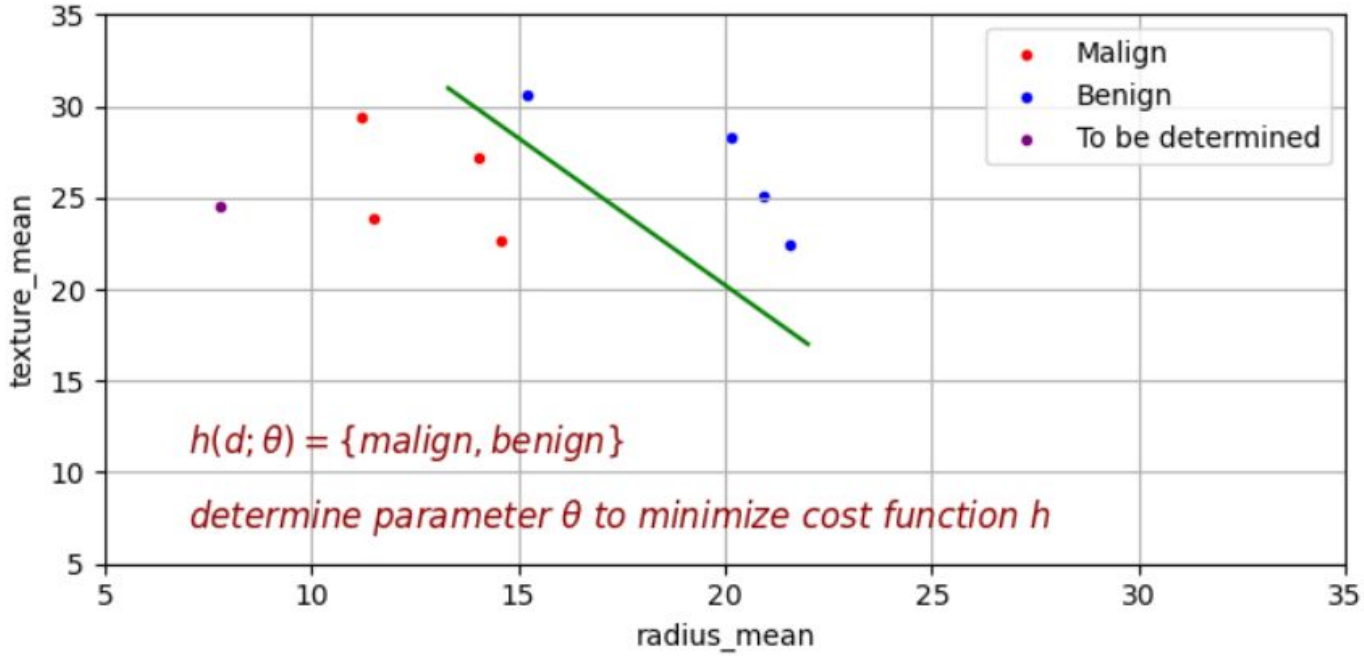
Cancer data simplified to 2D



Cancer data solution - with rules



The green line separates the data

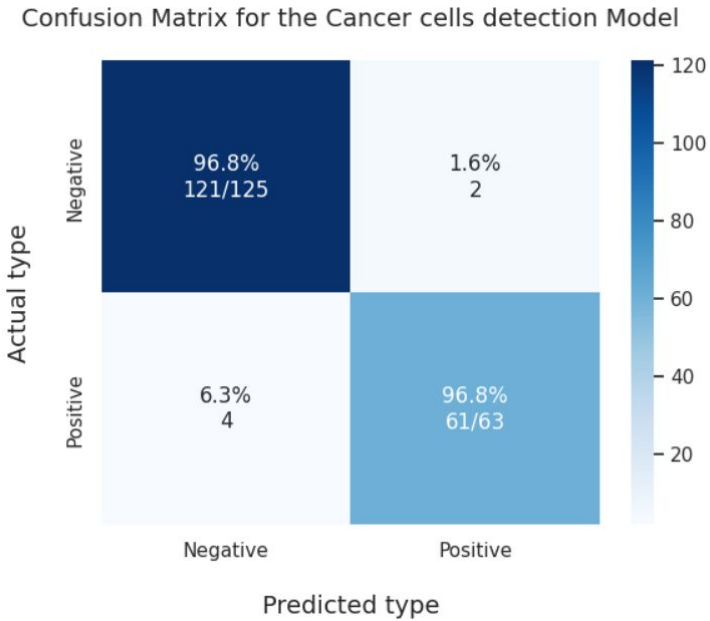


(Classification) solution template

- Load data
- Split the data:
  - 80% train
  - 20 % test data sets
- Train algorithm with train data set
- Test model with test data set
  - Check the classification accuracy

