Machine learning algorithms: Lecture 1.

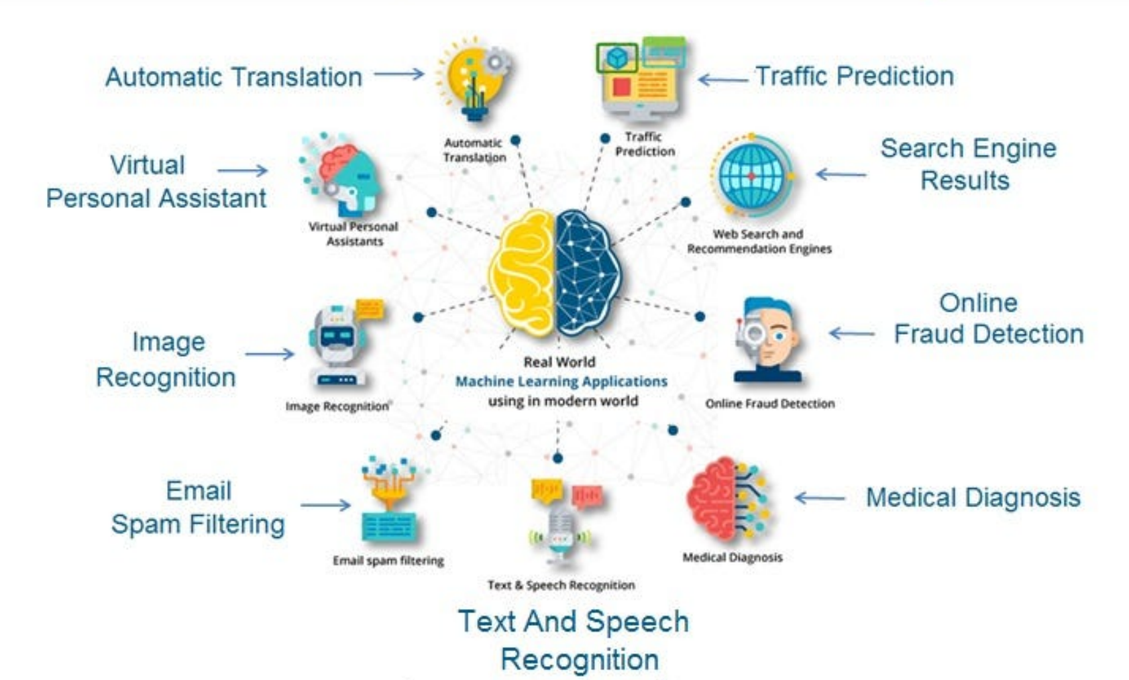
Intro to Machine learning

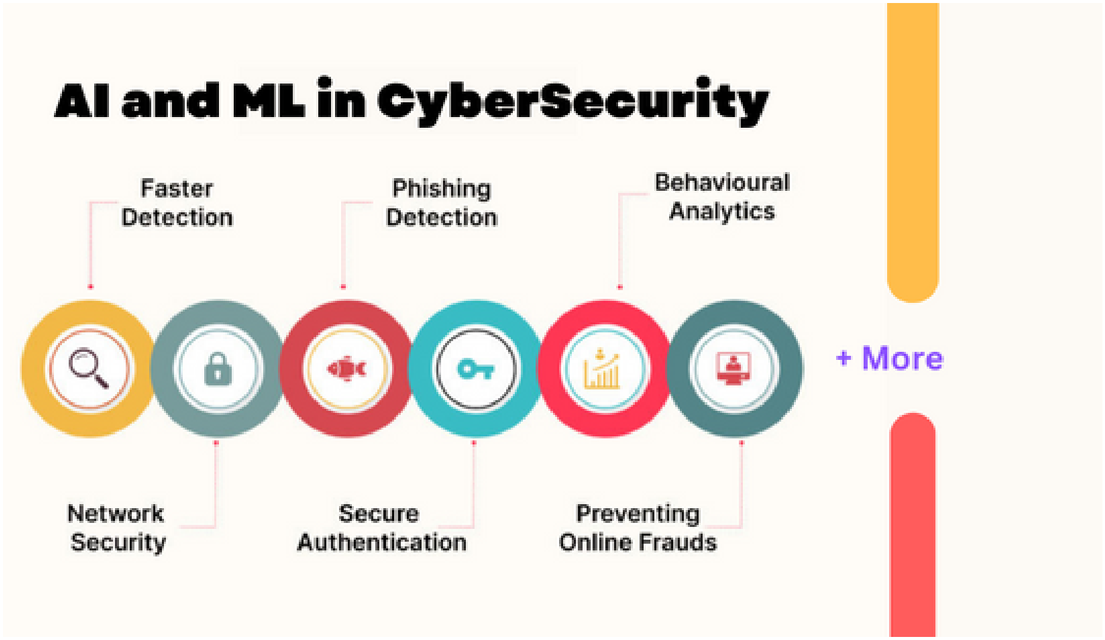
Alexandr Gavrilko
MLA Course for CS-2227

