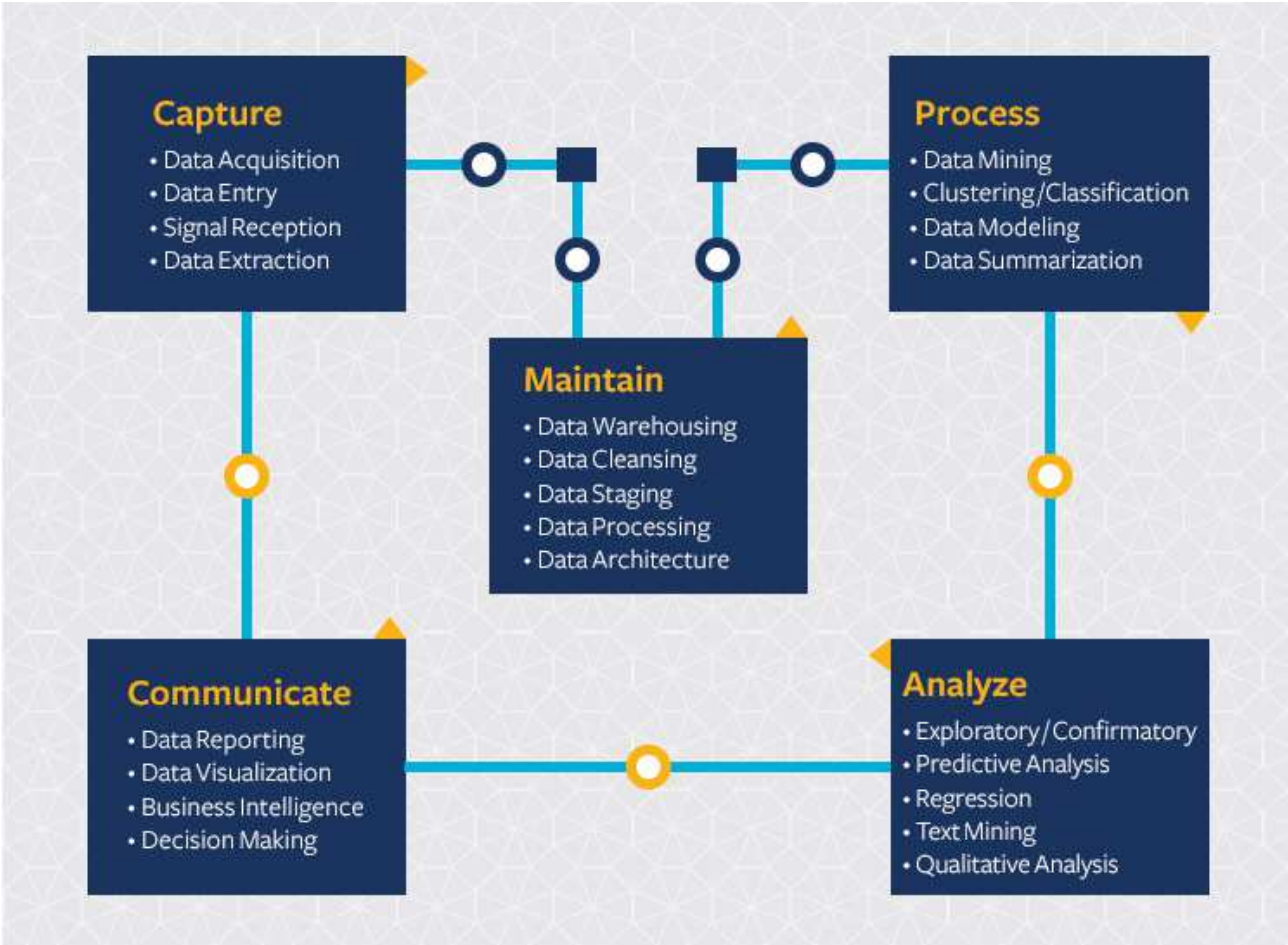
Machine Learning

Introduction to Data Science and Machine Learning

Data Science

- What is Data science?
  - Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to **extract knowledge and insights from structured and unstructured data**
  - Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.
  - Ref: https://en.wikipedia.org/wiki/Data_science

