# Preface

## Welcome {-}

This is work has had more lifetimes than a bag full of cats. It started as a hands-on Linux command line guide for new biology students.

However, as I was reading on the topic of Linux, I read a quote that describes the philosophy of Unix.

>The Unix philosophy is one of the simple tools, each doing one job well...

>Anonymous

\*\*This is not the philosophy of R nor RStudio.\*\* I am currently looking for an R package that fits the Unix motto.

As I stepped back-and-forth \*and\* back-and-forth, I realized I was more interested in teaching myself machine learning using Caret, R and RStudio.

### Who This Book Is For {-}

This book is for me and Worcester Polytechnic Institute for the fulfillment of my Master's degree.

It is very informal. It is my record of what I believed would help me shortly.

### Acknowledgments {-}

I would like to thank Doctors Dmitry Korkine and Elizabeth Ryder for their patience and generosity.

### Dedication {-}

To Richard and Susan Curcio

### About the Author {-}

Matthew C. Curcio is a chemist, biochemist, teacher, and student who loves science. I loved teaching but I loved science more.

The first time I realized Data Science was a bonafide subsection of science was the day I saw a fellow scientist make a kludgey mishmash with Visual Basic and an Excel spreadsheet.

>It is wisdom which is seeking for wisdom.

>

>Shunryu Suzuki, "Zen Mind, Beginner's Mind", Shambhala Publications, 1970, ISBN 978-1-59030-849-3.

<div style="page-break-after: always;"></div>

# Abstract

Nine machine learning methods (both supervised and unsupervised) have been introduced in 'Top 10 Algorithms In Data Mining' by X. Wu et al, 2008.[^1] Unfortunately, there are no easy rules to follow for which models to use in all situations. The comically named \*No Free Lunch Theorems for Optimization\* "demonstrates the danger of comparing [general-purpose] algorithms by their performance on a small sample of problems."[^2] Therefore the author felt it is necessary to carry out an empirical study of 9 of the 10 machine learning algorithms which were deemed influential in the field of data mining by the International Conference on Data Mining, which was polling source for the Wu paper. These algorithms include C5.0, k-Means, SVM, Apriori, EM, AdaBoost, kNN, Naive Bayes, and CART. PageRank was not considered due to its in-applicability. These methods can be carried out using one all-encompassing program known as \*caret\* written by Max Kuhn et al.[^3] \*caret\* is an R language package that allows one to use the same syntax for machine learning libraries written by over 200 different researchers also in the R language. These nine machine learning methods will be used on the classification of Oxygen binding proteins, eg. Hemoglobin, into six groupings as well as a negative control supplementing work carried out by S. Muthukrishnan et al, which only investigates the utility of SVM to differentiate the 6 groups.[^4] The machine learning tests were will be compared primarily using two measures, CPU-run time, and Cohen's kappa.

[^1]: \*Top 10 algorithms in data mining\*, X. Wu, Knowl Inf Syst (2008) 14:1–37,

DOI 10.1007/s10115-007-0114-2

[^2]: \*No Free Lunch Theorems for Optimization\*, IEEE Transactions On Evolutionary Computation, Vol. 1, NO. 1, April 1997, D. Wolpert et al

[^3]: \*The caret Package\*, M. Kuhn, 2019-03-27, https://topepo.github.io/caret/index.html

[^4]: \*Oxypred: Prediction and Classification of Oxygen-Binding Proteins\*, S. Muthukrishnan et al, Geno. Prot. Bioinfo., Vol. 5 No. 3–4, 2007

<div style="page-break-after: always;"></div>

# Intromission

> Intromission - Noun. the state of being allowed to enter; admittance

## Prerequisites

To get the most from this booklet, you should have:

1. Access to an Internet connection.

2. A computer with \>1 Gigabyte of hard drive space available.

3. Download R (version: >= 3.4.4 - 2018-03-15) at [cran.r-project.org](https://cran.r-project.org)

4. Download RStudio (version: >= 1.1.463 - 2018) at [rstudio.com](https://www.rstudio.com/products/rstudio/download/)

- Choose the free desktop version - the server software and the commercial license are unnecessary.

5. Have the ability/permission to install the software (i.e. R & RStudio).

---

## How to Load R and RStudio???

# Introduction {#Section-Introduction}

```{r, include=FALSE}

library(knitr)

knitr::opts\_chunk$set(cache = TRUE,

warning = FALSE,

message = FALSE,

cache.lazy = FALSE)

```

In early 2019, I was fortunate and came across the book "Applied Predictive Modeling" by Max Kuhn and Kjell Johnson.[^n1] The focus of this book is the implementation and optimization of predictive modeling. However "Applied Predictive Modeling" is also a demonstration of the power of the R package `caret` written by Max Kuhn and others.

