Machine Learning

What is Machine Learning?

Ai

RAWAN GHNEMAT

Textbooks

- 1. Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow: Concepts: Tools, and Techniques to Build Intelligent Systems, 2nd Edition, O’Reilly Media, Oct 2019.
-
- 2. François Chollet, Deep Learning with Python, Manning Pub. 2018

YouTube Video from Google Cloud

https://youtu.be/HcqpanDadyQ

After finishing this part of the course

- Solve an AI problem by developing an appropriate ML system.

- Use Python and its specialized libraries to develop programs for solving ML problems.

Outline

- Introduction
- Python programming language
- Data preparation and regression
- Classification
- Training models
- Classical techniques: SVM, decision trees and ensembles
- Neural networks
- Deep neural networks
- Convolutional neural networks
- Recurrent neural networks
- Reinforcement learning
- Recommendation systems

YouTube Video for ML

https://youtu.be/z-EtmaFJieY

https://www.youtube.com/watch?v=WXHM_i-fgGo

https://www.youtube.com/watch?v=IpGxLWOIZy4

Ai

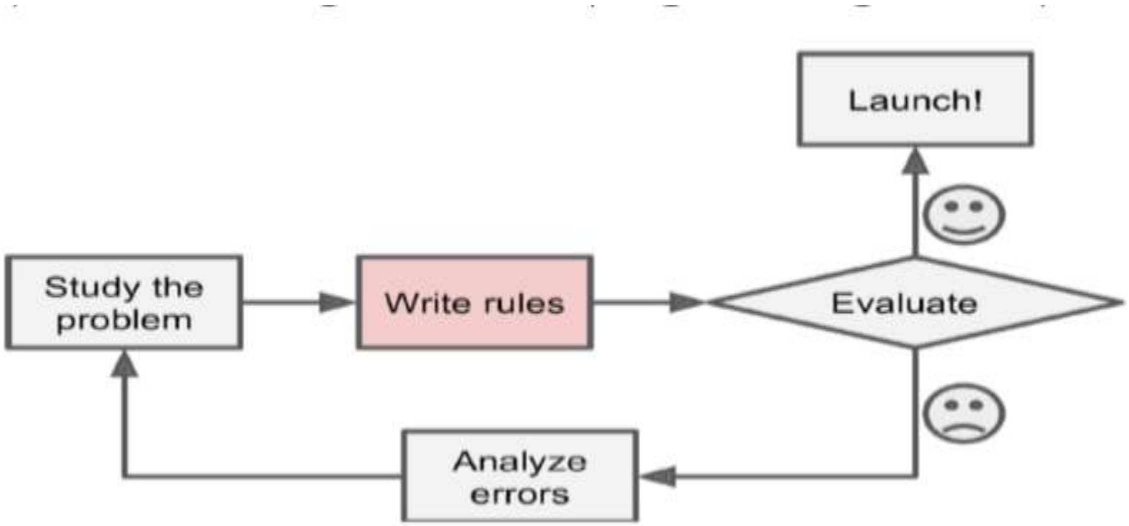
RAWAN GHNEMAT

What Is Machine Learning?

- The science (and art) of programming computers so they can learn from data.
- The field of study that gives computers the ability to learn without being explicitly programmed. Arthur Samuel, 1959
- A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. Tom Mitchell, 1997
- E: Training set made of training instances (samples)
- T: Test set • P: Such as accuracy

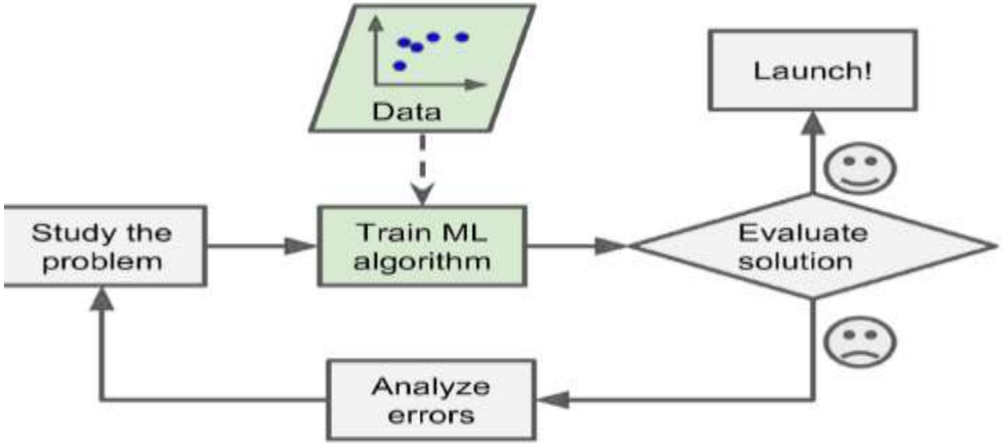
Why Use Machine Learning? Spam filter using traditional programming techniques

- Spam filter using traditional programming techniques



Why Use Machine Learning?