Data Science



- Ref: https://datascience.berkeley.edu/about/what-is-data-science/

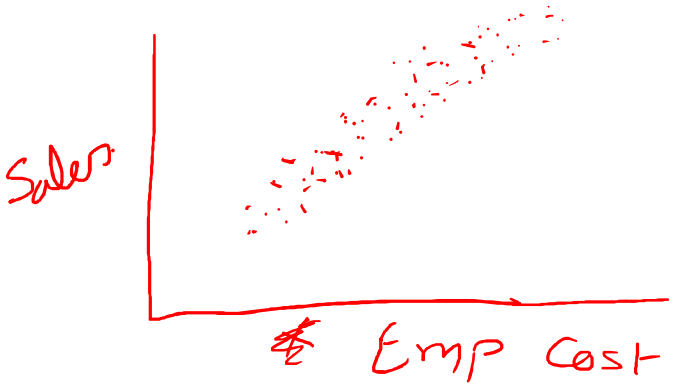
Data Science

- Why is there a sudden increased interest Data Science?
  - Burst in Data – Internet, electronic devices
  - Technological advancements – data storage, processing power, cloud based storage and computing
  - Businesses looking to use data to gain competitive advantage
  - “The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that’s going to be a hugely important skill in the next decades.”

    - Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

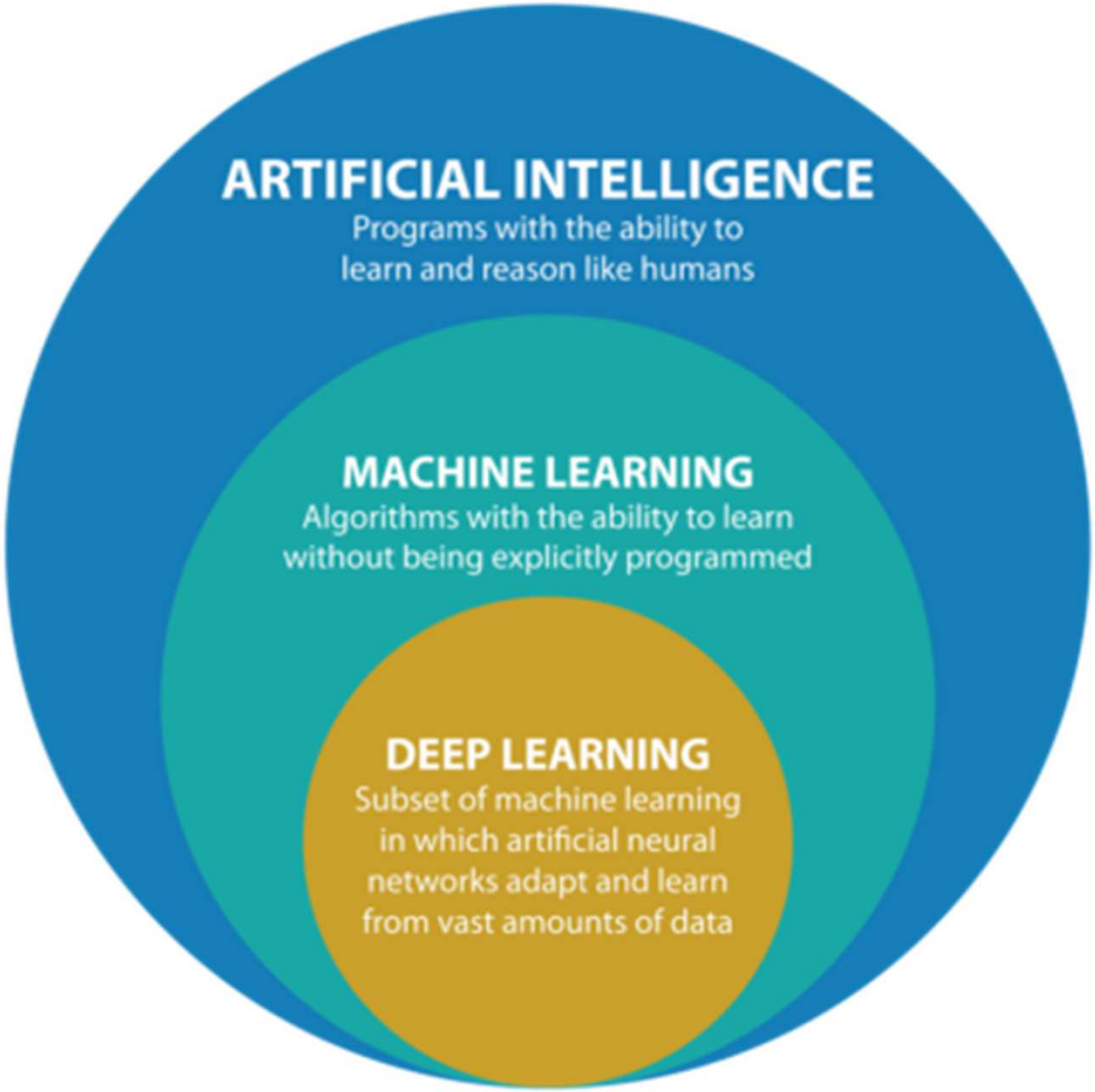
Machine Learning

- What is Machine Learning?

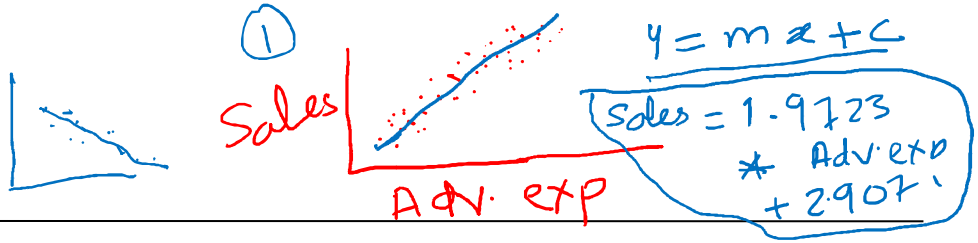


- Machine learning is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively **without using explicit instructions** i.e. **without being explicitly programmed, relying on patterns instead**

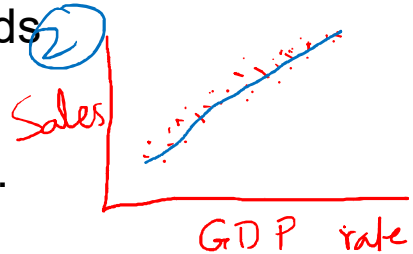
AI, ML, Deep Learning



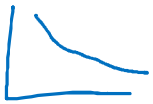
Machine Learning



- Process of enabling computer to learn to do tasks (for example, prediction) based on well defined statistical and mathematical methods.



- The ability to do the prediction is built in form of a “model”.
- A model is the result of the learning process
- The model represents the process which generated the data used to build the model



- The more representative data is of the real world in which the process is executed, the better the model would be

Machine Learning

- How does machine learning work?
  - It searches through data to look for patterns
  - The patterns are expressed as statistical / mathematical structures, for example polynomial equations
  - These statistical / mathematical structures, which can be used to perform predictions, are called models