[^n1]:"Applied Predictive Modeling", Max Kuhn and Kjell Johnson, Springer Publishing, 2018, (http://appliedpredictivemodeling.com/)

>"The caret package, short for classification and regression training, focuses on simplifying model training and tuning across a wide variety of modeling techniques.[^n2]

[^n2]:Building Predictive Models in R Using the caret Package, Max Kuhn, J. Stat. Soft., Nov. 2008, 28, 5, www.jstatsoft.org/v28/i05/paper

The R package `caret` was written, in part, to overcome the syntactical differences between numerous R machine learning packages as it helps automate the process of model scouting and development.

I plan to discuss "Applied Predictive Modeling" and its tight connection to `caret` in this work since it outlines a methodical process for machine learning.

As one might suspect, many R machine learning packages having been developed by many different scientists have different variable input and output formats. In the spirit of harmonization and simplicity, `caret` standardizes the process of model development. As of August 2019, `caret` can pass information to/from 238 R packages.[^n3] What `caret` does to facilitate machine learning process development will be discussed in more detail.

[^n3]:The caret Package (https://topepo.github.io/caret/index.html)

Additionally, while searching for information regarding superior Machine Learning techniques, I became aware of an article co-authored by Ross Quinlan and others, with the grand title, "Top 10 Algorithms in Data Mining"[^n4]. These 10 algorithms were identified by the IEEE, International Conference on Data Mining in December 2006. The algorithms were voted by the members of this organization as having the highest impact. As one would expect, these 10 algorithms are the mainstays of Predictive Modeling. After further research, I have also learned that this article and its companion book[^n5], have become the subject of an MIT course.[^n6]

[^n4]:"Top 10 Algorithms in Data Mining", X. Wu et al, Knowl Inf Syst, 2008, 14:1–37, DOI 10.1007/s10115-007-0114-2

[^n5]:"The Top Ten Algorithms in Data Mining", edited by X. Wu, V. Kumar, CRC Press, 2009, [Book available on Google Books](https://books.google.com/books?id=\_kcEn-c9kYAC&printsec=frontcover&source=gbs\_ViewAPI#v=onepage&q&f=false)

[^n6]:Cynthia Rudin, 15.097 Prediction: Machine Learning and Statistics. Spring 2012. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu/, Creative Commons BY-NC-SA.

---

At this point, \*\*Discuss ANTI-CANCER PEPTIDES\*\*

The 672 proteins were taken from six different classes of oxygen-binding proteins consisting of 6 groups.

\*\*7 Class Accuracies Via Support Vector Machine Models\*\*

| Protein Name | n | Accuracy (%) |

|:----------------------|:----|-------------:|

| C/Control/negtive80 | 266 | n |

| P/Positive/positive80 | 266 | n |

I realized that this paper was a great choice to model for several reasons. I postulated the high percent accuracies would very likely drop given a larger dataset of Oxygen binding proteins. Since the paper is from 2007, the number of proteins available to any researcher on Uniprot or the Internet, in general, would have been small, 672 as mentioned above. As of July 2019, the number of the possible available number of proteins in the same class as Globins was 89,064.

For optimum scouting, Kuhn and Johnson suggest that a much broader framework of half a dozen or more machine learning techniques were commonly used in their opinion for screening in Bioinformaticists before finally choosing a good model candidate. The "Oxypred" paper did not discuss any model development, therefore, I speculated that a more thorough search using the "The Top Ten Algorithms in Data Mining" may provide an excellent chance to study the differences between the machine learning approaches.

By using these three papers, I would be able to easily generate data that should yield intuitive insights regarding the characteristics of the proteins of interest. But more importantly, the work would allow any reader to learn how a class of very diverse and impactful machine learning algorithms would stack up in the face of a small dataset. The preparation of the dataset used in this research is discussed in Appendix A.

- Data mining is all about extracting patterns from an organization's stored data. These patterns can be used to gain insight into aspects of the organization's operations, and to predict outcomes for future situations as an aid to decision-making.

### Learning Algorithm Selection

1. Explainability

1. In-memory vs. out-of-memory

1. Number of features and examples

1. Categorical vs. numerical features

1. The nonlinearity of the data

1. Training speed

1. Prediction speed

### 4 Big Challenges in Machine Learning (ft. Martin Jaggi)

https://www.youtube.com/watch?v=v3QGgtmAZTE&t=0s&index=10&list=WL

Supervised Vs Unsupervised

Problem 1: The vast majority of information in the world is unlabeled so it would be advantageous to have good Unsupervised machine learning algorithms to use.