Demo with cancer data (svm)

- Run script from Jupyter notebook
  - Data pre-processing
    - Remove columns
    - Transform rows to numerical values
  - Train
    - Minimize error function
    - Examples with feature and corresponding label
  - Classification accuracy
    - Confusion matrix



(ML)Supervised learning - problem 2

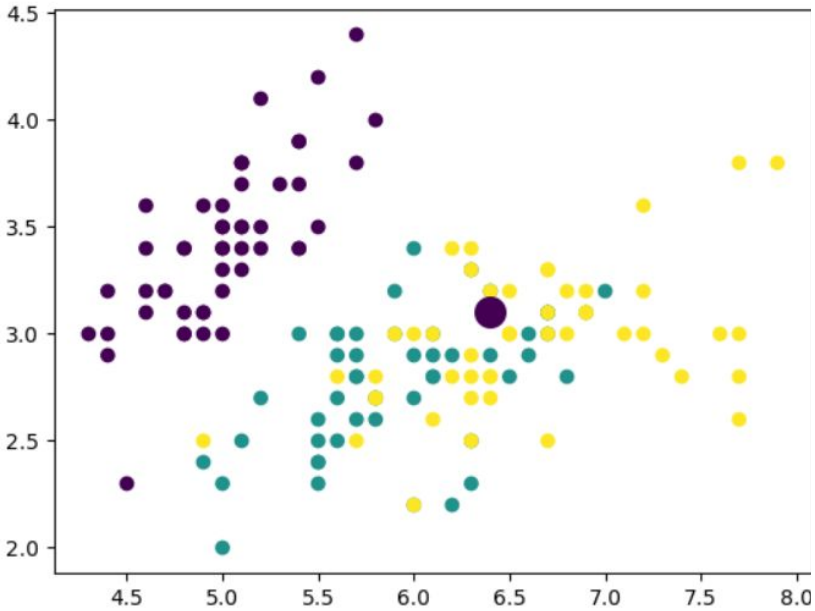
- Classification vs regression
  - Classification methods find the class from a finite domain
  - Regression methods predict a continuous value: temperature, C02 emissions, etc
- Classification problems
  - Detect iris flower classes