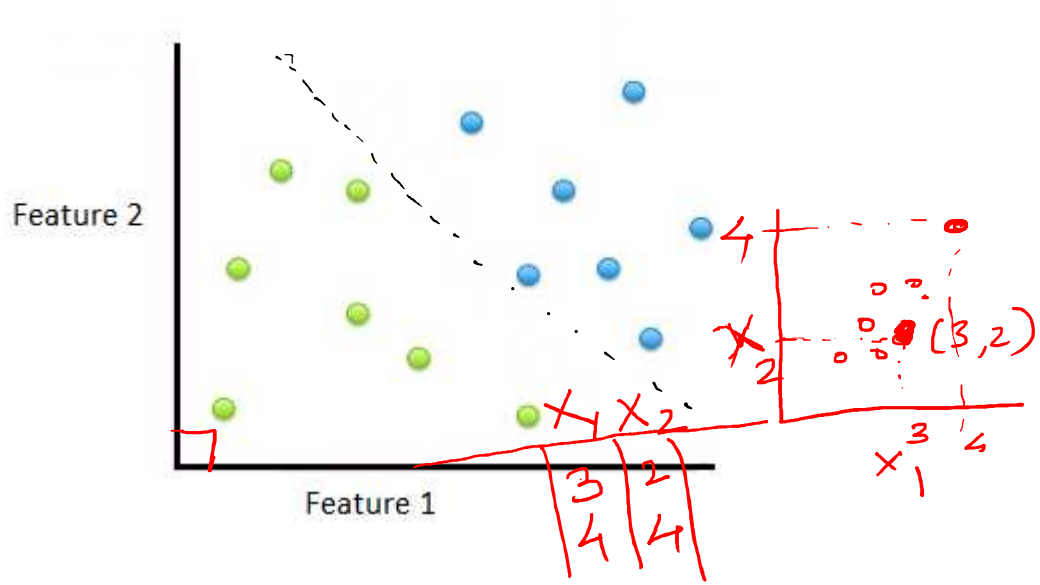
Use of Machine Learning

- Machine learning is useful when
  - Data patterns are too complex and constantly changing. E.g. weather forecasting
  - We find it hard to express our knowledge about patterns as a program. e.g. Character recognition
  - We do not readily have an algorithm to identify a particular pattern e.g. spam mail detection

Feature Space

- Each record represents data collected on various attributes
- These values, when plotted, are called feature space or mathematical space
- Following is an example of 2-dimensitonal feature spce

Feature 1	Feature 2	Class
2	1	Green
3.1	2.5	Green
8	7.2	Blue
3.5	2.9	Green
2.8	6	Green
6.8	5.5	Blue
.....	.....	.....
.....	.....	.....

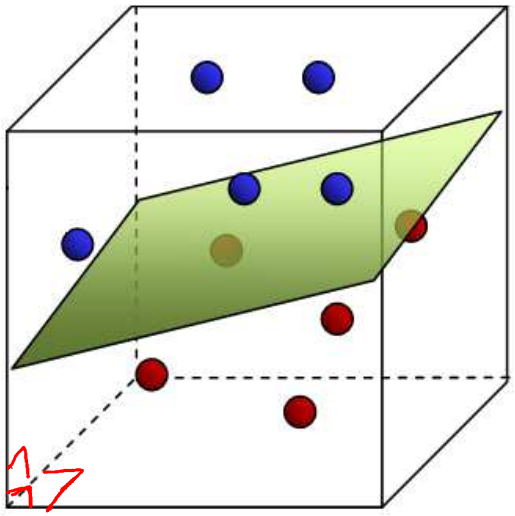


Feature Space

- In a feature space, each attribute becomes a dimension and each record becomes a point in the space

- Feature space can be 3-dimensional or multi-dimensional. In real world, typically there will be multi-dimensional feature space.

- Beyond 3-dimension, we cannot visualize the feature space and depend on statistical and mathematical concepts to derive meaning from it