Problem 2: Algorithms are very specialized, too specific.

Problem 3: Transfer learning to new environments

Problem 4: Scale, the scale of information is huge in reality and we have computers that work in gigabytes, not the Exabytes that humans may have available to them. The scale of distributed Big Data...

### 100\_page\_ml\_book\_NOTES

Machine learning can also be defined as the process of solving a

practical problem by

1. gathering a dataset, and

2. algorithmically building a statistical model based on that dataset.

That statistical model is assumed to be used somehow to solve

the practical problem.

### Supervised Learning

- dataset is the collection of labeled examples, sometimes called a predictor

- xi among N is called a feature vector

- label yi can be either an element belonging

- to a finite set of classes {1, 2, . . . , C}, eg. {type1, type2}

- f(x1, x2, x3,...xn)= yi, {(x i , y i )}

### Unsupervised Learning

- In unsupervised learning, the dataset is a collection of

- unlabeled examples {x i } N i=1 .

- For example, in clustering, the model returns the id of the

- cluster for each feature vector

- How Supervised works

Any classification learning algorithm that builds a model implicitly or explicitly creates a decision boundary. A decision boundary can be straight, or curved, or it can have a complex form.

In practice, there are two other essential differentiators of learning algorithms to consider:

1. speed of model building and

2. prediction processing time.

Why the Model Works on New Data

- If training examples are selected randomly & independently using the same procedure, then, \*\*statistically\*\*, it is more likely that the new negative example will be located closer to negative examples.

- For less likely situations, our model will make errors, because these situations are run across less often.

\*HOW TO REPHRASE:\* The number of errors will more likely be smaller than the number of correct predictions.

So it tends to reason that the larger the dataset the larger are the chances of describing all or most of the probable outcomes that one would encounter.

### "Probably approximately correct" learning theory by Leslie Valiant

- We cannot expect a learner to learn a concept exactly. There are only a small fraction of examples that describe our available instance/experimental space. It may be very difficult to describe the entire space with a small set.

- Cannot always expect to learn a close approximation to the target concept, the training set will not be representative of your model. It will contain uncommon examples.

- The only realistic expectation of a good learner is that with high probability it will learn a close approximation to the target concept/model.

- The only reason we can hope for this is the \*consistent distribution assumption\*.

### Parameters vs. Hyperparameters

- A hyperparameter is a property of a learning algorithm, usually (but not always) having a numerical value. That value influences the way the algorithm works. Hyperparameters aren't learned by the algorithm itself from data. They have to be set by the data analyst before running the algorithm.

### Classification vs. Regression

- \*\*Classification\*\* is the problem of trying to assign a label to a previously unlabeled example.

- the classification problem is solved by usually using a collection of examples/objects \*already\* labeled and producing a model that can NOW take unlabeled samples and label them with the information at hand.

- Is the labeled set binary? {0, 1}

or

- Multi-class (3 or more labels or classifications)

- \*\*Regression\*\* is the problem of predicting a real-value label (often called a target) given an unlabeled example.

- The regression problem is solved by a regression learning algorithm that takes a collection of labeled examples as inputs and produces a model that can \*then\* take an unlabeled example as a new input and output target value.

### Model-Based vs. Instance-Based Learning

\*\*DEFINE: Instance-Based Learning\*\*

1. Model-based learning algorithms use the training data to create a model that has parameters learned from the training data. In SVM, the two parameters we saw were w`\*` and b`\*`.

2. Instance-based learning algorithms use the whole dataset as the model. k-Nearest Neighbors (kNN) is very common.

### Shallow vs. Deep Learning

GET GOOD DEFINITION!!!

### Parametric Approach to Dimension Reduction, Feature Reduction/Selection

Forward Selection Method Approach

- Add a new Dimension one at a time then look at the goodness of fit and error statistics. if the feature is good to keep it else remove it from the model.

- add new dim if an error is low and the goodness of fit goes up, could use BIC of AIC.

- good for RF: logistic regression, neural network, SVM.

Backward Selection Method

- one variable at a time is deleted and the model goodness of fit is compared.

if the deletion of a feature reduces error and the goodness of fit increases that then the variable is dropped else it is kept.