Machine learning

Machine learning is a subfield of artificial intelligence that gives computers the ability to learn without explicitly being programmed

Set the goal ⟶ Prepare data ⟶ Train the model ⟶ Predict outcomes for new data





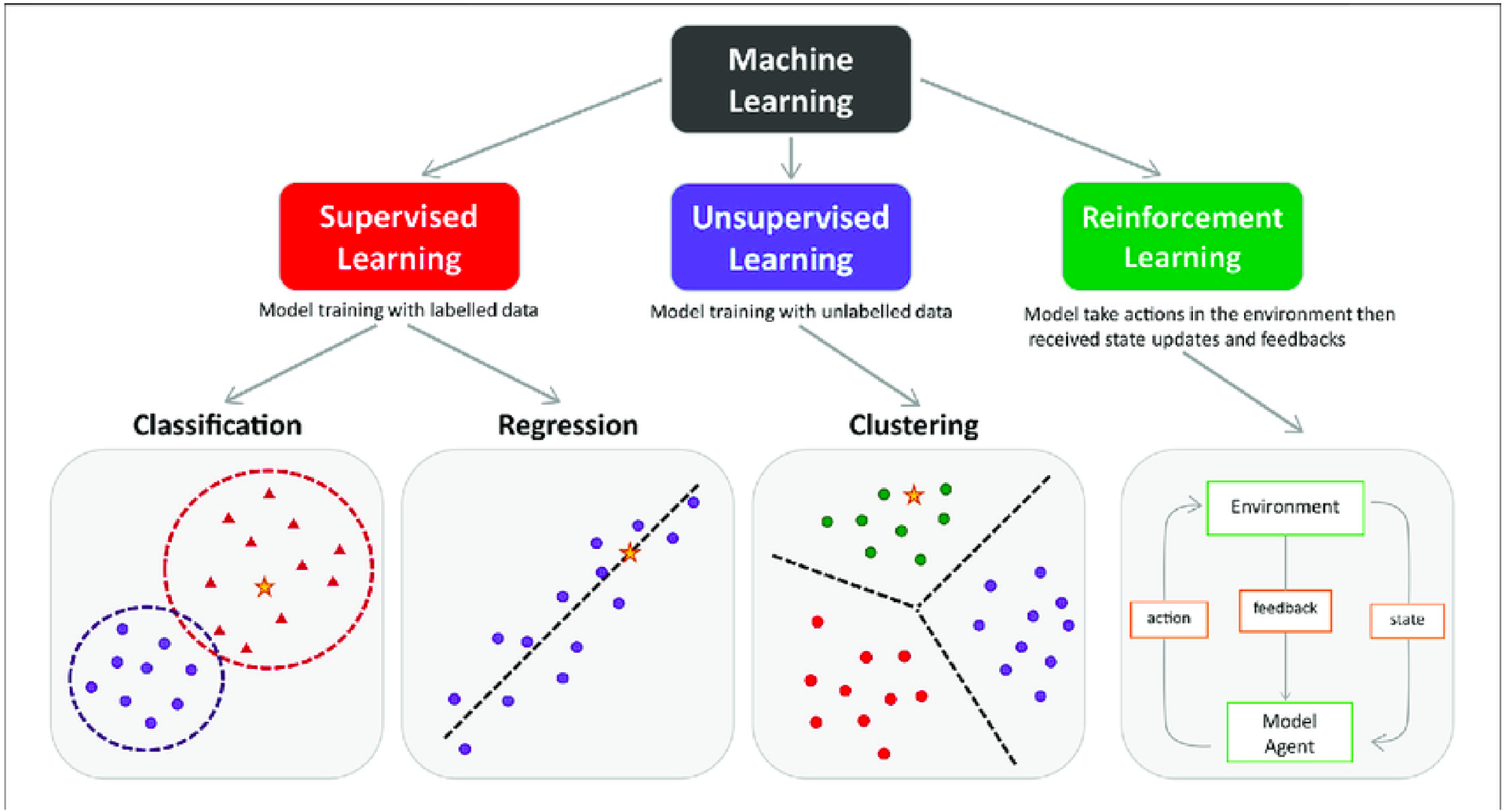
When to use machine learning?