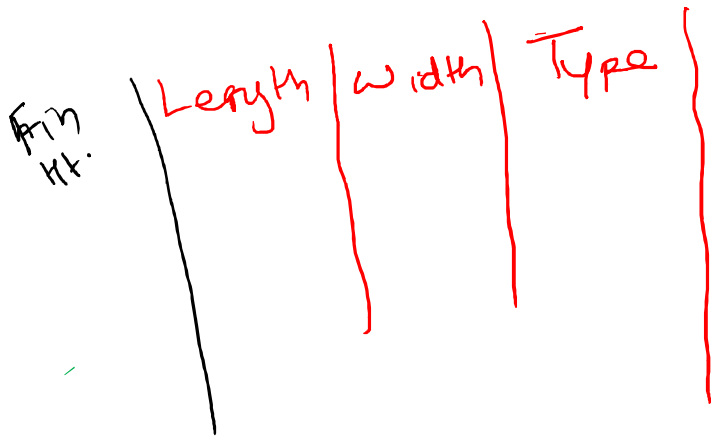
Terminology

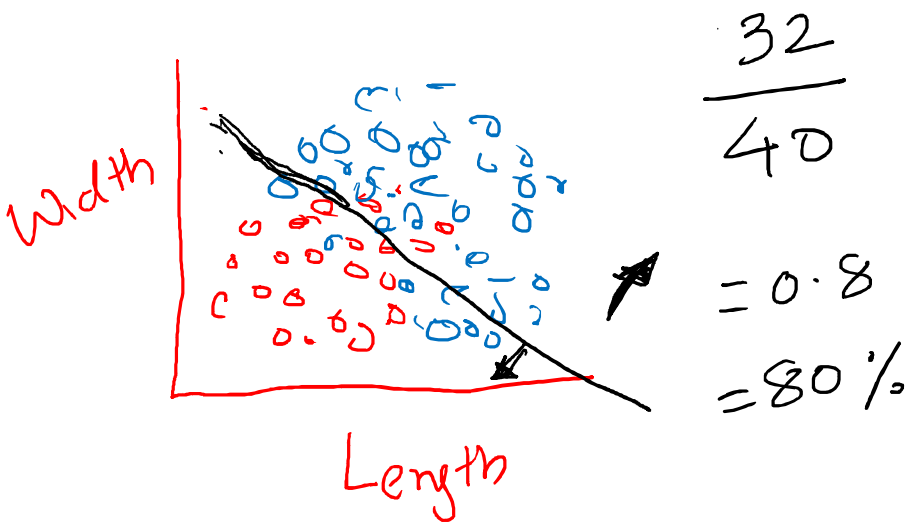
The value which we want to predict:

- Target variable, Dependant variable, Y, Predicted variable, Label

The values using which we will attempt to predict:

- Features, Dimension, Independent Variables, Xs, Predictor variable

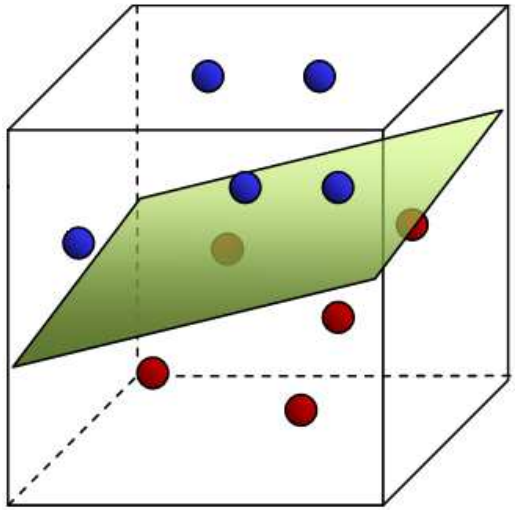




A Model

Classification : Accuracy = $\frac{\#\text{ of obs. correctly predicted}}{\text{Total \# of obs}} = \frac{10}{10} = 1 = 100\%$

- A plane shown in the diagram below is in example of a classification model
- This model attempts to classify data points as Blue or Red (e.g. diabetic or non diabetic)
- The equation of the place can be used to predict the classification of new records
- Thus, if we provide the three dimensions of a point, i.e. values of three attributes, then the model can predict classification of the point
- Proportion of the records that are correctly classified by a model decides accuracy of the model