- More can be found Book: The comparisons of data mining techniques for the predictive accuracy of probability of default credit card clients, Expert Systems with Application,

Yeh & Lien 2009, 36(2), 2473-2480

Dimensionality Reduction Methods Include:

1. PCA

2. SVD

3. Iterative feature Selection method(s?)

- Dim Reduction Techniques

1. Parametric or model based

- PCA

- Forward feature selection

- Backward feature selection

- Parametric methods, like iterative feature selection

Too many variables may be detrimental. For example, if there are thirty D that affect sales but the company cannot control all of them (such as weather) then it may not be worth trying to model them.

By reducing Dimensions it can also save space and cost.

Highly correlated variables mean redundant data.

### Dealing with Missing Values

Missing data should not be confused with censored data where the exact

value is missing but something is known about its value. For example, a

company that rents movie disks by mail may use the duration that a customer

has kept a movie in their models. If a customer has not yet returned a movie,

we do not know the actual period, only that it is as least as long as the

current duration. Censored data can also be common when using laboratory

measurements. Some assays cannot measure below their limit of detection.

In such cases, we know that the value is smaller than the limit but was not

precisely measured.

- For predictive models, it is more common to treat these data as simple missing

data or use the censored value as the observed value. missing data can be imputed.

- One popular technique for imputation is a K-nearest neighbor model.

If the number of predictors affected by missing values is small, an exploratory analysis of the relationships between the predictors is a good idea.

For example, visualizations or methods like PCA can be used to determine if

there are strong relationships between the predictors. If a variable with missing values is highly correlated with another predictor that has few missing

values, a focused model can often be effective for imputation.

- knn - A new sample is imputed by finding the samples in the training set "closest"

to it and averages these nearby points to fill in the value.

### Removing Predictors

There are potential advantages to removing predictors before modeling.

First, fewer predictors relate to decreased computational time and complexity.

\*\*zero variance predictor\*\* - A tree-based model (Sects. 8.1 and 14.1) is impervious, a model such as linear regression would find these data problematic.

- The fraction of unique values over the sample size is low (say 10 %).

- The ratio of the frequency of the most prevalent value to the frequency of the second most prevalent value is large (say around 20).

- Between-Predictor Correlations - multicollinearity,

- Fig. 3.10: A visualization of the cell segmentation correlation matrix. The order of the variables is based on a clustering algorithm.

- A "scree plot" where the percentage of the total variance explained by each component is shown

- Since collinear predictors can impact the variance of parameter estimates in this model, a statistic called the variance inflation factor (VIF) can be used to identify predictors that are impacted (Myers 1994).

- A less theoretical, more heuristic approach to dealing with this issue is to remove the minimum number of predictors to ensure that all pairwise correlations are below a certain threshold. While this method only identifies collinearities in two dimensions, it can have a significantly positive effect on the performance of some models.

The algorithm is as follows:

1. Calculate the correlation matrix of the predictors.

2. Determine the two predictors associated with the largest absolute pairwise correlation (call them predictors A and B).

3. Determine the average correlation between A and the other variables. Do the same for predictor B.

4. If A has a larger average correlation, remove it; otherwise, remove predictor B.

- Suppose we wanted to use a model that is particularly sensitive to between-predictor correlations, we might apply a threshold of 0.75.

5. Repeat Steps 2–4 until no absolute correlations are above the threshold.

# Topics Useful to ME:

## Data Pre-processing, p27

- Data Transformations for Individual Predictors

1. Centering, the predictor has a zero mean.

2. Scaling, each value of the predictor

variable is divided by its standard deviation.

3. Transformations to Resolve Skewness

1. Replacing the data with the log, square root, or inverse may help to remove the skew.

2. Box and Cox Transformation can be useful too. p32

1. http://onlinestatbook.com/2/transformations/box-cox.html

Skew formula

Data Transformations for Multiple Predictors, p33

1. Transformations to Resolve Outliers

2. Investigate why they are different...

3. Several predictive models are resistant to outliers. Tree-based classification models create splits of the training data and the prediction equation

4. If a model is considered to be sensitive to outliers, one data transformation that can minimize the problem is the spatial sign (Serneels et al. 2006).

5. Data Reduction and Feature Extraction - PCA is a commonly used data reduction technique

- Dealing with Missing Values

- Removing Predictors

- Adding Predictors

- Binning Predictors

- Computing

## Chapter 4 Over-Fitting and Model Tuning

- Model Tuning

- Data Splitting

- Resampling Techniques

- Choosing Between Models

### Chapter 20 Factors That Can Affect Model Performance

- Type III Errors

- Measurement Error in the Outcome

- Measurement Error in the Predictors

- Discretizing Continuous Outcomes

- When Should You Trust Your Model's Prediction?