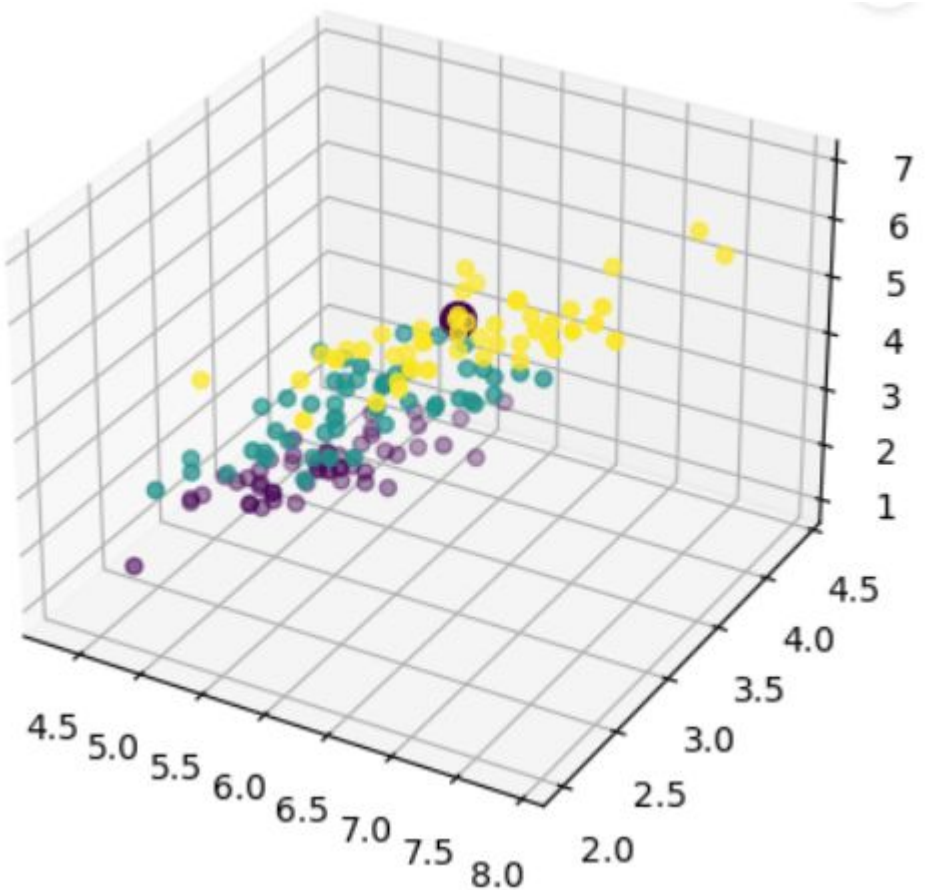
Detect Iris flowers

- Features: SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm
- Classes: 'setosa', 'versicolor', 'virginica'

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa

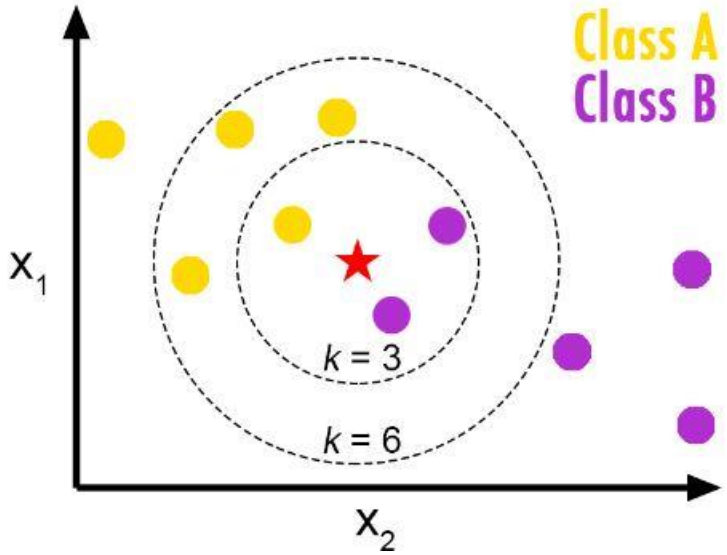
Problem solution - KNN





KNN classification overview

- Check the closest K neighbours
- KNN
  - Pros:
    - Easy to understand & implement
    - No training phase, does not build a model
  - Cons
    - Speed declines as dataset grows
    - Optimal K for new data
    - Imbalanced data favors the predominant class



Classification Recap

- Classification definition

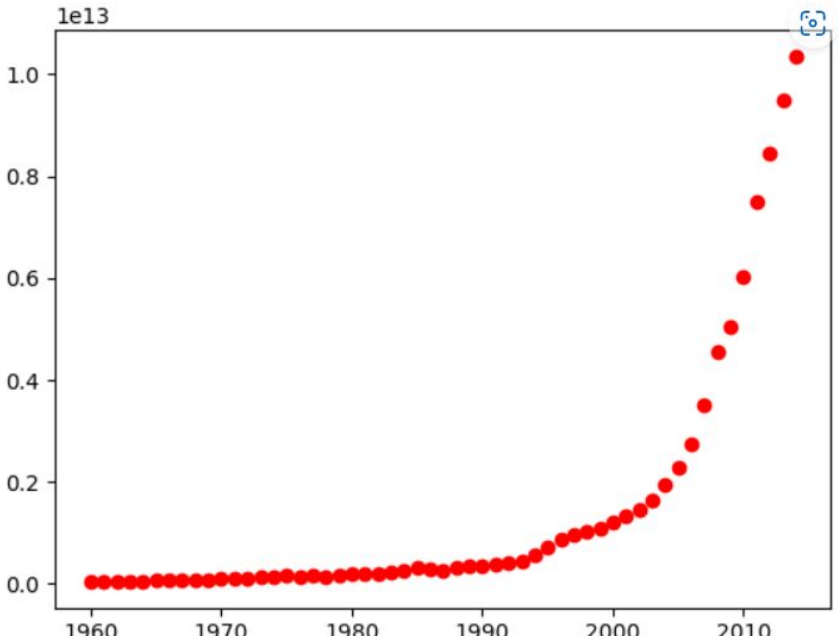
  $S = \{(x^i, y^i) | i = 1..n\}$

  $x^i \leftarrow R^a, y^i \leftarrow \{-1, +1\}$

- Y^i can have more than two classes

How to tackle continuous values?