- When the goal is too complicated to be coded: spam-filtering
- When the goal is continuously changing: Recommendation system
- When the goal is connected with perception: images, videos, speech, etc.
- When the goal is connected with unexplored phenomenon: predict social behavior of humans
- When it is economically beneficial: AI-Assistant for Customer Service

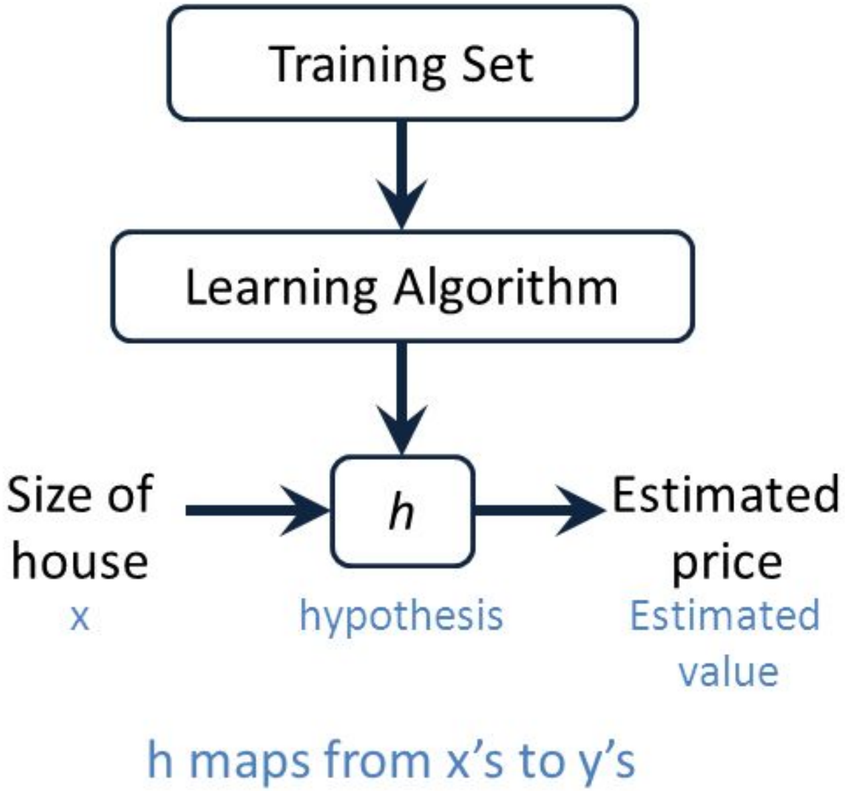
When not to use machine learning?