- The Impact of a Large Sample

#### Three Pieces of Advice (1) be patient, (2) be creative, and (3) be practical.

Supervised learning problems vs Unsupervised

Following are two different kinds of supervised learning problems which are later associated with different solution approaches:

Numerical related problems in which one predicts the quantity (represented using numbers). Algorithms such as regression, SVM, neural network, decision trees, etc are used to solve these kinds of problems.

Classification related problems in which one predicts classes such as yes/no, positive/negative, good/bad/ugly, etc. Algorithms such as logistic regression, SVM, neural network, K-NN, decision trees, etc are used to solve these kinds of problems

No matter how much data we have, it is very unlikely that we will see those exact examples again at test time.

Doing well on the training set is easy (just memorize the examples).

- cross-validation; this is very important

VERY INTERESTING

- Notice that generalization being the goal has an interesting consequence for machine learning. Unlike in most other optimization problems, we don't have access to the function we want to optimize! We have to use training error as a surrogate for test error.

- Every learner must embody some knowledge or assumptions beyond the data it's given to generalize beyond it. This was formalized by Wolpert in his famous "no free lunch" theorems, \*\*according to which no learner can beat random guessing over all possible functions to be learned\*\*.

- But this is alright considering we don't want to predict ALL possible values with our function just a relatively small subset.

- Very general assumptions—like smoothness, similar examples having similar classes, limited dependences, or limited complexity are often enough to do very well

OVERFITTING HAS MANY FACES

- Bias and Variance

- One way to understand overfitting is by decomposing generalization error into bias and variance. Bias is a learner's tendency to consistently learn the same wrong thing. Variance is the tendency to learn random things irrespective of the real signal.

- decision trees learned on different training sets generated by the same phenomenon are often very different, when in fact they should be the same.

---

At the intersection between Statistics and Computer Science, there is a portion outside of Bioinformatics. This intersecting area is where the field of Machine Learning or Predictive Modeling resides.

Machine Learning, Predictive Modeling, Artificial Intelligence, and even intelligent machines are all synonyms for pretty much the same tool(s). I plan to use Machine Learning or Predictive Modeling interchangeably. The technology boils down to the use of mathematics, specifically statistics coupled with computer science.

This booklet will discuss ?supertwo of the most commonly studied facets of machine learning, Supervised and Unsupervised learning approaches.

- Definitions:

1. \*\*Supervised learning\*\* is used whenever we want to predict a certain

outcome from a given input, and we have examples of input/output pairs.

2. In \*\*unsupervised learning\*\*, the learning algorithm is just shown the input data

and asked to extract knowledge from this data.[^1]

[^1]:Introduction to Machine Learning with Python (2017), Andreas C. Müller & Sarah Guido, O'Reilly press, 2017

TheA second set of definitions might be helpful although from n different view.

1. Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.[^2] The algorithm determines a pattern from the input information and groups this with its title or classification. It may also produce regression lines or models that can be used for predictions.

[^2]:Stuart J. Russell, Peter Norvig (2010) Artificial Intelligence: A Modern Approach, Third Edition, Prentic- Hall

- Supervised learning

\*\*Input:\*\*

1. (X, Y) is a \*\*labeled pair\*\*

Where:

- $\langle X \rangle ~~ is~ a~ matrix(m,n)~ \in \mathbb{R}^n$,

- $\langle Y \rangle ~~ is~ a~ matrix(m,1),~~~ label~y\_i \in \{1, 2, 3, ...,C\}~~$

\*\*Output:\*\*

$f(x\_i, x\_j, ... ~| Y) = Classification(s)$

OR

$f(x\_i, x\_j, ... ~| Y) = Regression ~model(s)$

- Unsupervised learning

\*\*Input:\*\*

1. (X) is a \*\*unlabeled data\*\*,

Where:

$\langle X \rangle ~~ is~ a~ matrix(m,n)~ \in \mathbb{R}^n$

\*\*Output:\*\*

$f(x\_i, x\_j, ...) = Classification(s)$,

OR

$f(x\_i, x\_j, ...) = Regression ~model(s)$

Machine Learning from Podcast MLGuide

3 Steps of ML

1. infe predict

2. erro loss; sometimse aka cost functionis taken from business

3. train / learn

define:

-The aAlgorithm as software which spells out the steps for the mathematics and procedure.

- rows equal observations, n's

- columns equal Attributes or Features or Variables

-The Last column is typically the Y - predictor.

- the model is the stored pattern learned the algorithm the value of numerical values is the model.

