Machine Learning - Key Distinctions

Machine Learning vs Data Mining vs Statistical Modelling

ML vs Normal algorithms

- Machine Learning algorithms are a class of algorithms that performs tasks without being explicitly programmed. i.e, they learn from data.
- In contrast to normal algorithms, Machine Learning Algorithm's output is probabilistic and normal algorithm's output is deterministic.

ML vs Statistical modelling

- A **statistical model** is a mathematical **model** that embodies a set of **statistical** assumptions concerning the generation of sample data. i.e, It tries to discover the underlying relationship in data with respect to target and features.
- A machine learning model is a model concerned with prediction rather than inference.
- Machine learning algorithms can learn from n-number of observations,***going over each one by one. It accesses each observation and learns over time to predict the exact combination of actions that can help in achieving the end-goal.*** Usually, machine learning deals with data that has numerous attributes and a large number of observations.
- On the contrary, statistical models encompass ***series of assumptions that deal with a long lineage of observed and similar data.*** It does not rely on predictive accuracy like machine learning. In fact, you can define machine learning as model and assumptions checking without the statistical touch in it.
- Machine learning ***requires almost no human intervention because it is about enabling a computer to learn on its own from a large set*** of data without any set instructions from a programmer.
- That’s not how things work in statistical modeling. It is about ***mathematics equations and requires the modeler to accurately understand the relationship between two variables*** before they are fed into the data.
- Machine Learning algorithms need training and testing. The goal is to create the best predictor. We are not worried about the underlying relationship in the dataset. This process involves lot of fields such as statistics, linear algebra, optimization etc.
- Statistical modeling doesn't require training and testing. It assumes some properties of data and the model is evaluated to prove or disprove the assumption. We are more concerned about inference we can draw rather than prediction. Statistical model generally has a confidence interval for forecasting or prediction while machine learning models doesn't.
- Machine learning algorithms may involve less statistics as well, example: Decision Trees and its family, K nearest neighbors, Neural Networks, Agglomerative clustering or DBSCAN etc.

- Statistical Models include Linear Regression, Logistic Regression, ANOVA and its family, Factor Analysis, Discriminant Analysis etc.
- Most statistical models require hypothesis significance testing whereas machine learning models don't.
- Machine Learning borrows many models from statistics such as linear regression, logistic regression, LDA etc. But these differ in some ways. They involve training, testing and optimization.

ML vs Data Mining

- **Data Mining** is the process of discovering knowledge or underlying patterns in large data sets. Clearly, data mining is a task. These tasks ranges from classification, regression, clustering, association analysis etc.
- Machine learning is one way to do these tasks. We can use statistical models as well.
- Data Mining's goal is to get underlying patterns in data. To do this, we can use many methods.
- Generally its a pipeline involving data collection, preprocessing, exploratory analysis, model building and evaluation and result analysis/report generation.

Data Science

Data Science can mean many things. It is a science of dealing with data. This is an umbrella terms for all data related concepts such as Machine Learning, Data Mining, Statistics, Big Data etc.

Algorithms derived from Statistics

- Linear Regression
- Logistic Regression
- Support Vector Machines
- LASSO and Ridge Regression
- Naïve Bayes Classifier
- Expectation Maximization Algorithm (Gaussian Mixture Models)
- Hidden Markov Models (Reinforcement Learning)
- Monte Carlo Methods (Reinforcement Learning)

Algorithms derived from Linear Algebra

- Principle Component Analysis
- Singular Value Decomposition

List of Libraries for ML and Data Science in Python

Data Collection/Web Scraping/DataBase

- urllib
- beautifulsoup4
- scrapy
- sqlite3

Mathematical operations

1. math
2. sympy
3. statistics (Built in)
4. cmath
5. fractions
6. random

Numerical Computing and Data Exploration

1. NumPy
2. SciPy
3. Pandas
4. agate
5. bottleneck - Cython Numpy calculations
6. json

Visualization

1. Matplotlib
2. Seaborn
3. bokeh
4. plotly
5. pydot

Machine Learning and Statistical Modelling

1. scikit-learn
2. statsmodels
3. xgboost
4. pyearth
5. pybrain
6. pyspark
7. pgmpy
8. Eli5 - ML result visualizer