- When each action of the model should be interpretable
- When change in the behavior of the model should be interpretable
- Cost of the error is too high
- Collect the correct data is too hard or impossible
- When the goal can be achieved by traditional programming with less spendings
- When you can prepare mapping for all pairs “input -> output”

Types of Machine learning



Regression



How do we represent *h* ?

$h_\Theta(x) = \Theta_0 + \Theta_1 x$

Linear regression with one variable.
Univariate linear regression.

One variable

Andrew Ng

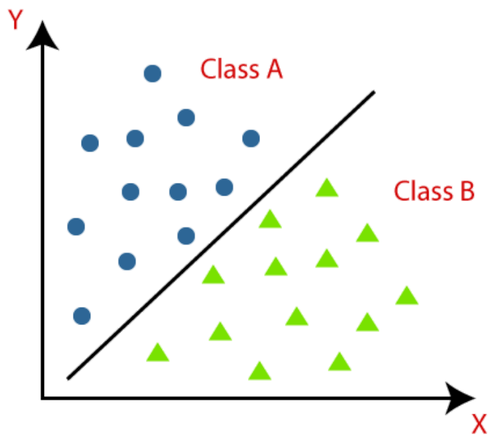
Regression models:

- Linear regression
- Support Vector Regression
- Lasso Regression
- Ridge Regression
- Decision Tree Regression

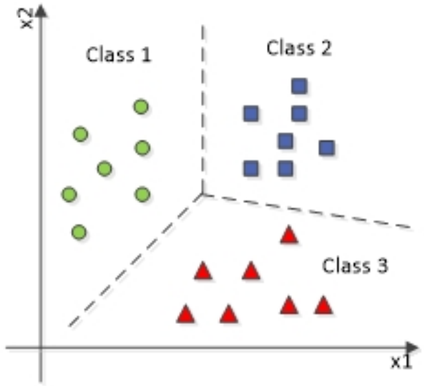
Classification

Classification types:

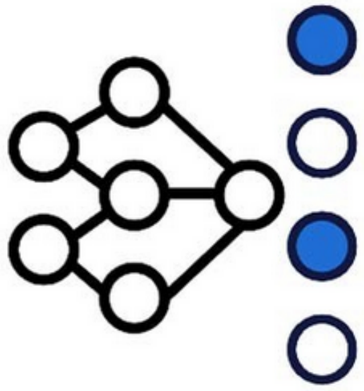
- Binary classification - only two possible outputs: 0 and 1
  For example: spam or not spam? dog or cat?
- Multi-class classification - three more possible outputs
  For example: species of flowers, digit recognition
- Multi-label classification - two more classes can be assigned to one example
  For example: genres for movie, tags for message



Binary classification



Multi-class classification

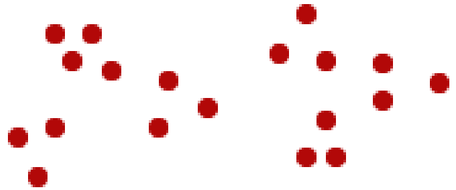


Multi-label classification

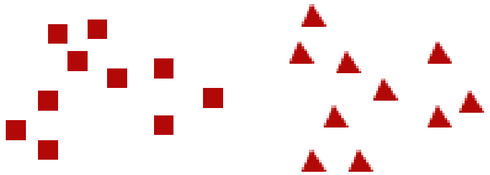
Classification models:

- Logistic regression
- Naive Bayes Classifier
- Decision Trees
- Support Vector Classifier
- K-Nearest Neighbors