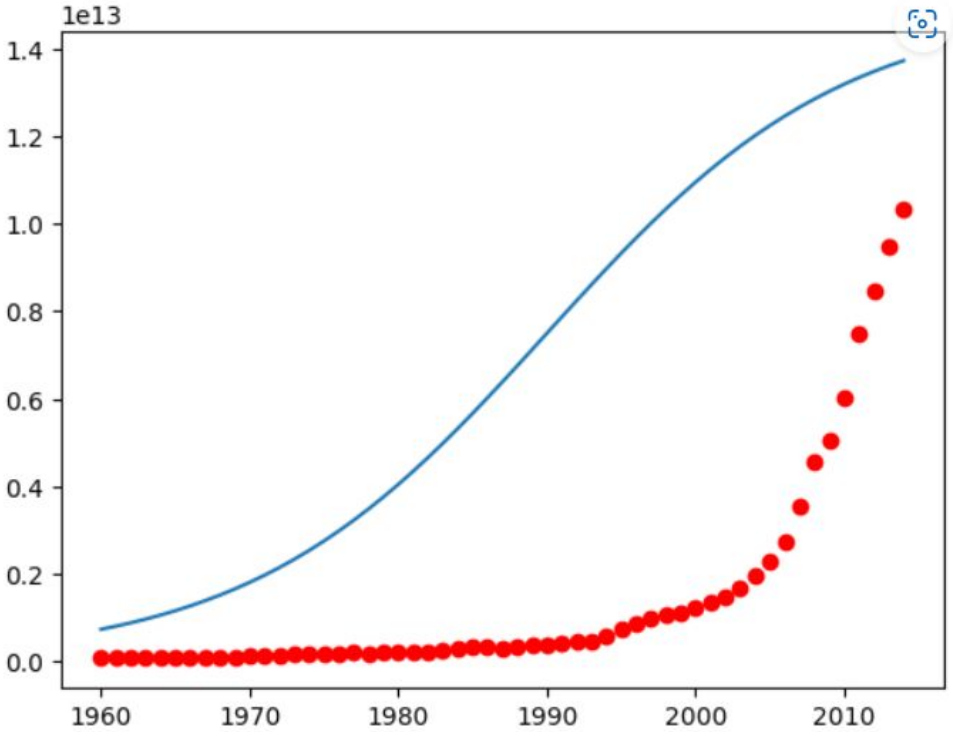
- How to predict ?
  - Temperature
  - Stock price
  - CO2 emission
  - etc..



	Year	Value
0	1960	5.918412e+10
1	1961	4.955705e+10
2	1962	4.668518e+10
3	1963	5.009730e+10
4	1964	5.906225e+10
5	1965	6.970915e+10
6	1966	7.587943e+10
7	1967	7.205703e+10
8	1968	6.999350e+10
9	1969	7.871882e+10