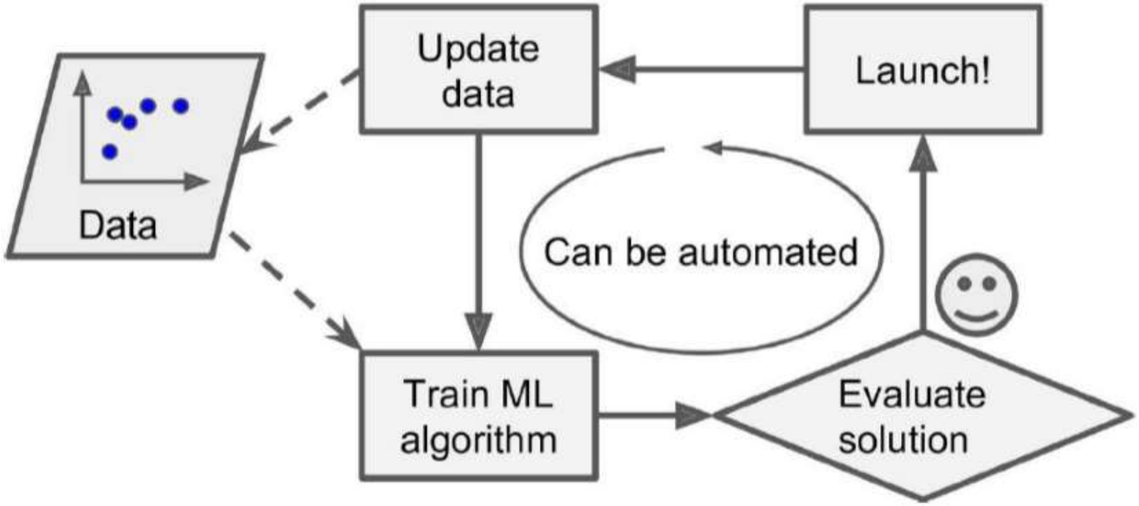
- Spam filter using machine learning techniques 1/2



RAWAN GHNEMAT

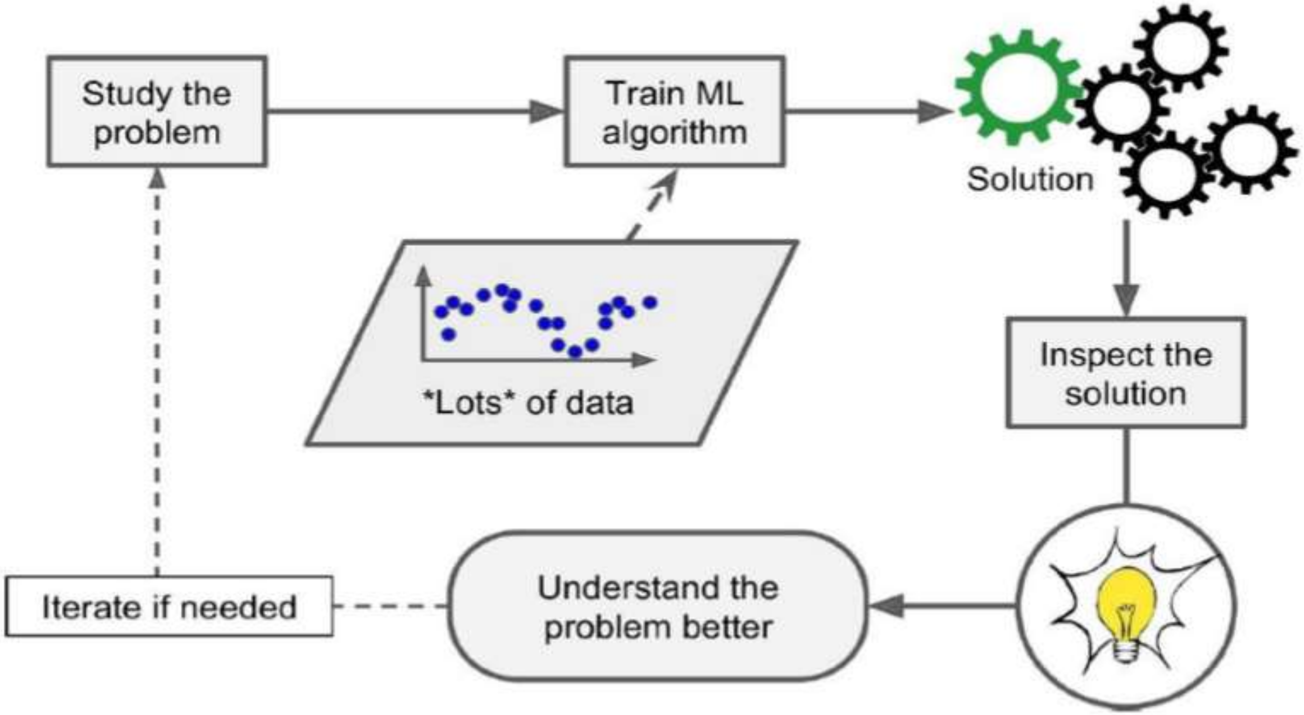
Why Use Machine Learning?