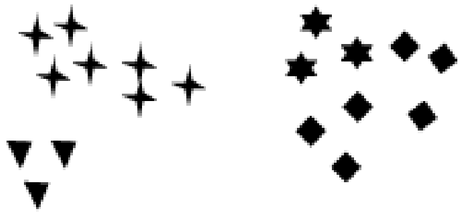
Clustering



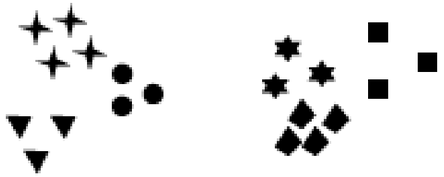
(a) Original points



(b) Two clusters



(c) Two clusters



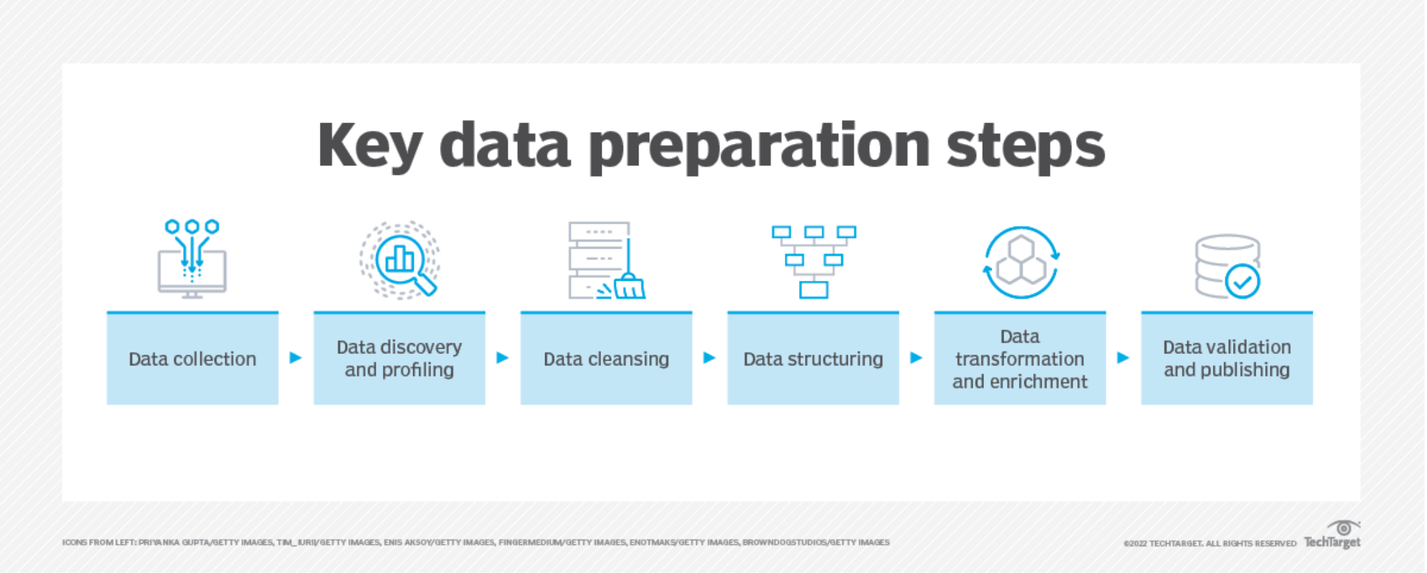
(d) Six clusters

Clustering models:

