DS

Intro to ML

Intro - intro link, (YT), intro video

Machine learning is behind chatbots and predictive text, language translation apps, the shows Netflix suggests to you, and how your social media feeds are presented. It powers autonomous vehicles and machines that can diagnose medical conditions based on images.

When companies today deploy artificial intelligence programs, they are most likely using machine learning. Machine learning is a subfield of artificial intelligence that gives computers the ability to learn without explicitly being programmed.

What is machine learning?

Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems.

Machine learning starts with data. The data is gathered and prepared to be used as training data, or the information the machine learning model will be trained on. The more data, the better the program. From there, programmers choose a machine learning model to use, supply the data, and let the computer model train itself to find patterns or make predictions.

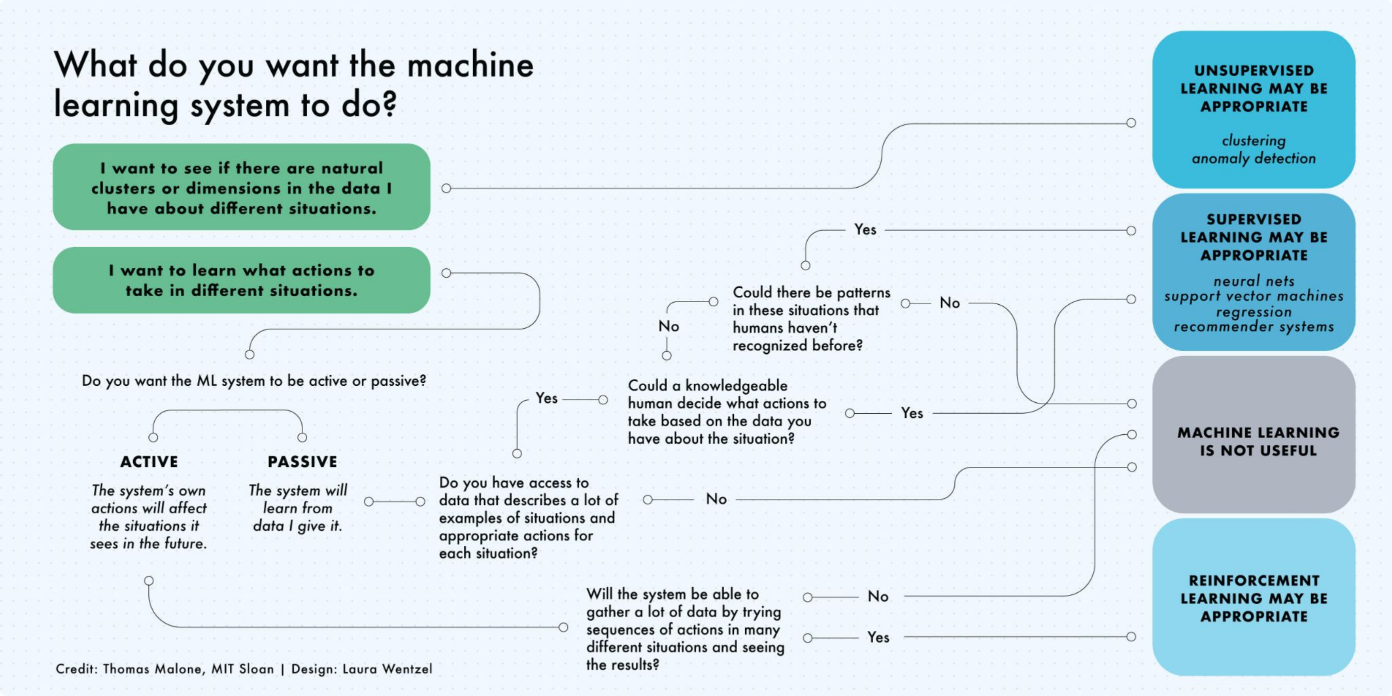
Some data is held out from the training data to be used as evaluation data, which tests how accurate the machine learning model is when it is shown new data. The result is a model that can be used in the future with different sets of data.