- Automatically adapting to change 2/2



RAWAN GHNEMAT

Why Use Machine Learning? ML can help humans learn (Data mining)



Ai
RAWAN GHNEMAT

Types of Machine Learning Systems

Involves human supervision?

1. Supervised learning

2. Unsupervised learning

3. Semi-supervised learning

4. Reinforcement learning

- Learns incrementally?

1. Batch learning 2. Online learning

Generalization approach

1. Instance-based learning

2. Model-based learning

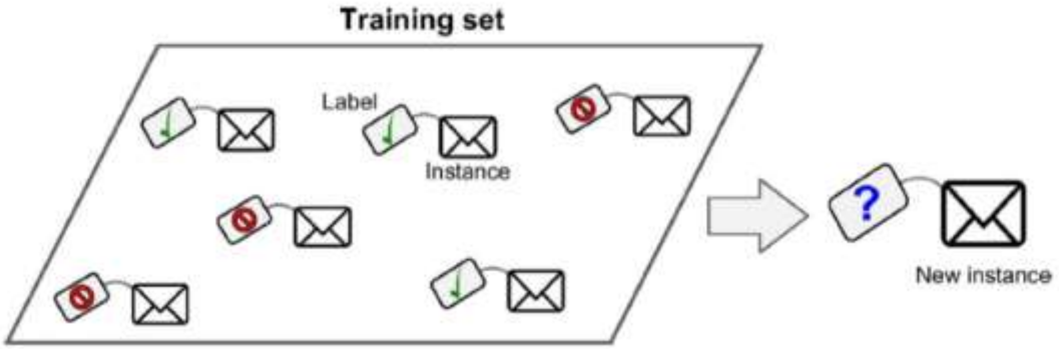
Ai

RAWAN GHNEMAT

Supervised Learning

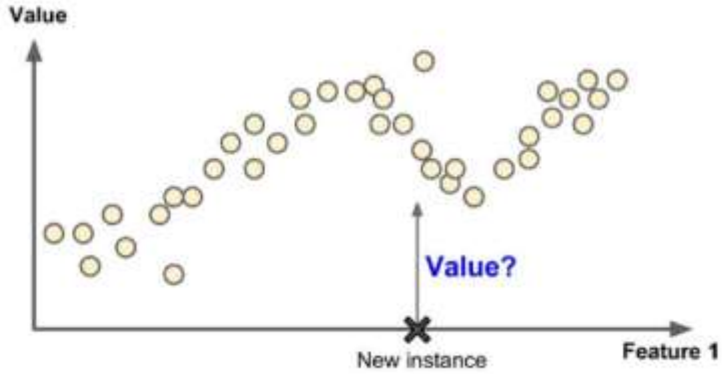
The training data you feed to the algorithm includes the desired solutions, called labels

Classification: finds the class, e.g., email type (spam or ham)