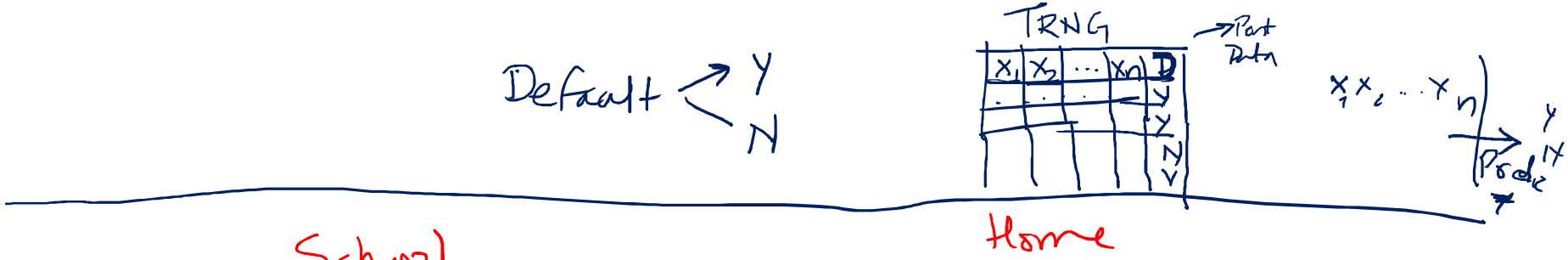


Machine Learning Categories

- Popular Machine Learning models:

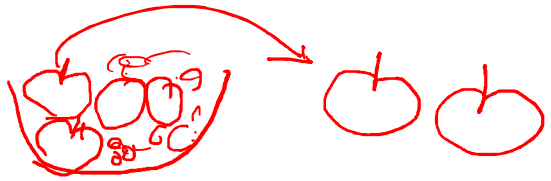
Supervised Learning	
Regression	**Classification**
Linear Regression Artificial Neural Network	K-nearest Neighbors Logistic Regression Decision Tree Naïve Bayes classifier Support Vector Machine Artificial Neural Network

Unsupervised Learning	
Cluster Analysis	**Dimension Reduction**
K-Means Clustering Hierarchical Clustering	Principle Component Analysis



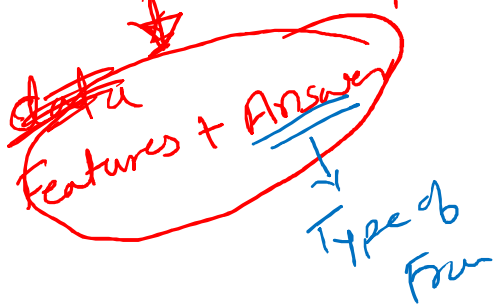
School

Learnt



① Learning process

Training Data → Find pattern

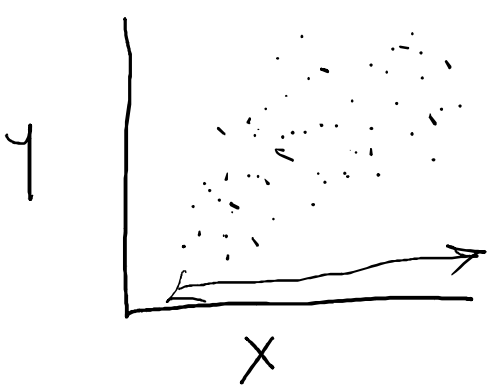


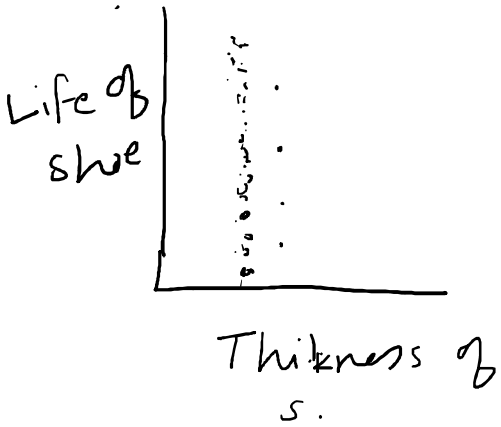
② Testing – We know the answer

Features → Predict → Type of Fruit

Compare

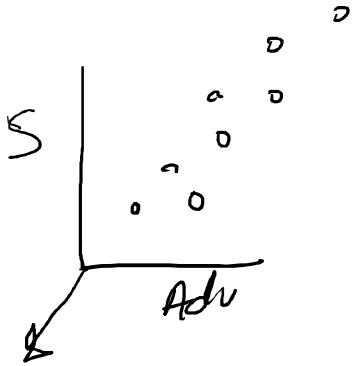
what patterns / groups exist in the data?