- sometimes the model is ALL the data. For example, KNN does not learn deeply. KNN learns shallowly because it does not generalize but KEEPS the original data for predicting values in the future. It does not keepthe formula in the same way a regression line is the distillation of all the data points.

- uses linear algebra to solve given all observations

- Features can be numerical or nominal, categorical or even binary.

- The Y value or outcome could also be a numerical value or a categorical or even binary too.

- the model produces the coefficients for the linear regression for example, but in machine learnin,g it is more often called a weight.

Procedure:

1. load data

2. clean data, delete id and number of aa, set class asa factor

3. further split the data into 80:20 parts,

- 80%of training data

- 20%of testing data

4. Use training data to learn or train the model.

- Set up training grid to tuning parameters

- Set up cross-validation

5. Obtain times for training

6. Using Testing data set Predict classes and generate CM.

7. Compile CM output (Accuracy, MCC, etc.) for comparing 10 models.

### monothetic or polythetic

- Turn categorical into numerical values

- Consider all features and pick machine learning algo to doa job.

### Modeling Algorithms

- Regression:

- Simple linear regression

- Multiple linear regression

- GLM witha proper understanding of assumptions, multicollinearit,y etc

-The goodnesss of fit measures such as R^2, RMSE, Adjusted-R^2, Predicted-R^,2 etc

- Lasso regression

- Ridge regression

- Concept of regularization

- Concept of over-fitting and under-fitting, learning rate

- Concept of bias and variance

- Classification

- K-NN

- Logistics regression with ROC/AUC curve and selection of optimal threshold

- Decision tree

- Concept of Bagging: Random forest

- Support vector machine(s)

- Baye's classifier

- Concept of Boosting: Xgboost, Adaboost

- Artificial neural networks:

- Perceptron, Multi-layer perceptron(s) (Feed forward neural networks)

- Classification performance measures:

- confusion matrix, accuracy, sensitivity, specificity, precision, F-measure

- Clustering:

- K-means witha fair understanding of distance and dissimilarity measures for non-numerical type of variables as well, eg. text

- Hierarchical clustering

- Dimensionality reduction techniques, eg. PCA

---

title: "ML-Types"

author: "Matthew Curcio"

date: "6/23/2019"

output: html\_document

---

Title: ML-Types

```{r message=FALSE, warning=FALSE, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE, cache = TRUE)

library(knitr)

```

At the intersection between Statistics and Computer Scienc,e there is a portion outside ofBioinformaticss. This intersecting area is where the field of Machine Learning or Predictive Modeling resides.

Machine Learning, Predictive Modeling, ArtificialIntelligence,e and even intelligent machines are all synonyms for prettymuchc the same tool(s). I plan to use Machine Learning orPredictivee Modelinginterchangeablyy. The technology boils down to the use of mathematics, specifically statistics coupled with computer science.

Thisbooklett will discuss supertwo of the most commonly studied facets of machine learning, Supervised and Unsupervised learning approaches.

```{r, echo=FALSE, fig.cap="Two Types Of Machine Learning", out.width = '70%', fig.align="center"}

knitr::include\_graphics("types\_of\_ML\_small\_size.png")

```

##### Definitions:

1. \*\*Supervised learning\*\* is used whenever we want to predict a certain

outcome from a given input, and we have examples of input/output pairs.

2. In \*\*unsupervised learning\*\*, the learning algorithm is just shown the input data

and asked to extract knowledge from this data.[^1]

[^1]:Introduction to Machine Learning with Python (2017), Andreas C. Müller & Sarah Guido, O'Reilly press, 2017The

A second set of definitions might be helpful although froaan different view.

1. Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.[^2] The algorithm determines a pattern from the input information and groups this with its title or classification. It may also produce regression lines or models that can be used for predictions.

[^2]:Stuart J. Russell, Peter Norvig (2010) Artificial Intelligence: A Modern Approach, Third EditionPrentice-Hallll

---

### Supervised learning

\*\*Input:\*\*

1. (X, Y) is a \*\*labeled pair\*\*

Where:

- $\langle X \rangle ~~ is~ a~ matrix(m,n)~ \in \mathbb{R}^n$,

- $\langle Y \rangle ~~ is~ a~ matrix(m,1),~~~ label~y\_i \in \{1, 2, 3, ...,C\}~~$

\*\*Output:\*\*

$f(x\_i, x\_j, ... ~| Y) = Classification(s)$

OR

$f(x\_i, x\_j, ... ~| Y) = Regression ~model(s)$

### Unsupervised learning

\*\*Input:\*\*

1. (X) is a \*\*unlabeled data\*\*,

Where:

$\langle X \rangle ~~ is~ a~ matrix(m,n)~ \in \mathbb{R}^n$

\*\*Output:\*\*

$f(x\_i, x\_j, ...) = Classification(s)$,

OR

$f(x\_i, x\_j, ...) = Regression ~model(s)$

---

EOF

A Few Useful Things to Know about Machine Learning

Pedro Domingos

Department of Computer Science and Engineering

University of Washington

Seattle, WA 98195-2350, U.S.A.

pedrod@cs.washington.edu

Definition of Classifier:

- A classifier is a system that inputs (typically) a vector of discrete and/or continuous feature values and outputs a single discrete value, the class.