The function of a machine learning system can be descriptive, meaning that the system uses the data to explain what happened; predictive, meaning the system uses the data to predict what will happen; or prescriptive, meaning the system will use the data to make suggestions about what action to take

There are **three subcategories** of machine learning:

- **Supervised** machine learning models are trained with labeled data sets, which allow the models to learn and grow more accurate over time. For example, an algorithm would be trained with pictures of dogs and other things, all labeled by humans, and the machine would learn ways to identify pictures of dogs on its own. Supervised machine learning is the most common type used today.
- **Unsupervised** machine learning, a program looks for patterns in unlabeled data. Unsupervised machine learning can find patterns or trends that people aren't explicitly looking for. For example, an unsupervised machine learning program could look through online sales data and identify different types of clients making purchases.
- **Reinforcement** machine learning trains machines through trial and error to take the best action by

  establishing a reward system. Reinforcement learning can train models to play games or train autonomous vehicles to drive by telling the machine when it made the right decisions, which helps it learn over time what actions it should take.


What do you want the machine learning system to do?
I want to see if there are natural clusters or dimensions in the data I have about different situations.
I want to learn what actions to take in different situations.
Do you want the ML system to be active or passive?
ACTIVE
The system's own actions will affect the situations it sees in the future.
PASSIVE
The system will learn from data I give it.
Do you have access to data that describes a lot of examples of situations and appropriate actions for each situation?
Yes
No
Could a knowledgeable human decide what actions to take based on the data you have about the situation?
No
Yes
Could there be patterns in these situations that humans haven't recognized before?
Yes
No
Will the system be able to gather a lot of data by trying sequences of actions in many different situations and seeing the results?
No
Yes
UNSUPERVISED LEARNING MAY BE APPROPRIATE
clustering
anomaly detection
SUPERVISED LEARNING MAY BE APPROPRIATE
neural nets
support vector machines
regression
recommender systems
MACHINE LEARNING IS NOT USEFUL
REINFORCEMENT LEARNING MAY BE APPROPRIATE

Credit: Thomas Malone, MIT Sloan | Design: Laura Wentzel

It takes a strong focus on statistical, mathematical and problem solving skills to get ahead in the field of Machine Learning.

Problem solving approach:

1. Recommender systems
2. Security system

Programming

:The basics of programming:

1. Data structures and algorithms
2. Basic python programming: (link)(Anaconda set-up)(basic installation)
    1. Modules
    2. Arithmetics
    3. Functions
    4. Strings
    5. Exceptions (Error Handling - Python 3 Programming Tutorial p.9)
    6. Lists

7. Tuples, sets (Tuples, Strings, Loops - Python 3 Programming Tutorial p.2)
8. Dictionaries
9. Control flow
10. Regular expressions
11. Generators and iterators
12. Lambda (link)
13. Object-oriented programming

Database

SQL Resource:(link)

1. Create table and insert
2. Update, delete, select, groupby, order by, and join etc queries
3. Nested queries or subqueries
4. Query optimization
5. Introduction to NoSQL

Mongo Resource

1. TBD

Data Handling Libraries

1. Introduction to Numpy (https://www.youtube.com/watch?v=Rbh1rieb3zc - hindi)
2. pandas (link,Python Pandas Tutorial (Part 1): Getting Started with Data Analysis - Installation and Loading Data)

Visualizing Data

1. Matplotlib: (matplotlib, seaborn) - (yt)
   1. Basics : (https://www.youtube.com/watch?v=UO98lJQ3QGI&list=PL-osiE80TeTvipOqomVEeZ1HRrcEvtZB_&index=1)
   2. Bar Charts (https://www.youtube.com/watch?v=nKxLfUrkLE8&list=PL-osiE80TeTvipOqomVEeZ1HRrcEvtZB_&index=2)
   3. Scatter Plots : https://www.youtube.com/watch?v=zZZ_RCwp49g&list=PL-osiE80TeTvipOqomVEeZ1HRrcEvtZB_&index=7
2. Seaborn https://www.youtube.com/watch?v=TLdXM0A7SR8

Basics of Linear algebra

:

1. Scalar, vector, matrices, and tensors (Linear Algebra-What is Scalar and Vectors And Its Practical Applications In Machine Learning? ☆☆☆☆☆☆)
2. Linear dependence and span
3. Norms and determinants
4. Eigen decomposition

Basics of probability

1. Dependence and independence
2. Random variables
3. Conditional probability
4. Bayes’s theorem
5. Normal distribution
6. Central limit theorem
7. Chain rule of conditional probability
8. Expectation, variance, and covariance
9. Marginal probability