Supervised Learning

- Regression: finds the value, e.g., car price

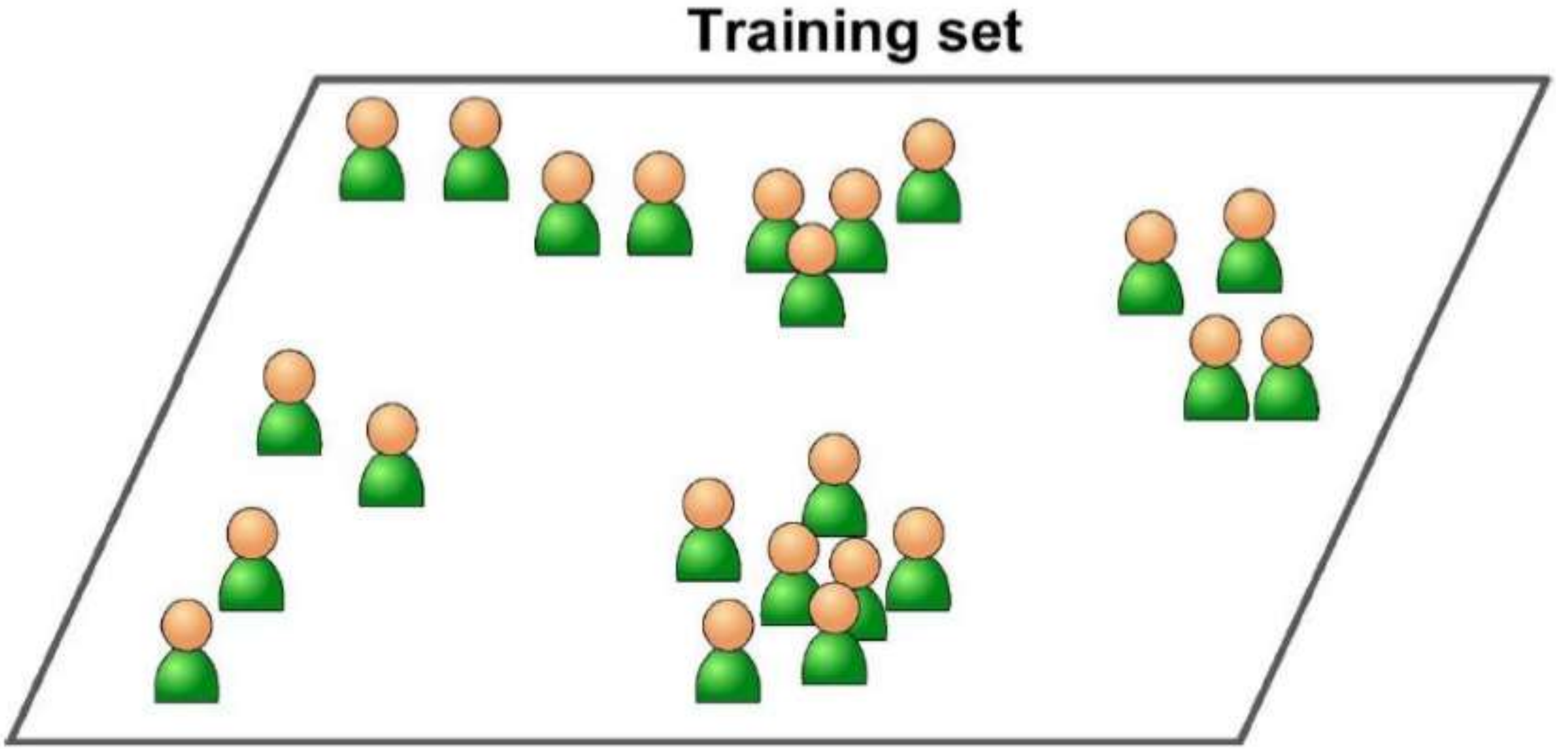


Ai
RAWAN GHNEMAT

1. Supervised learning algorithms

Algorithm	Type
k-Nearest Neighbors	Classification
Linear Regression	Regression
Logistic Regression	Regression
Support Vector Machines (SVMs)	Classification
Decision Trees	Classification
Random Forests	Classification
Neural Networks	Both

2. Unsupervised Learning



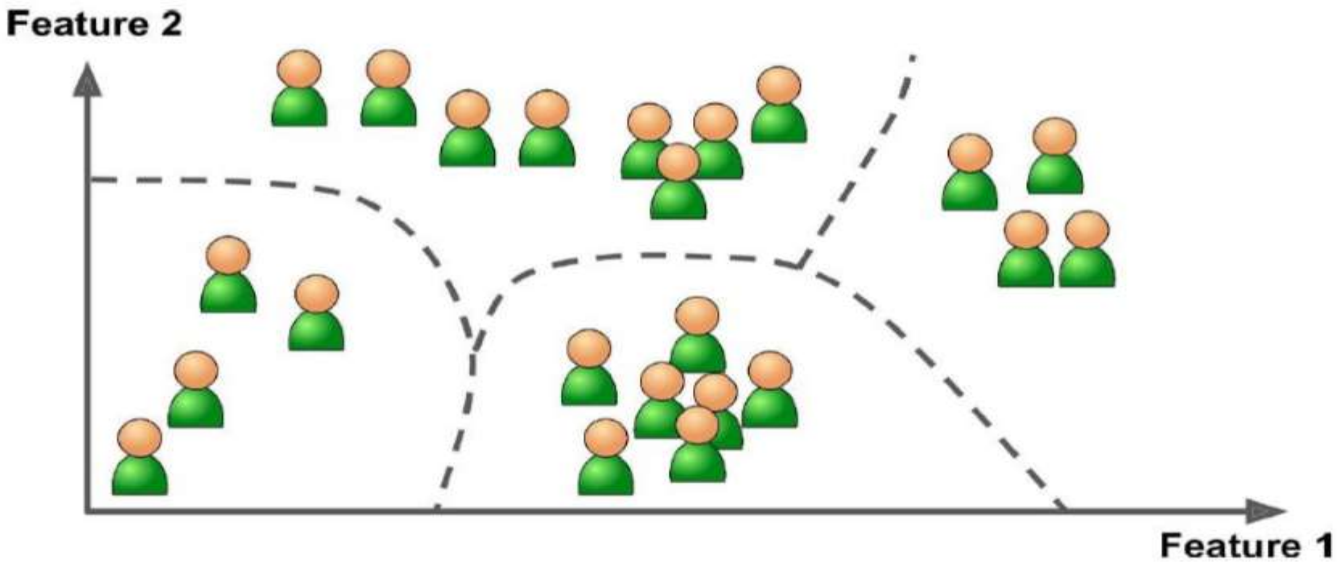
The training data is unlabeled.