- A learner inputs a training set of examples ($x\_i , y\_i$), where $x\_i = (x\_i,1 , . . . , x\_i,d )$ is an observed input and $y\_i$ is the corresponding output, and outputs a classifier.

- The test of the learner is whether this classifier produces the correct output $y\_t$ for future examples $x\_t$ (e.g., whether the spam filter correctly classifies previously unseen emails as spam or not spam).

LEARNING = REPRESENTATION + EVALUATION + OPTIMIZATION

- The first problem facing you is the bewildering variety of learning algorithms available.

- There aly thousands available, and hundreds more are published each year. The key to not getting lost in this huge space is to realize that it consists of combinations of just three components. The components are:

REPRESENTATION:

- A classifier must be represented in some formal language that the computer can handle. Conversely, choosing a representation for a learner is tantamount to choosing the set of classifiers that it cly learn.

- This set is called the hypothesis space of the learner.

- If a classifier is not in the hypothesis space, it cannot be learned.

- It begs the question, what features to use?

EVALUATION:

- An evaluation function (also called objective function or scoring function) is needed to distinguish good classifiers from bad ones.

- The evaluation function used internally by the algorithm may differ from the external one that we want the classifier to optimize.

OPTIMIZATION:

- The choice of optimization technique is key to the efficiency of the learner, and also helps determine the classifier produced if the evaluation function has more than one optimum. It is common for new learners to start out using off-the-shelf optimizers, which are later replaced by custom-designed ones.

For example,

- k-nearest neighbor classifies a test example by finding the k most similar training examples and predicting the majority class among them.

- Hyperplane-based methods form a linear combination of the features per class and predict the class with the highest-valued combination.

- Decision trees test one feature at each internal node, with one branch for each feature value, and have class predictions at the leaves.

---

SEE: TABLE: Three-components of Learning Algos.png

![Three-components of Learning Algos](Three-Components-of-Learning-Algos.png)

---

Algorithm Pseudocode: Learn D.T. (TrainSet)

1. if all examples in Training Set have the same class $y\_∗$ then

- return Make Leaf($y\_∗$)

2. if No feature $x\_j$ has InfoGain$(x\_j ,y) > 0$ then

- $y\_∗$ ← Most frequent class in Training Set

- return Make Leaf$(y\_∗)$

3. $x\_∗$ ← argmax $x\_j$ InfoGain$(x\_j ,y)$

4. $TS\_0$ ← Examples in Training Set with $x\_∗ = 0$

5. $TS\_1$ ← Examples in Training Set with $x\_∗ = 1$

- return Make Node$(x\_∗)$, Learn D.T. $(TS\_0)$, Learn D.T.$(TS\_1)$)

---

Most textbooks are organized by representation, and it's easy to overlook the fact that the other components are equally important. There is no simple recipe for choosing each component.

- The fundamental goal of machine learning is to generalize beyond the examples in the training set.

- No matter how much data we have, it is very unlikely that we will see those exact examples again at test time.

- Doing well on the training set is easy (just memorize the examples).

- The most common mistake among machine learning beginners is to test on the training data and have the illusion of success. If the chosen classifier is then tested on new data, it is often no better than random guessing.

- HOLD back some data for more testing

- if you hire someone to build a classifier, be sure to keep some of the data to yourself and test the classifier they give you on it.

- cross-validation; this is very important

- Picture of equally divided partthatch are tested and scored then the scores are analyzed by boxplot for example of mean +/- 95% C.I.

![CV diagram](CV diagram.png)

```{r}

accuracy = c(86,93,95,83,73,84,81,78,91,89)

boxplot(accuracy,

main="Boxplot of Possible Distribution of Test Results",

xlab="Accuracy (%)",

col="orange",

border="brown",

horizontal = T)

```

VERY INTERESTING

- Notice that generalization being the goal has an interesting consequence for machine learning. Unlike in most other optimization problems, we don't have access to the function we want to optimize! We have to use training error as a surrogate for test error.