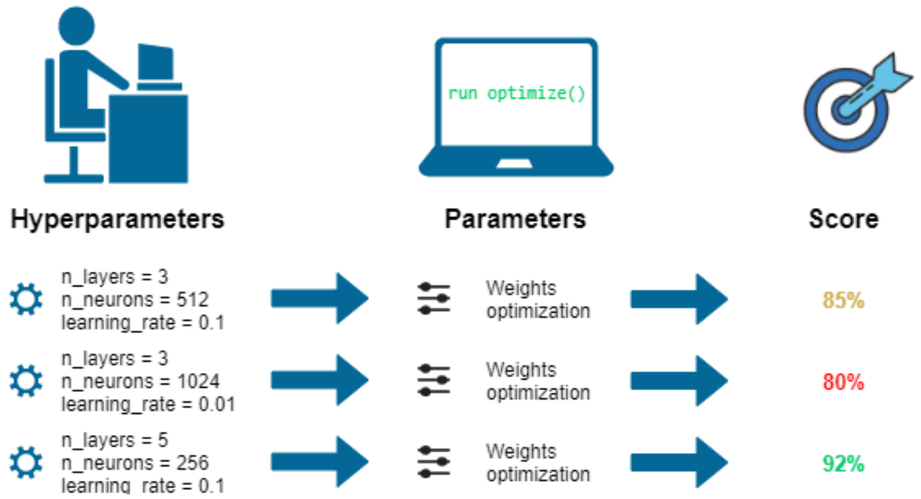
- K-Means clustering
- Agglomerative hierarchical clustering
- DBSCAN
- Support Vector Classifier
- K-Nearest Neighbors