Deep Learning

1. tensorflow

2. keras
3. pytorch
4. theano
5. fast.ai
6. NeuPy

Reinforcement Learning

1. OpenAI Gym
2. Google Dopamine
3. Deepmind TRFL (truffle)
4. keras-rl

Natural Language Processing/Text Processing

1. NLTK
2. SpaCy
3. GenSim
4. Stanford CoreNLP
5. Textblob
6. re - built in for pattern matching using regular expressions
7. string

Image Processing and Computer Vision

1. OpenCV
2. Scikit-Image
3. PIL
4. simpleCV

Saving Models

1. Pickle
2. Joblib

List of Libraries for ML in R

Data Collection/Web Scraping/DataBase

1. rvest - webscraping
2. rjdbc - database
3. rsqlite
4. rcrawler

Mathematical operations

Most are built in

Numerical Computing and Data Exploration

1. dplyr
2. tidyr
3. data.table
4. lubridate - date
5. jsonlite
6. Matrix

Visualization

1. plot
2. ggplot2
3. corrplot
4. lattice
5. plotly
6. ggvis
7. datatables //html tables
8. rcharts
9. Esquisse
10. shiny
11. rbokeh

Machine Learning and Statistical Modelling

1. mlr
2. caret
3. prophet (time series)
4. randomForest
5. rpart -decision trees
6. lm - linear regression (built in)
7. glm (logistic regression) (built in)
8. e1071 - SVM,Naive Bayes
9. knn
10. cluster
11. stats

12. Rweka

Deep Learning

1. keras
2. deepnet
3. mxnet
4. h2o
5. neuralnet
6. darch
7. nnet

Reinforcement Learning

1. nproellochs/ReinforcementLearning

Natural Language Processing/Text Processing

1. quanteda
2. stringr
3. nlp
4. opennlp
5. tm
6. languageR

Image Processing and Computer Vision

1. magick
2. imager

Saving Models

saeRDS() readRDS() built in

List of Libraries for ML in other languages

1. tensorflow.js (javascript)
2. mljs (javascript)
3. keras.js (javascript)
4. brain.js (javascript)
5. D3.js, Chart.js, Plotly.js (javascript)
6. Deeplearning4j (Java)
7. weka (Java)
8. javaML (Java)
9. Mahout (Java)
10. Spark (Java)
11. Darknet (C/C++)
12. Caffe (C++)
13. CNTK (C++)
14. tensorflow (C++)
15. Dlib (C++)
16. openNN (C++)
17. ML.Net (C#)
18. mlc++

List of Algorithms

Algorithm	Type	Python Package	R Package
K Nearest Neighbors	Instance Based	sklearn.neighbors	knn,caret
Learning Vector Quantization	Instance Based	sklearn-lvq 1.1.0 or NeuPy	caret
Self Organizing Map	Instance Based	NeuPy	kohonen
Logistic Regression	Regression	sklearn.linear_model	glm,caret
Linear Regression	Regression	sklearn.linear_model	lm
Lasso Regression	Regression	sklearn.linear_model	glm
Ridge Regression	Regression	sklearn.linear_model	glm
Ordinary Least Square Regression	Regression	sklearn.linear_model	lm
MARS	Regression	pyearth.Earth.Earth()	mars
Elastic Net	Regression	sklearn.linear_model	lm
LARS Regression	Regression	sklearn.linear_model	lm
Naive Bayes	Probability based	sklearn.naive_bayes	e1071,caret
Decision Trees	Tree Based	sklearn.trees	caret,rpart
Bagging	Tree Based	sklearn.ensemble	caret
Boosting	Tree Based	sklearn.ensemble	caret
Random Forests	Tree Based	sklearn.ensemble	caret,randomforest

Algorithm	Type	Python Package	R Package
SVM	Decision Boundary	sklearn.svm	e1071
Linear Discriminant Analysis	Linear Classifier	sklearn.discriminant_analysis	caret
PCA	Dimensionality Red	sklearn.decomposition	caret
SVD	Dimensionality Red	sklearn.decomposition	base
EFA	Dimensionality Red	sklearn.decomposition	
Neural Networks (MLP)	ANN	keras/tensorflow	keras
Convolutional Neural Nets	ANN	keras/tensorflow	keras
Deep Belief Nets	ANN	keras/tensorflow	keras
RecurrentNeural Nets	ANN	keras/tensorflow	keras
Restricted Boltzmann Machines	ANN	keras/tensorflow	keras
AutoEncoders	ANN	keras/tensorflow	keras
Long Short Term Memory	ANN	keras/tensorflow	keras
Recursive Neural Tensor Nets	ANN	keras/tensorflow	keras