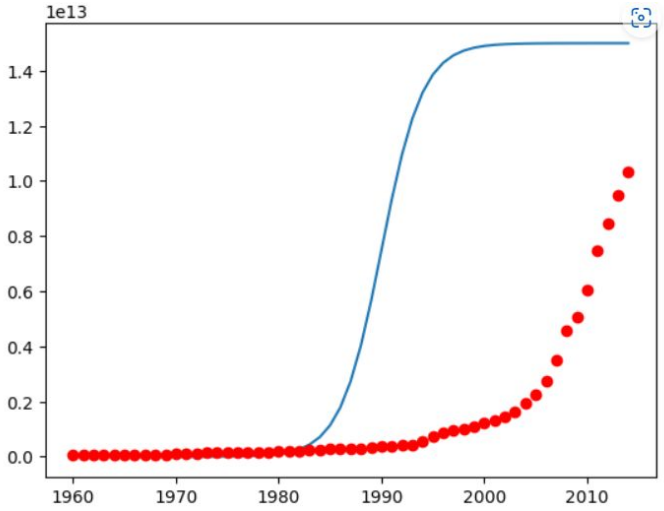
Predict China GDP

- Sigmoid Function – GeoGebra
  - Use sigmoid function as a base

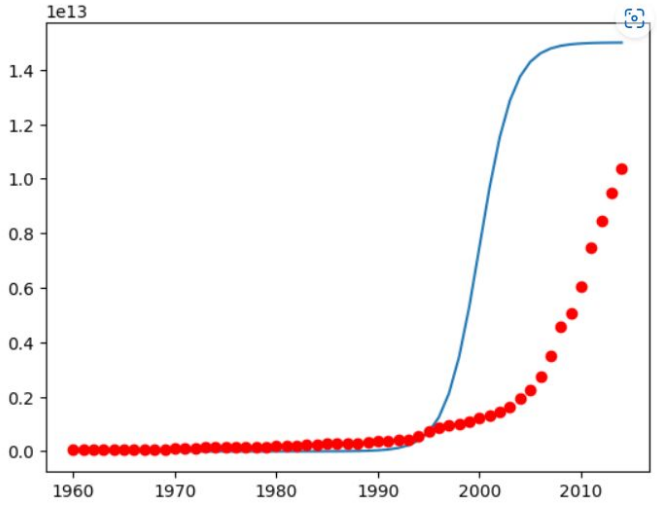


Sigmoid function

- beta_1 = 0.5, beta_2 = 1990.0

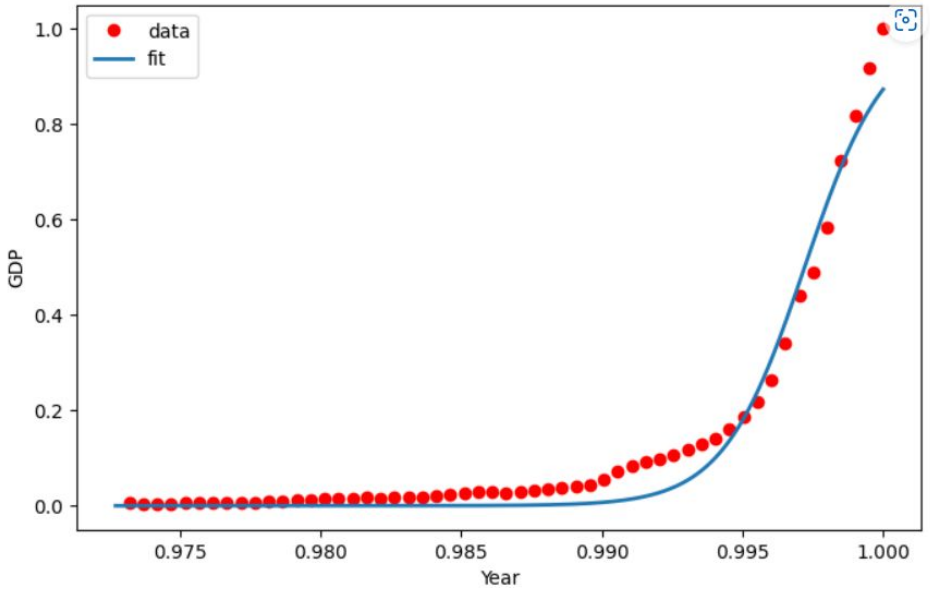


- beta_1 = 0.6, beta_2 = 2000

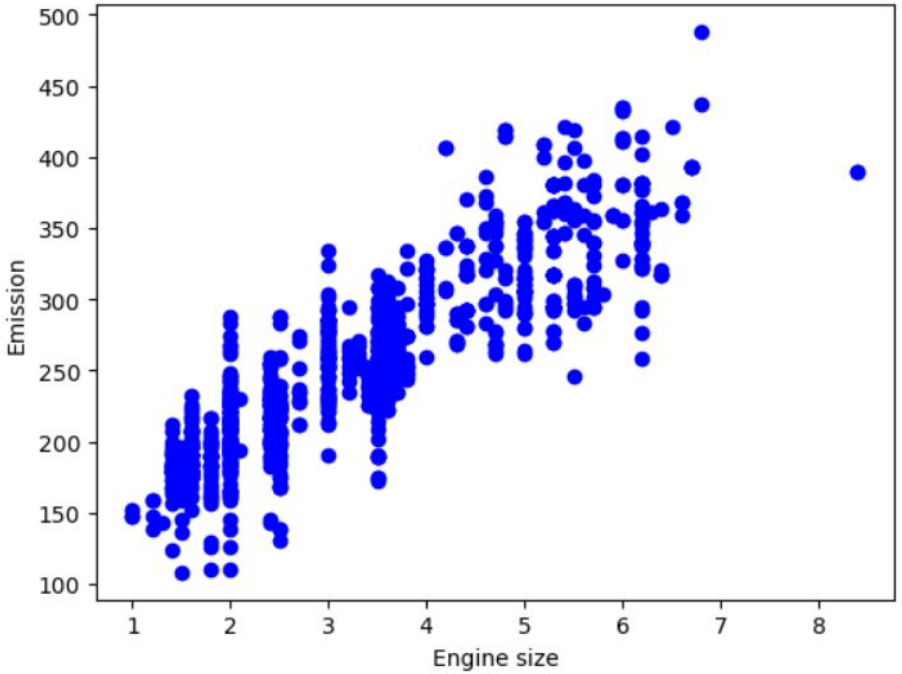


Fit the sigmoid parameters

- Run the Jupyter code
- beta_1 = 690.451709, beta_2 = 0.997207

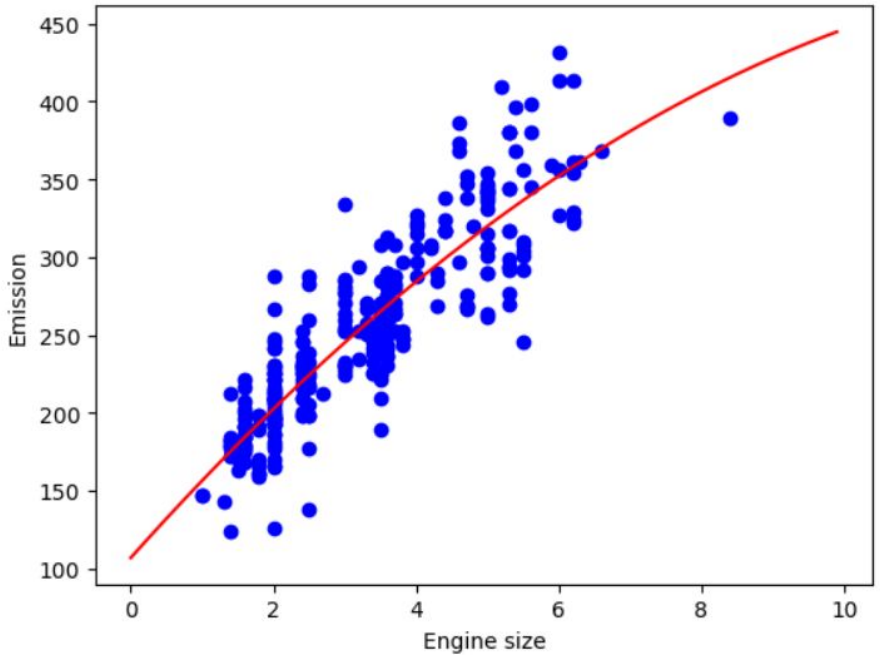


Predict CO2 emissions



Predict CO2 emissions with Linear regression

- Run Jupyter notebook script



Major machine learning techniques

- Regression/estimation
  - Predict continuous values
    - Predict the price of a house based on its characteristics
    - Predict the CO2 emissions of a car engine
- Classification
  - Predict the class or category of a case
    - Detect the letters from handwritten format
- Clustering
  - Finding the structure of the data; summarization
- Association
  - Associating frequently co-occurring items or events

Major machine learning techniques

- Anomaly detection
  - Discover abnormal and unusual cases
- Sequence mining
  - Predict next events; click-stream (Markov model, HMM)
- Dimension reduction
  - Reduce the size of the data (PCA)
- Recommendation system
  - Recommending items