2. Unsupervised learning algorithms

- Clustering
    - k-Means
    - Hierarchical Cluster Analysis (HCA)
    - Expectation Maximization
- Visualization and dimensionality reduction
    - Principal Component Analysis (PCA)
    - Kernel PCA
    - Locally-Linear Embedding (LLE)
    - t-distributed Stochastic Neighbor Embedding (t-SNE)
- Association rule learning
    - Apriori
    - Eclat

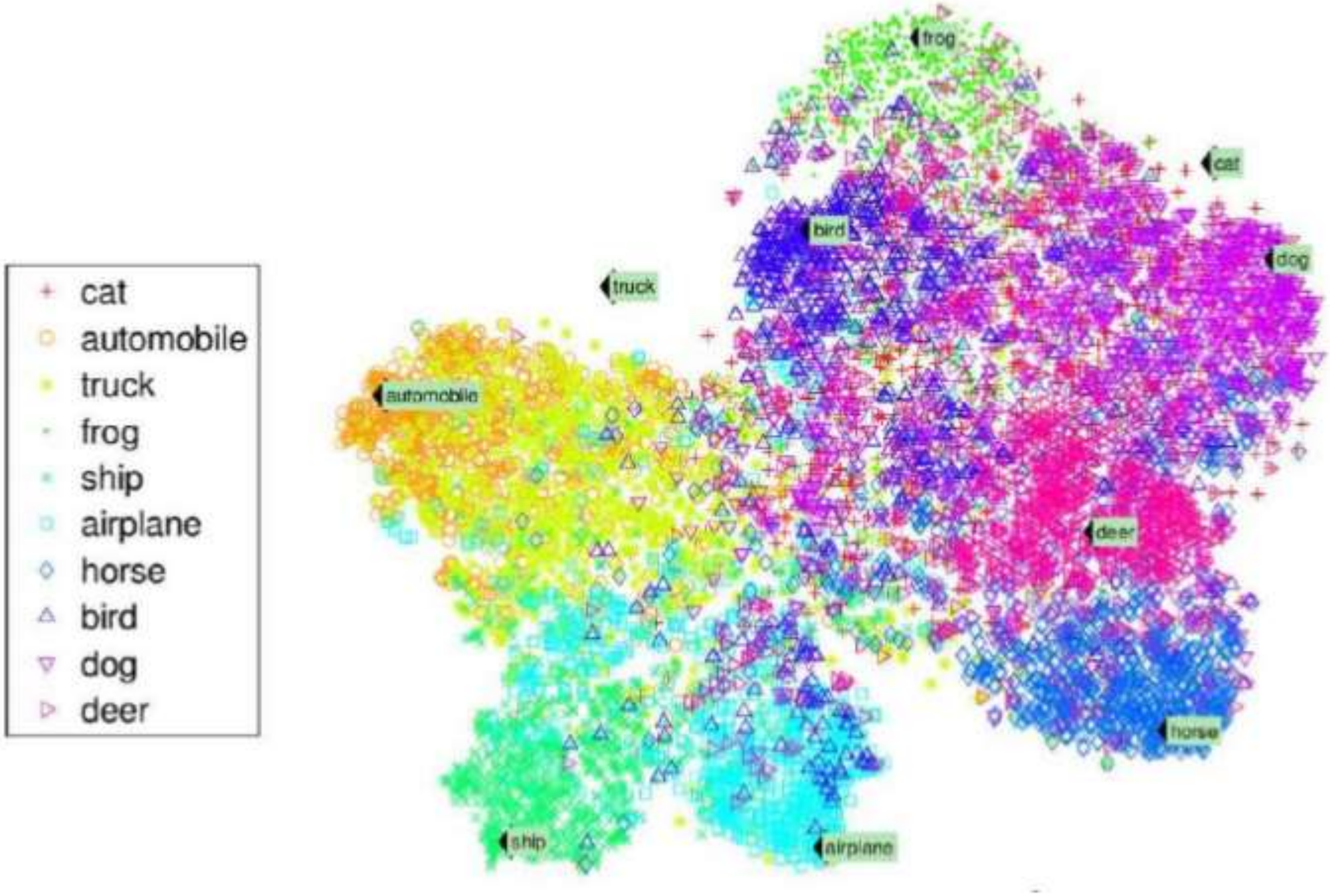
Clustering

2.a Clustering



RAWAN GHNEMAT

2.b Visualization



2.c Dimensionality Reduction

- The goal is to simplify the data without losing too much information.
- One way to do this is to merge several correlated features into one.

For example, a car’s mileage may be very correlated with its age, so the dimensionality reduction algorithm will merge them into one feature that represents the car’s wear and tear.

