Sane's STATS Academy of Statistics

Types of ML

Types of Machine Learning Techniques

- The machine learning algorithms which we will be covering are
  - Supervised learning algorithms
  - Unsupervised learning algorithms

Models for Supervised Learning

- We identify strong links between variables of a data table (columns).
- Such a link may translate into an expression between one variable y (the so-called "dependent" or "response" or “label” variable) and a group of other variables {xi} (the so-called “independent variables" or "predictors“ or “features”) :

$y = f(x_1, x_2, ..., x_p)$ + Small random noise

Models for Supervised Learning

- When the response variable is numerical, predictive modeling is called Regression.
- When the response variable is nominal / categorical, predictive modeling is called Classification. The values of the response variable can be considered as “class labels” in this case.

Examples

- **Regression Case**: Sales are influenced by the variables like advertisement expenses, manpower deployed for sales, cost of products, number of dealers etc. Hence we see here

  Sales = function (Adv. Exp , Manpower , Cost , Dealers , …)

- **Classification Case**: The customer may purchase a particular product based on some conditions like his need, his age, his income, his place of residence etc. Hence we see here

  Prob(Customer Purchases) = function(Age, Income, Residence,…)

Short Quiz: Identify the type

1. An e-commerce company using labeled customer data to predict whether or not a customer will purchase a particular item.
2. A healthcare company using data about cancer tumors (such as their geometric measurements) to predict whether a new tumor is benign or malignant.
3. A factory wanting to predict the time before a break-down of its production machines.
4. A restaurant using review data to ascribe positive or negative sentiment to a given review.
5. A bike share company using time and weather data to predict the number of bikes being rented at any given hour.

Short Quiz: Answers

1. An e-commerce company using labeled customer data to predict whether or not a customer will purchase a particular item. **--- Classification**
2. A healthcare company using data about cancer tumors (such as their geometric measurements) to predict whether a new tumor is benign or malignant. **--- Classification**
3. A factory wanting to predict the time before a break-down of its production machines. **--- Regression**
4. A restaurant using review data to ascribe positive or negative sentiment to a given review. **--- Classification**
5. A bike share company using time and weather data to predict the number of bikes being rented at any given hour. **--- Regression**

Examples: Supervised Learning

- Naïve Bayes
- K-NN
- Decision Trees
- Regression Models
- Neural Nets
- Support Vector Machines

Partitioning in Supervised Learning

- In Supervised Learning, we partition the data
- We typically deal with two or three partitions:
  - a training set,
  - a Test set,
  - and sometimes an additional test set.