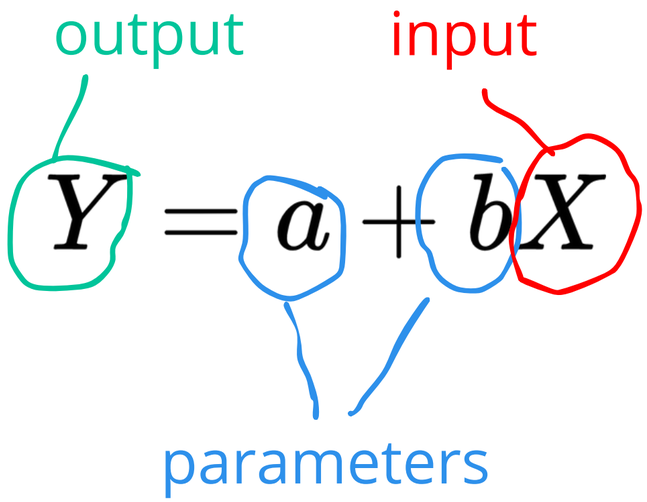
Terminology of Machine Learning

Data preparation.
Raw and accurate data



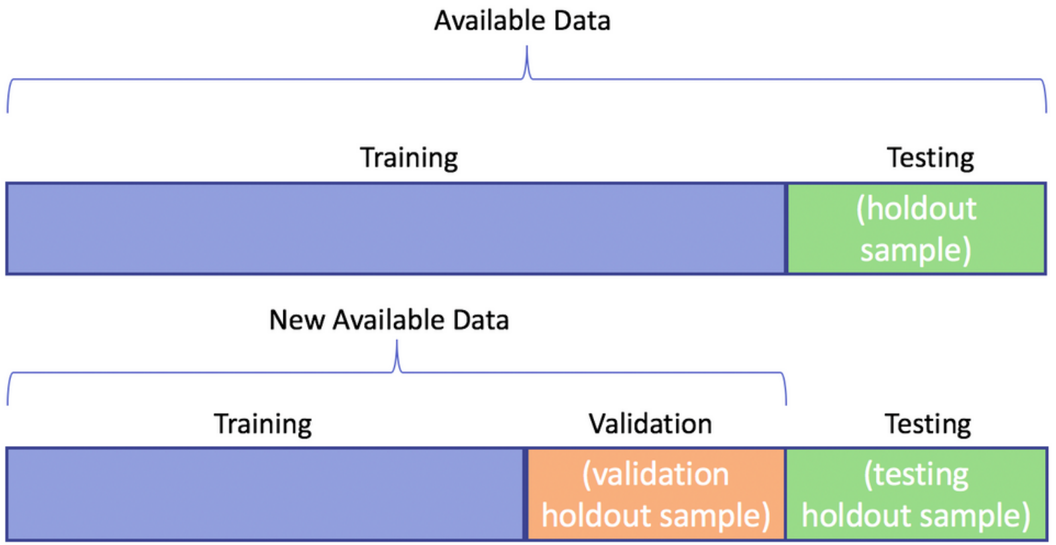
Parameters and hyper-parameters





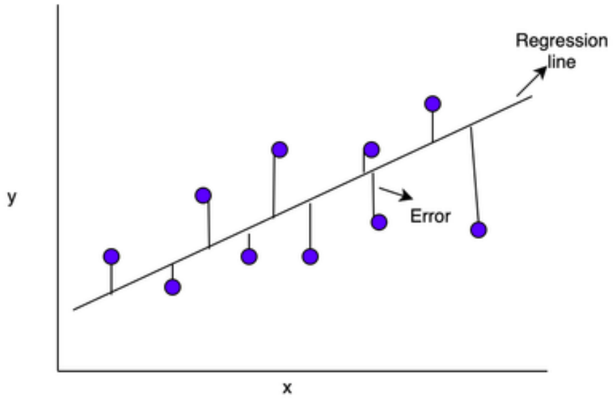
Terminology of Machine Learning

Training, validation and test splits



Metrics in Machine Learning

Regression



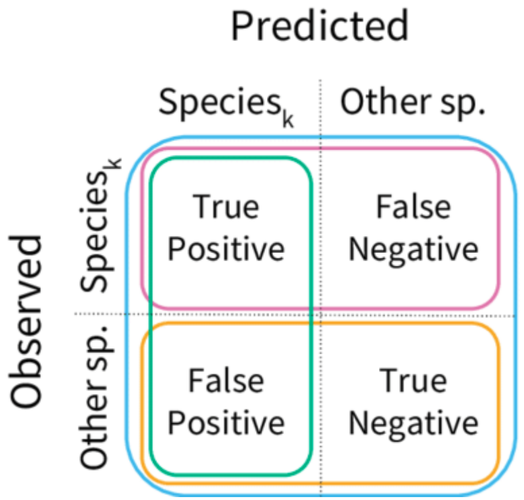
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Mean · Error · Squared

$$MAE = \frac{1}{n}\Sigma\left|y - \hat{y}\right|$$

Divide by the total number of data points · Predicted output value · Actual output value · Sum of · The absolute value of the residual

Classification



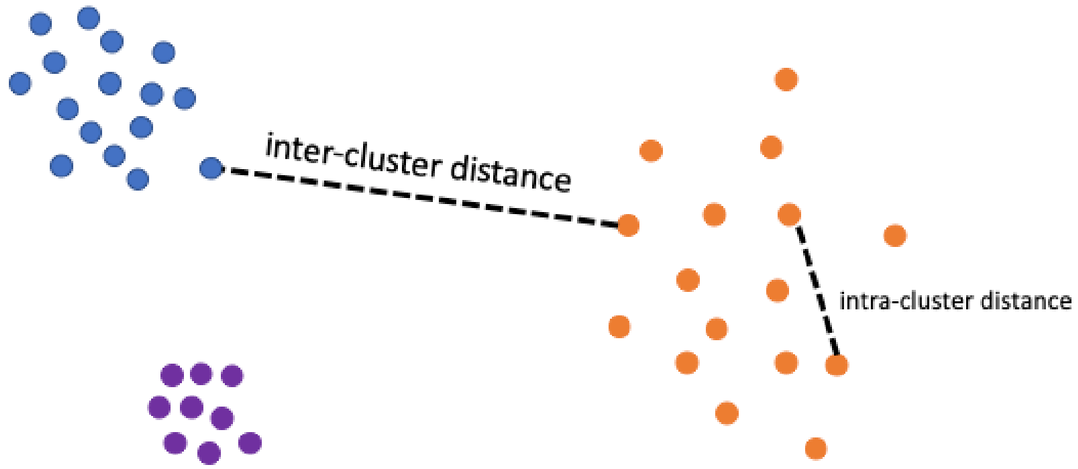
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Clustering



Machine Learning Lifecycle Pipeline