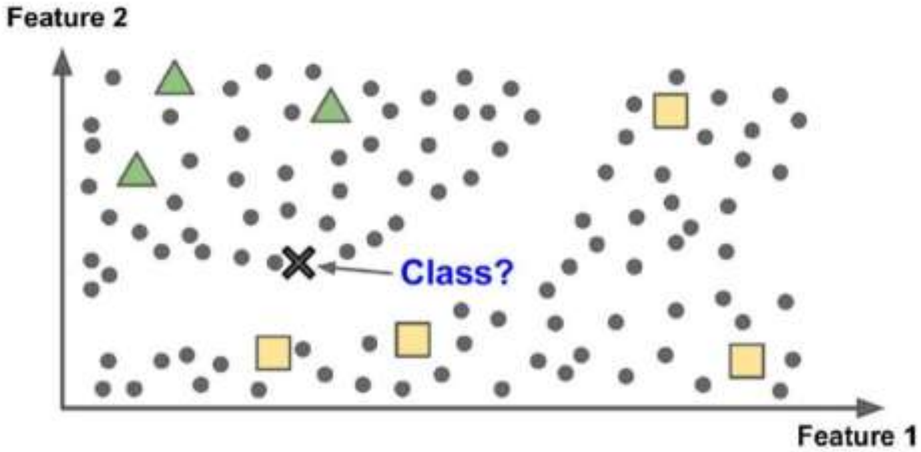
- Also called feature extraction.

2.e Association Rule Learning

- The goal is to dig into large amounts of data and discover interesting relations between attributes.
- For example, suppose you own a supermarket. Running an association rule on your sales logs may reveal that people who purchase barbecue sauce and potato chips also tend to buy steak. Thus, you may want to place these items close to each other