Basics of Statistics

1. Describing a single set of data - distribution, central tendency, dispersion etc
2. Correlation
3. Causation
4. Statistical hypothesis testing - ANOVA, p-values, t-test etc
5. Confidence intervals
6. Running A/B testing
7. Bayesian inference

Multi Variate

Getting and understanding Data

1. Reading files - basic text files, delimited files (CSV)
2. Scraping the web

1. Using APIs -- advanced
2. Exploring the data - 1D, 2D, and multi dimensional data

1. Cleaning and parsing - finding missing data, outliers etc (link)

1. Rescaling - MinMax normalization or Z-normalization

1. Data manipulation and augmentation
2. Dimensionality reduction - PCA, t-SNE etc (PCA)
3. Handling categorical data (link)
4. Partitioning datasets into training and testing datasets
5. Feature selection - L1, L2 regularization, and sequential feature selection algorithms
6. Feature engineering -- ??

Featurizing

Basic ML/AI/ Data Science concepts

(Andrew Ng) - https://arxiv.org/pdf/1609.04747.pdf

1. Linear regression - Simple regression, multiple regression, Housing dataset prediction (link)
2. Logistic regression - modeling class probabilities, learning the weights with logit cost function, training a logistic regression model with scikit-learn, tackling overfitting with regularization (link)(l2)

1. Support vector machine learning - Maximum margin concept, kernel methods for linearly inseparable data, implementations in scikit-learn (video)

1. Decision tree learning - Maximizing information gain, combining multiple decision trees via random forest, implementation (RF)
2. K-nearest neighbor - lazy learning, implementation (KNN)

Learning with ensemble learning

1. Combining classifier via voting - implementation of a simple majority vote classifier, using the majority vote principle to make predictions, evaluating and tuning the model
2. Bootstrapping - Bagging concept, applying bagging to classify Wine dataset
3. Adaptive boosting - How boosting works, Scikit-learn implementation
4. Hyper params - (link)

Learning Mechanism

1. **Gradient descent**
2. **Momentum**
3. **Adam**
4. **Back propagation**
5. **BPTT**
6. **Contrast divergence (Blog)**
7.

Clustering

1.

1. PCA (Blog) (dimensionality reduction)
2. K-nn (Blog), dynamic time warping
3. GMM (Blog)
4. Autoencoders (Book Link)
5. DBSCAN (Blog)
6. Agglomerative hierarchical clustering (Blog)

Adversarial and Reinforcement

1. **Adversarial learning**
    1. **GANs and its variants**

1. **Reinforcement learning (List of resources)**
    1. **Markov decision process**
    2. **Q-learning**

Multilayer Artificial Neural Networks

1. (from scratch): (yt)
    1. Training simple ML algorithm for classification - glimpse of perceptron, implementation from scratch, classification on IRIS dataset
    2. Modeling complex functions with ANNs - MLP architecture, activation functions, forward propagation, back propagation, gradient descent (l1,l2,l3)(SGD)
    3. Classifying MNIST dataset - obtaining dataset, implementing MLP,
    4. Training an MLP - computing the logistic cost function, training from backpropagation, convergence of the network

Model evaluation

1. Cross validation - holdout method, k-fold, and stratified
2. Debugging with learning and validation curves - RMS, MAE, MAP, MAPE etc
3. Fine-tune models with grid search - tuning hyperparameters (link)
4. Reading performance indices - Confusion matrix, precision, Recall, F1, ROC (confusion matrix, ROC,)

1. Raja (Book Link)
    1. Evaluation with Performance metrics - RMS, MAE, MAP, MAPE, Accuracy, Precision, recall, F1, Kappa, BLUE1-4, METEOR etc (Blog)
    2. Validation: Unbiased estimator, least squares, Cross entropy (Blog), Softmax, Likelihood, Convolution, Forgetting factors
    3. Hyperparameter selection - Cross validation, batch learning, stratified validation, overfitting, regularization, underfitting (Book Link, BookLink)

Explainable ML - SHAP, LIME (Blog)

Advanced Architecture & Concepts

1. **CNN : evolution of CNN architecture.**

2. **Encoder-Decoder Architecture : blog**
3. **Attention : blog**
4. **Transformer : blog**
5. **Evolution of GANs: CycleGAN, Style GAN etc blog**