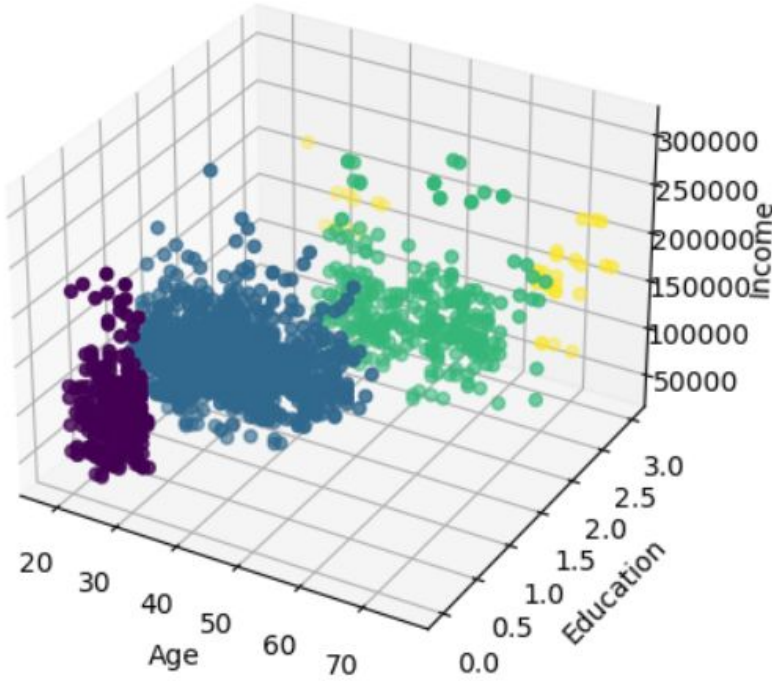
ML clustering - unsupervised learning

- Detect patterns in data
- E.g:
  - Customer data

	ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
0	100000001	0	0	67	2	124670	1	2
1	100000002	1	1	22	1	150773	1	2
2	100000003	0	0	49	1	89210	0	0
3	100000004	0	0	45	1	171565	1	1
4	100000005	0	0	53	1	149031	1	1

Clustering customer data

- The dataset consists of information about the purchasing behavior of 2,000 individuals from a given area when entering a physical ‘FMCG’ store
- Group similar customers



1 dim Similarity/distance calculation

Customer 1
Age
54

Customer 2
Age
50

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

Multi-dimensional similarity/distance

Customer 1		
Age	Income	education
54	190	3

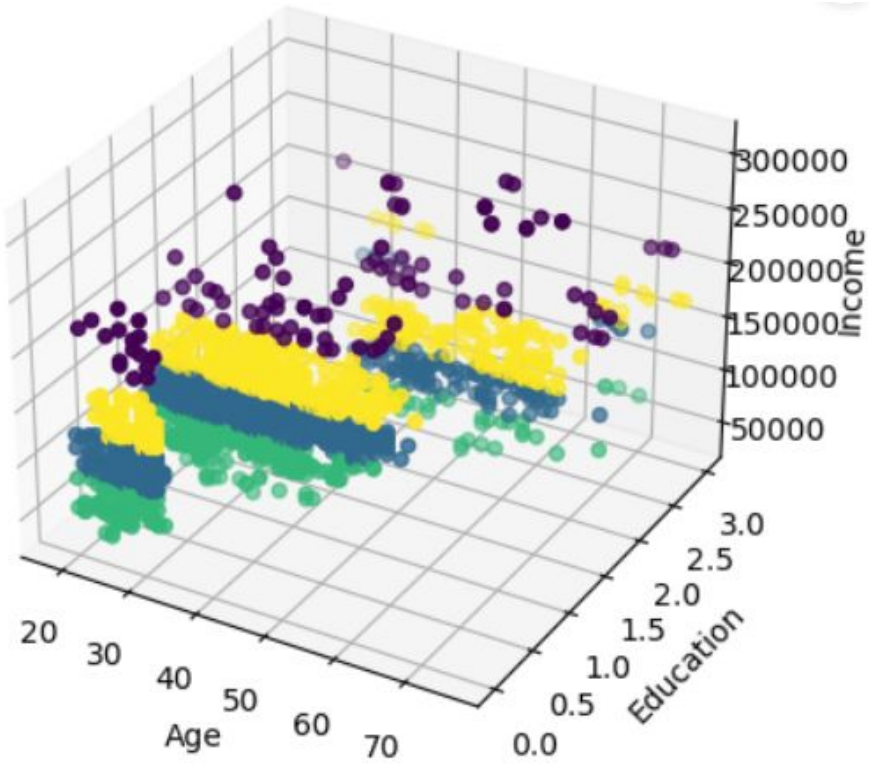
Customer 2		
Age	Income	education
50	200	8

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(54-50)^2+(190-200)^2+(3-8)^2} = 11.87$$

Clustered with K-means

- Data labeled by associated cluster
- Customers within cluster are similar
- Customers between clusters are very different
- K-Means tries to minimize the intra-cluster distances and maximize the inter-cluster distances.
- Run demo app