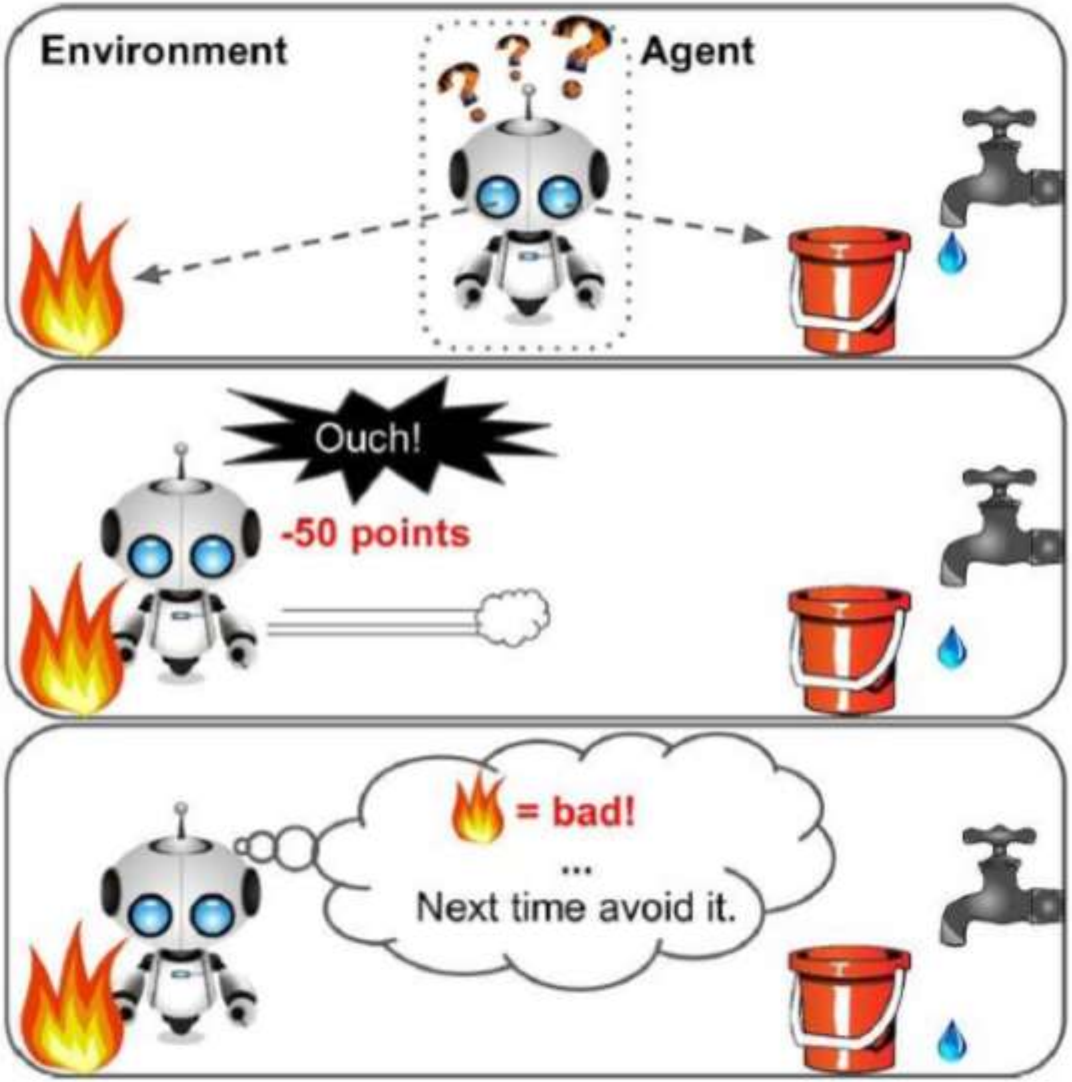
3. Semi-supervised Learning

- Partially labelled training data, usually a lot of unlabelled data and a little bit of labelled data. E.g., Google Photos.



RAWAN GHNEMAT

4. Reinforcement Learning

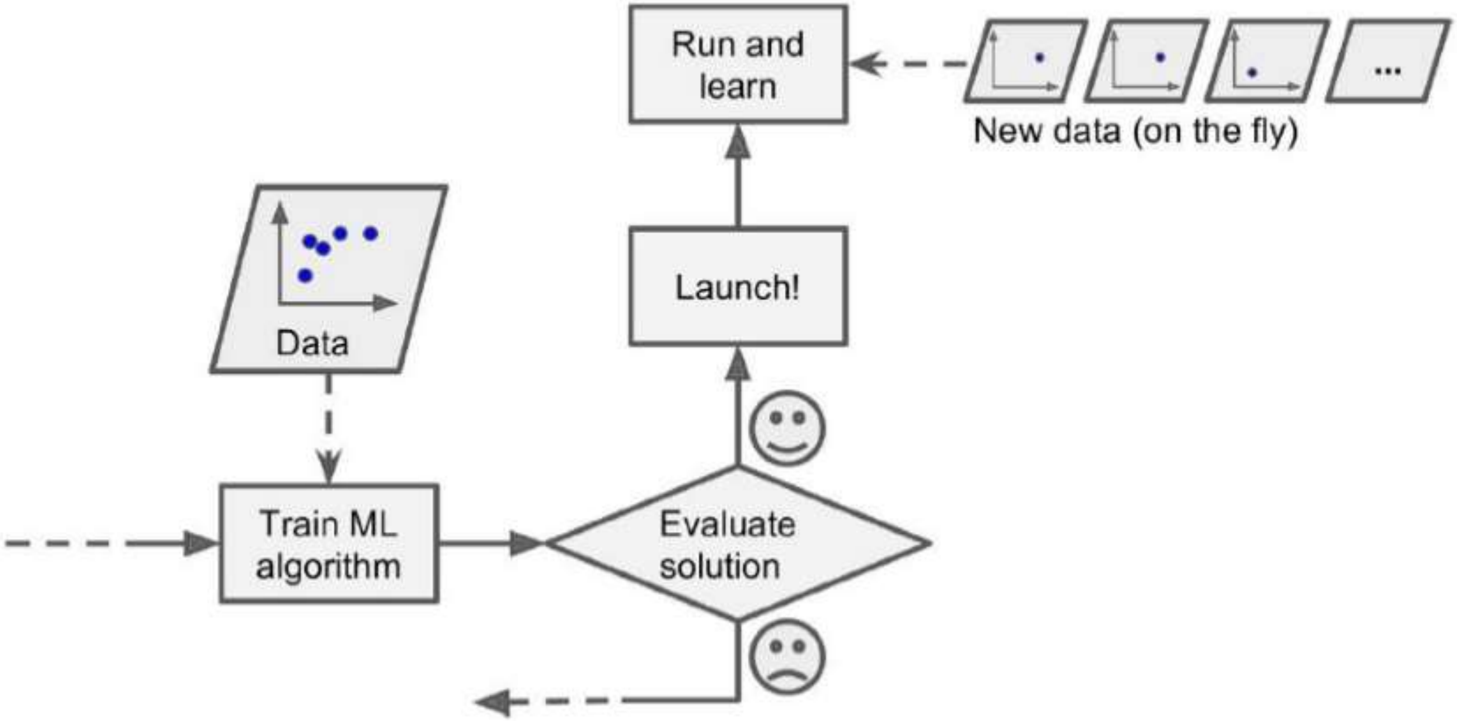


1. Observe
2. Select action using policy
3. Action!
4. Get reward or penalty
5. Update policy (learning step)
6. Iterate until an optimal policy is found

1. Batch (offline) Learning

- Must be trained using all the available data.
- This will generally take a lot of time and computing resources, so it is typically done offline.
- First the system is trained, and then it is launched into production and runs without learning anymore; it just applies what it has learned.