VAE : blog

Deep learning

overall - (link)

The mechanics of tensorflow

1. Tensorflow ranks and tensors
2. Understanding tensorflow’s computational graphs
3. Placeholders in tensorflow
4. Variables in tensorflow
5. Building a regression model
6. Executing objects in tensorflow graph
7. Saving and restoring a model in tensorflow
8. Visualizing the graph with tensorboard

Deep Convolutional Neural Networks

1. Building blocks of CNN - Understanding CNN and learning feature hierarchies (video) - (link)
2. Convolution, padding, subsampling

3. Building a CNN - theory
4. Implementing a CNN with tensorflow
5. Image classification with the built CNN model

Modeling sequential data with recurrent neural networks - RNN

(LSTM,Video LSTM)

1. Introduction to sequential data - text, time series etc
2. Modeling the sequences -- Understanding the structure and flow of RNNs. computing activations in a RNN,(blog)
3. Performing sentiment analysis of IMDB movie review
4. Stock market prediction with RNN

MLOps and Engineering

1. **Engineering Best Practices for ML: https://se-ml.github.io/practices**
2. **Why ML OPS and how is it different from regular DevOPS? https://hackernoon.com/why-is-devops-for-machine-learning-so-different-384z32f1, https://towardsdatascience.com/ml-ops-machine-learning-as-an-engineering-discipline-b86ca4874a3f**
3. **ML OPS Principles: https://ml-ops.org/content/mlops-principles**
4. **General ML OPS Resource repository: https://ml-ops.org**
5. **Sample Data Science Project Directory structure: https://drivendata.github.io/cookiecutter-data-science/#directory-structure, https://github.com/cmawer/reproducible-model**
6. **Basic Task sheet for most Data science projects: https://towardsdatascience.com/task-cheatsheet-for-almost-every-machine-learning-project-d0946861c6d0**
7. **Model Monitoring, What, Why?: https://www.analyticsvidhya.com/blog/2019/10/deployed-machine-learning-model-post-production-monitoring/**
8. **Metrics used for Model Monitoring: https://towardsdatascience.com/mlops-model-monitoring-101-46de6a578e03**
9. **Out-of-Box solutions for Model monitoring: https://neptune.ai/blog/ml-model-monitoring-best-tools**

Practice Project Links

1. Basic python and statistics

Pima Indians :- https://www.kaggle.com/uciml/pima-indians-diabetes-database

Cardio Goodness fit :- https://www.kaggle.com/saurav9786/cardiogoodfitness

Automobile :- https://www.kaggle.com/toramky/automobile-dataset

2. Advanced Statistics

Game of Thrones:-https://www.kaggle.com/mylesoneill/game-of-thrones

World University Ranking:-https://www.kaggle.com/mylesoneill/world-university-rankings

IMDB Movie Dataset:- https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset

3. Supervised Learning

a) Regression Problems

How much did it rain :- https://www.kaggle.com/c/how-much-did-it-rain-ii/overview

Inventory Demand:- https://www.kaggle.com/c/grupo-bimbo-inventory-demand

Property Inspection predictiion:- https://www.kaggle.com/c/liberty-mutual-group-property-inspection-prediction

Restaurant Revenue prediction:- https://www.kaggle.com/c/restaurant-revenue-prediction/data

IMDB Box office Prediction:-https://www.kaggle.com/c/tmdb-box-office-prediction/overview

b) Classification problems

Employee Access challenge :- https://www.kaggle.com/c/amazon-employee-access-challenge/overview

Titanic :- https://www.kaggle.com/c/titanic
San Francisco crime:- https://www.kaggle.com/c/sf-crime
Customer satisfcation:-https://www.kaggle.com/c/santander-customer-satisfaction
Trip type classification:- https://www.kaggle.com/c/walmart-recruiting-trip-type-classification
Categorize cusine:- https://www.kaggle.com/c/whats-cooking

4. Unsupervised Learning

Vehicle Identification:- https://www.kaggle.com/c/st4035-2019-assignment-1

Open Data Sets

Find Open Datasets and Machine Learning Projects | Kaggle
Kaggle

OpenML
OpenML: exploring machine learning better, together.

The NLP Index
Quantum Stat

Time Series

Sources - World and regional statistics, national data, maps, rankings
Knoema