List of Data Mining Softwares (Free)

1. Orange
2. Weka
3. RapidMiner
4. KNIME
5. Tableau (Visualization)
6. OpenRefine

List of Useful Websites

1. Kaggle
2. UCI ML Repo
3. towardsdatascience
4. medium
5. analyticsvidhya
6. machine learning mastery
7. pyimagesearch
8. fast.ai
9. datasciencecentral
10. subreddits
11. datacamp community

List of MOOCS

Free Open Courses with certifications :

- cognitiveclass.ai : An IBM Initiative which provides free courses with certifications (on passing the tests with 70%)on fields such as Data Science, Big Data, Block Chain, IBM Cloud and Machine Learning. It also includes courses for Python, Scala and R. You'll also get verifiable badges by IBM upon completion of a learning path.
- IBM Skills Gateway and Skills Network : Lot of free courses (some are paid) on different technologies.
- Udemy: A lot of free courses are available on Udemy on almost any topic. Although the quality of courses differ greatly on each course. So, be careful while choosing the course you want. You can use the ratings to decide which course you want to pursue. A Certificate of completion will be given by Udemy after completing the course.
- DataCamp: DataCamp offers free courses on Python, R, SQL, Git and Shell. It also gives a hands on approach to the concepts rather than a normal series of videos offered by others. It also has a good community for Data Science and has lot of other resources such as Cheat Sheets, Tutorials, Open Courses and Podcasts on Data Science.
- Freecodecamp.org: It offers 6 certifications upon completing courses and projects. Most of the courses revolve around web technology. It also has a blog which provides a lot of tutorials and articles and also a Youtube channel which has a lot of tutorials and open courses.
- Oracle DevGym: An Oracle initiative which provides free SQL courses with certifications and also provides a lot of competitions and quizzes
- Saylor Academy: Provides a lot of courses in a lot of domains. There are many computer science courses as well.
- Microsoft Learn: Free Azure Courses from Microsoft with certifications. They also plan to add other courses later.
- BitDegree: Lot of Free courses with certifications available.
- Google: Google Provides free digital marketing courses with certifcation. They also provide some open courses without any certificates.
- Gymnasium: Free Courses with certifcations
- SoloLearn: It is a fun app to learn programming. It does offer certificates on completion, but it is the knowledge and skill which is more important.
- OpenLearn: Free Courses from open university UK
- School of AI: A non profit initiative by Siraj Raval to teach world about AI and Machine Learning.
- CodeAcademy: Brief and Interactive courses on various computer science topics plus you'll get badges to show off.

Free Open Courses without certifications :

There are a lot of open courses which do not provide any certifications. But ultimately, it is the knowledge that matters and not the certification. Certifications are just documents that provide some "proof". But this proof needs to be justified by you. Internet has a lot of free resources that one can utilize and can improve his/her skill.

I cannot possibly list all the resources available but these are the ones which I am familiar with.

- Coursera: All Courses on Coursera are free if you do not want any certification. It also provides you with an option to upgrade and ger certificate. It also provides financial aid if you cannot afford the certification price.
- edx: Similar to Coursera, edx provides a free access to course contents if you do not need any certificates. You can later upgrade and ger certified if you meet the criteria.
- Khan Academy: Khan Academy has a lot of free courses on Mathematics, Science and Programming
- MIT OCW: MIT's Open courseware provides a lot of free resources for anybody who wants to learn.
- Youtube: There are a lot of resources on Youtube that I cannot possibly mention. Some of my favorite Youtube resources are CrashCourse (Absolutely fun way to learn), csdojo, Google's tutorial videos, Siraj Raval, Sentdex.
- Udacity: Offers few free courses from Industry's giants such as google.
- NPTEL: Offered by IITs of India. The courses are free. If you want certifications, you'll need to clear a written exam with a nominal fee (only available in India at the moment)
- OSSU Learning Curriculum: https://github.com/ossu
- Analytics Vidya: Offers some free courses and lot of articles for Data Science and Machine Learning
- towardsdatascience.com: A blog with lots of tutorials and articles about data science.

List of cloud resources for ML

1. Google Colaboratory
2. Kaggle Kernels
3. cognitiveclass labs