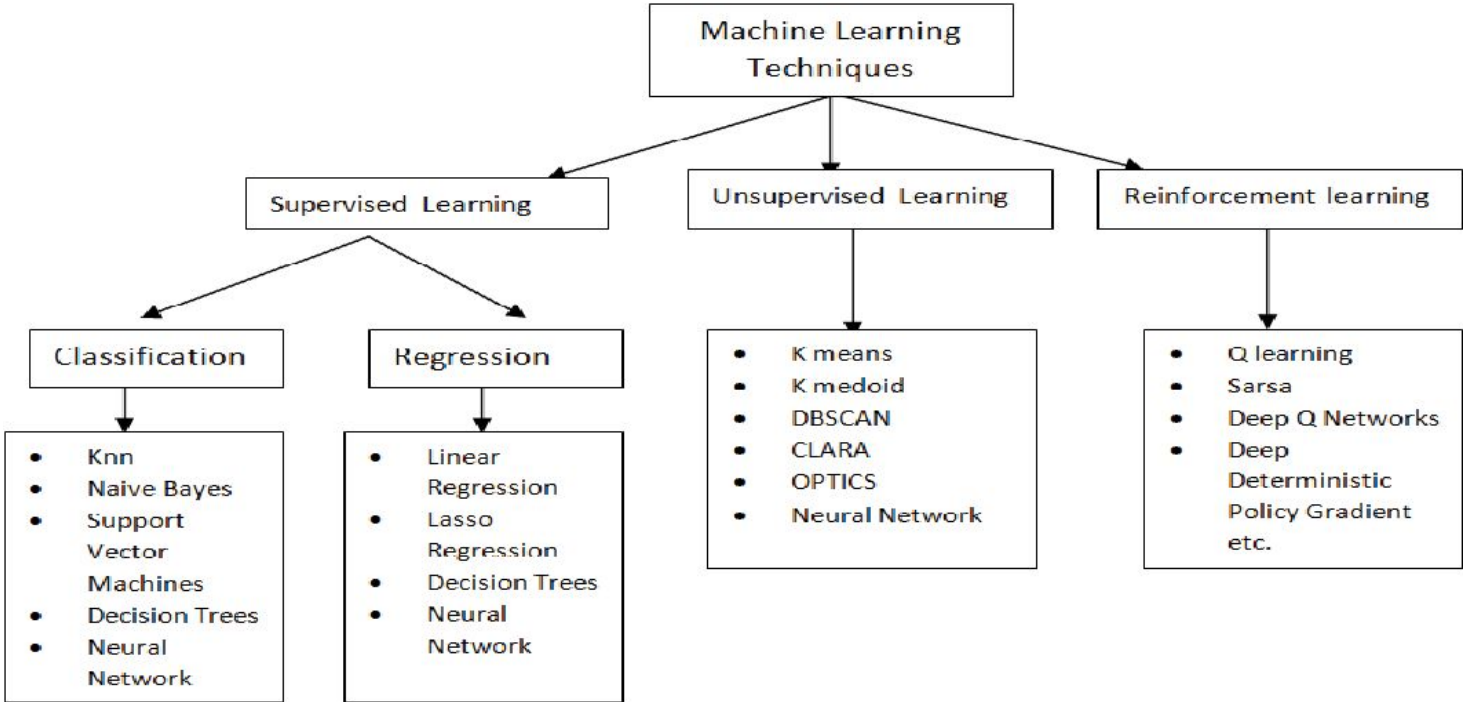
K-means clustering method

- Select randomly K cluster centroids
- Determine for each dataset point the closest centroid
- Update centroids
- Stop when no modification is performed
- Error calculated as
    - SSE=sum of squared differences between each point and its centroid

Demo with centroids change

- K-mean visual

ML fields


Machine Learning Techniques
Supervised Learning
Unsupervised Learning
Reinforcement learning
Classification
Regression
Knn
Naive Bayes
Support Vector Machines
Decision Trees
Neural Network
Linear Regression
Lasso Regression
Decision Trees
Neural Network
K means
K medoid
DBSCAN
CLARA
OPTICS
Neural Network
Q learning
Sarsa
Deep Q Networks
Deep Deterministic Policy Gradient
etc.

Toolkit

- Python
- Scikit-learn: scikit-learn: machine learning in Python — scikit-learn 1.2.1 documentation
- Numpy: NumPy
- Pandas: pandas - Python Data Analysis Library (pydata.org)

Online:

- Jupyter notebooks: JupyterLite

References

- Machine Learning with Iris Dataset | Kaggle
- Machine Learning with Python: A Practical Introduction