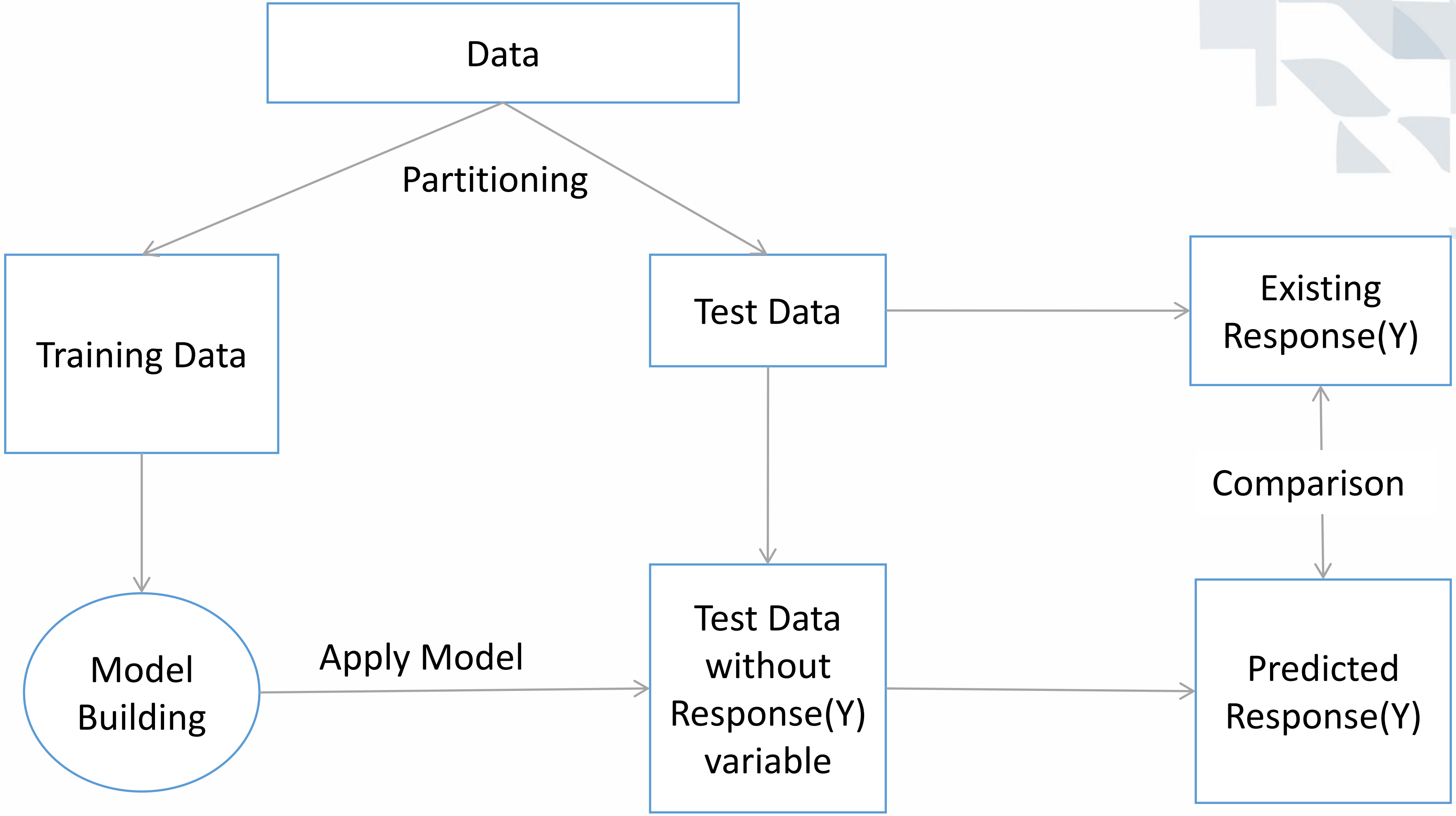
Training Partition

- Typically the largest partition
- Contains the data used to build the various models we are examining
- Generally used to develop multiple models.

Test Partition

- Used to assess the performance of each model so that you can compare models and pick the best one.
- This partition is used for internally verifying the performance of the models
- Important for measuring the goodness of fit

Supervised Learning Process with 2 partitions



Data

Partitioning

Training Data

Test Data

Existing Response(Y)

Comparison

Model Building

Apply Model

Test Data without Response(Y) variable

Predicted Response(Y)