Online Learning Examples: Stock prices, huge data



Main Challenges of Machine Learning Data

- 1. Insufficient quantity of training data
- 2. Non-representative
- 3. Poor-quality data that contains:
- • Errors •
- Outliers
- Noise
- 4. Irrelevant features: Need feature engineering: • Feature selection: selecting the most useful features. • Feature extraction: combining existing features to produce a more useful one. • Creating new features by gathering new data.

Main Challenges of Machine Learning Algorithm

1. Over-fitting the training data

- Regularization constrains the model's

hyper parameters to make it simpler and reduce the risk of over-fitting

2. Under-fitting the training data

3.Computer recourses

Ai

RAWAN GHNEMAT

Testing and Validating

- Split your data into two sets (cross validation): • The training set (80%)

  The test set (20%)

- Evaluate:

  The training error

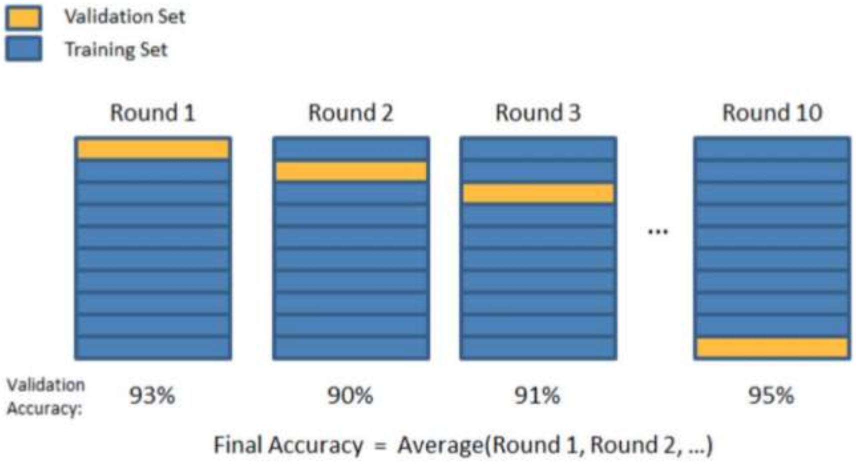
  The generalization error

- If the training error is low but the generalization error is high, it means that your model is overfitting the training data.

- When the ML algorithm is iterative, often we use a third set: validation set.

Cross Validation

- In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples



Summary

- ML is about making machines get better at some task by learning from data, instead of having to explicitly code rules.
- Types of ML systems: supervised or not, batch or online, and instance-based or model-based.
- A model-based algorithm tunes some parameters to fit the model to the training set, and then hopefully it will be able to make good predictions on new cases.
- An instance-based algorithm learns the examples by heart and uses a similarity measure to generalize to new instances. • The system will not perform well if your training set is too small, not representative, noisy, or polluted with irrelevant features.
- Your model needs to be neither too simple (under-fit) nor too complex (over-fit)

Exercises

- How would you define Machine Learning?
- What is a labelled training set?
- Can you name four common unsupervised tasks?
- What type of Machine Learning algorithm would you use to allow a robot to walk in various unknown terrains?
- What type of algorithm would you use to segment your customers into multiple groups?
- What is an online learning system?
- What is the difference between a model parameter and a learning algorithm’s hyper parameter?
- If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?
- What is the purpose of a validation set?

Project-Based Learning for Data Scientists

- https://towardsdatascience.com/project-based-learning-for-data-scientists-df6a8f74e4a1

Project Proposal by 20/12/2021

- One to two-page proposal
- Specify problem
- Specify sample size and source
- Structure

– Title

– Student name

– Problem definition

– Data description

– Samples

project Report by15/1/2022

- Four to 8-page report
- The introduction, include – Motivation – Problem definition – Literature review
- Describe your data and development environment.
- Describe any pre-processing, feature extraction and selection, techniques used, and post-processing.
- Give results and comments
- Give conclusions (work done, main results, future work)
- Include your source code in an appendix after the list of references.

Python Resources

- Official documentation: https://docs.python.org
- Tutorials:

https://www.learnpython.org/

- Python Books :

1. A Whirlwind Tour of Python, by Jake VanderPlas, https://www.oreilly.com/programming/free/files/awhirlwind-tour-of-python.pdf (short)
2. Python for Everybody, by Charles R. Severance, https://py4e.com/book.php (medium)
3. Fundamentals of Python Programming, by Richard L. Halterman,http://python.cs.southern.edu/pythonbook/pythonbo ok.pdf(long)

Kaggle

- https://www.kaggle.com/

Google Colab

- Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud. With Colaboratory you can write and execute code, save and share your analyses, and access powerful computing resources, all for free from your browser.

https://colab.research.google.com/

- Check the introduction in

https://colab.research.google.com/notebooks/intro.ipynb

Next lecture :Python Basics

Quick Python Syntax

- Variables and Objects
- Operators
- Built-In Types: Simple Values
- Built-In Data Structures
- Control Flow
- Defining and Using Functions
- Objects and Classes
- Errors and Exceptions
- Iterators
- List Comprehensions
- Generators