Machine Learning

Dimension Reduction

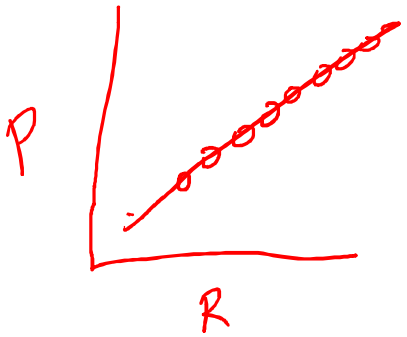


Dimensions

- Each attribute is called feature and forms dimension in mathematical space / feature space
- Single attribute data does not convey much but when combined with another dimension, reveals more actionable information
- Total variance in a dataset (seen together all dimensions) is often equated to information conveyed by the data set as a whole.
- More the dimensions, generally, more information the dataset reveals
- However, beyond a point, too many number of dimensions **can** pose challenge to machine learning algorithms

Dimension Reduction

- Just adding more dimensions, may cause increase in noise
- Dimension reduction intends to convert high dimensional data set into lesser dimension data **with minimal loss of information**.
- Low Variance Filter. Columns with little variance carry little information. Such columns can be **considered** for dropping..
- High Correlation Filter. Data columns with very similar trends are also likely to carry very similar information. In this case, only one of them will suffice to feed the machine learning model.



Principle Component Analysis

- Principle Component Analysis transforms existing dimensions to increase the signal to noise ratio.
- From original dimensions, it creates a new dimension
- Helps remove redundancy by eliminating dimensions which contain same information as another attribute

Machine Learning

Examples of Roles

What is needed to Build ML Models

- Good quality data that is representative of the real-world process is the key starting point. Without data, we cannot build machine learning models
- Domain knowledge – without domain knowledge, it is not possible to understand data, check data quality etc. which is essential while building a model
- Understanding of basic mathematics, statistics and machine learning algorithms
- Technical programming skills

Roles

- Data Scientist:
  - Understanding business challenges
  - Predictive analytics. Define models to be used.
  - Create valuable actionable insights using data.
  - Effectively communicate findings to the business.
  - Ability to understand Big picture, in-depth knowledge of Statistics techniques and also technical competency to work with data.

- Machine Learning Engineer:
  - Design and develop machine learning algorithms
  - Run machine learning tests and experiments
  - Optimize models
  - Implement appropriate ML algorithms

Other Related Roles

- Data engineer / Big Data engineer, Data Architect
- Business Analyst
- Visualization expert

Machine Learning

Python for Machine Learning

Machine Learning Languages

- Python and R are suited for data science functions.
- Go is emerging as an alternative but is not yet as well supported as Python.
- In practice, data science teams use a combination of languages to play to the strengths of each one, with Python and R used in varying degrees
- As of now, Python stands out as the preferred language for machine learning framework

Python

- NumPy for scientific computing, many other libraries use NumPy arrays to operate efficiently. It also supports multidimensional arrays and matrices, as well as mathematical and statistical functions that need little code

- SciPy builds on NumPy by adding a collection of algorithms and functions for computing integrals numerically, solving differential equations, optimization, and more.

- Pandas adds data structures and tools that are designed for practical data analysis. Pandas provides tools for shaping, merging, reshaping, and slicing datasets

- Matplotlib is the standard Python library for creating 2D plots and graphs.

- scikit-learn builds on NumPy and SciPy by adding a set of machine learning algorithms, including clustering, regression and classification.

Jupyter Notebook

- Jupyter Notebook extends the functionality of Python's interactive interpreter with a interactive shell

- It is highly preferred for conducting training programs

Machine Learning

Supervised Machine Learning

Supervised Machine Learning

- A class of machine learning algorithms that work on data in form of predictor attributes and associated target values
- Process of model building involves training and testing stages
- Training stage involves use of training data (a subset of the externally supplied data) supplied in form of predictor and target values
- They produce a model which is supposed to represent the real process that generated the data
- The model is tested for it's performance in test stage using test data. If satisfactory the model is implemented
- The model is used to predict target values for new data points

Supervised Machine Learning