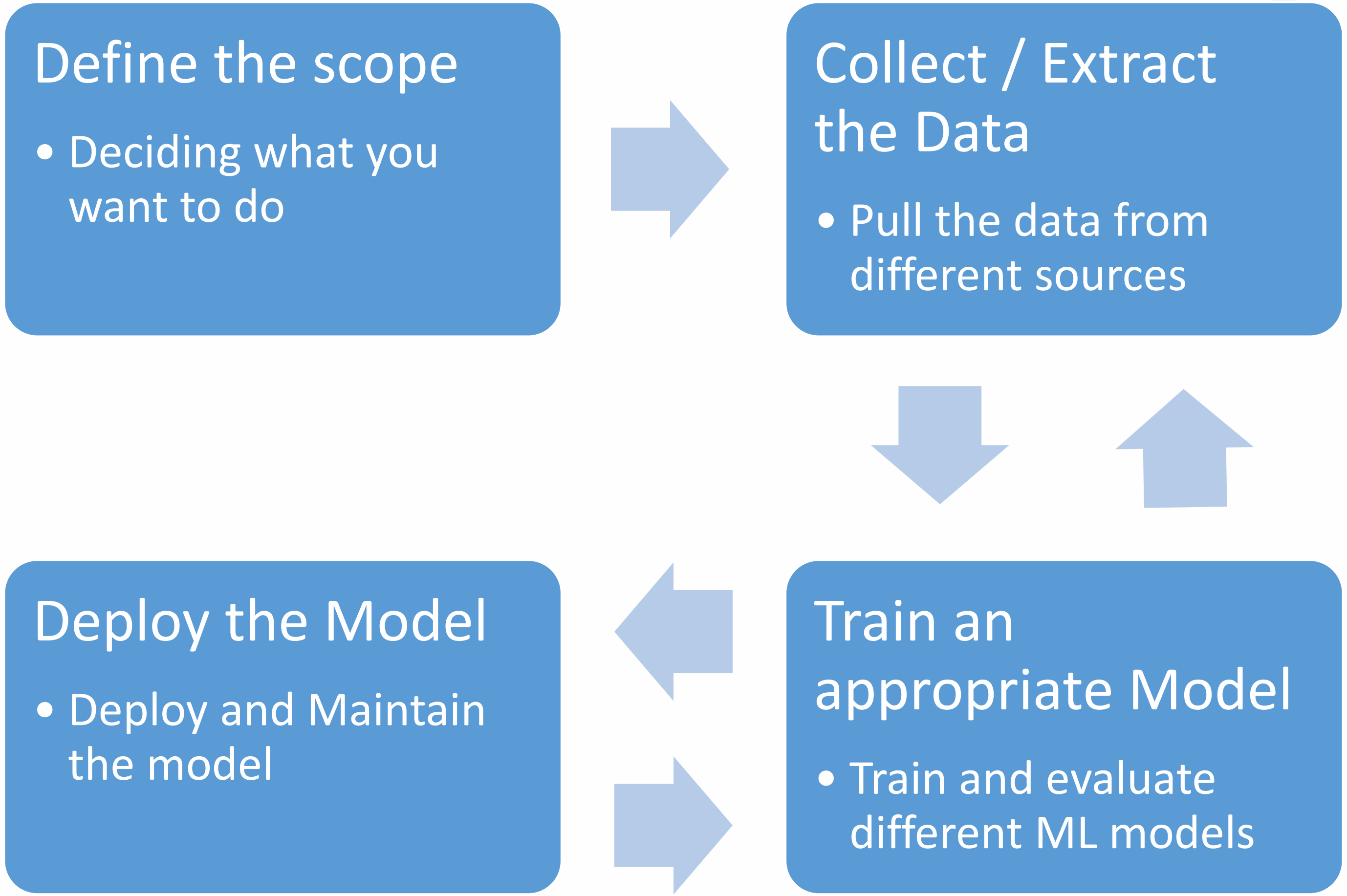
Unsupervised Learning

- Unsupervised learning algorithms are those used where there is no outcome variable to predict or classify.
- Association rules, data reduction methods, and clustering techniques are all unsupervised learning methods.

Examples of Unsupervised Learning

- Clustering Techniques
  - Hierarchical
  - K-means
  - DBSCAN
- Principal Component Analysis
- Association Rules

Lifecycle of any ML Project



Define the scope

- Deciding what you want to do

Collect / Extract the Data

- Pull the data from different sources

Train an appropriate Model

- Train and evaluate different ML models

Deploy the Model

- Deploy and Maintain the model

Sane's STATS Academy of Statistics

Technologies for ML

Desktop Software

- Click and Drag (Menu Driven)
  - KNIME
  - RapidMiner
  - SAS Enterprise Miner
  - IBM SPSS Modeller

Programming Languages

- R
- Python
- Julia
- Scala

- An open source project
- Fast on desktop with small sized data
- Add-ins (packages) available for every statistical/ML algorithm in the world
- Has been used since last 2 decades for statistical computing by statistical professionals community
- There are good IDEs available like RStudio, RTVS, R Commander, Tinn-R, STATET(Eclipse plug-in) etc.
- Among IDEs R Studio is most known
- Provides a scope for implementing our own algorithms being an open source language

python™

- An open source project
- Fast on desktop with small sized data
- Add-ins (packages) available for every statistical/ML algorithm in the world
- The statistical aspects of Python have been developed recently
- There are good IDEs available like Spyder(Anaconda Installation), PyCharm etc.
- Provides a scope for implementing our own algorithms being an open source language

Cloud-Based Platform

- Amazon Web Services
- Microsoft Azure
- Google Cloud AI

Large Scale Data Processing Libraries

- Libraries are such kind of modules which are language independent.
- Using libraries, one can code in R / Python / Java
- Well known libraries for ML are
  - Apache Spark
  - h2o (by h2o.ai)
  - TensorFlow (by Google)
  - Theano (by University of Montreal)
  - CNTK (by Microsoft)
- All of the above provide support for GPU-based operations for algorithms in Deep Learning
- The superb feature which these libraries provide is the fast speed that too at relatively low cost.

Thank You