- Every learner must embody some knowledge or assumptions beyond the data it's givetoto generalize beyond it. This was formalized by Wolpert in his famous "no free lunch" theorems, \*\*according to which no learner can beat random guessing over all possible functions to be learned\*\*.

- But this is alright considering we don't want to predict ALL possible values with our function just a relatively small subset.

Veryry general assumptions—like smoothness, similar examples having similar classes, limited dependences, or limited complexity are often enough to do very well

OVERFITTING HAS MANY FACES

- Bias and Variance

- One way to understand overfitting is by decomposing generalization error into bias and variance. Bias is a learner's tendency to consistently learn the same wrong thing. Variance is the tendency to learn random things irrespective of the real signal.

- decision trees learned on different training sets generated by the same phenomenon are often very different, when in fact they should be the same.

# - Exploratory Data Analysis

## What is Exploratory Data Analysis?

EDA ia s form of statistical data analysithatch begins by exploring data rather than testily formulated prior hypotheses. Exploratory data analysis does as it says: it explores the pattern of the data set under analysis, considering its range, level, outliers, batching it before graphing and transforming it. (https://encyclopedia2.thefreedictionary.com/exploratory+data+analysis)

---

\* it allowgettinget closer to the certainty that the future results will be valid, correctly interpreted, and applicable to the desired business contexts.

(https://medium.com/@InDataLabs/why-start-a-data-science-project-with-exploratory-data-analysis-f90c0efcbe49)

---

In Roger Peng's e-book entitled [Exploratory Data Analysis](https://leanpub.com/exdata), he uses the analogy of movie production to help describe the process. In the first stages of movie developme,nt a directory might have an idea that he/she wants to explore. The scene or idea to be captured might be in the beginning, middle or end of the final project, but that seed of an idea is most valuable to visualize. From there other ideas might flow. This is often true with a larger movie. Scenes are not always shot in chronological order. One viewpoint may spring from the preceding idea or experimental view of exploratory work.

---

Q. What are the Methods of Exploratory Data Analysis

A. The same may be true for "Exploratory Data Analysis", (EDA). When one obtains a data set it is important to get an intuitive feel for the numbers at hand.

1 - First approach to data

2 - Analyzing categorical variables

3 - Analyzing numerical variables

4 - Analyzing numerical and categorical at the same time

---

- It is often the situation when they realize, that they have a lot of data and no ideas of what value that data can bring to their business decision making. (https://medium.com/@InDataLabs/why-start-a-data-science-project-with-exploratory-data-analysis-f90c0efcbe49)

---

Key points in a basic EDA:

\* \*\*Data types\*\* - categorical variables, numerical variables, nominal

\* \*\*Outliers\*\*

\* Missing values

\* Distributions (numerically and graphically) for both, numerical and categorical variables.

\* (Steps and Key points frm: https://blog.datascienceheroes.com/exploratory-data-analysis-in-r-intro/)

---

(Bioinformatics with R Cookbook by Paurush Praveen Sinha)

The key issues that must be taken care of before we start with the formulation of learning models are as follows:

- \*\*Assessing data quality\*\*: Correctness and completeness of datareis necessary to get the model to work in the way that it is supposed to. This includes checking the consistency, the attribute values and types, the missing values, and finding the outliers.

- \*\*Data normalization\*\*: Depending on the input data and the chosen algorithm, sometimes we need to standardize the dattoto make it comparable. However, it might not be required in cases where the unique scaling of data is

significant. Similarly, the normalization method can be different depending on the types of features. Another important aspect is that one should first normalize the training data and then apply these normalization parameters to rescale the validation data.

- \*\*Feature selection\*\*: We need to get the features that show a dependency between the data instance and the objective (for example, class labels in the case of classification). Depending on the data, objective, and learning algorithm, all features might not be required to create modelde; they either add noise, redundany, or cause computational inefficiency for the model. Therefore, we need to identify useful features that can define the objective function. For example, avoiding a redundant feature or correlated feature is a good call.

---

Jason BrownleePh.D.hD, says;

Simple univariate and multivariate methods that give a vieofon the data can be used.

For example, five methods that I would consider must have are:(https://machinelearningmastery.com/understand-problem-get-better-results-using-exploratory-data-analysis/)

- Five number summaries (mean/median, min, max, q1, q3)

- Histogram graphs

- Line Charts

- Box and Whisker plots

- Pairwise Scatterplots (scatterplot matrices)

- Lattice plots

---

1. Univariate visualization — provides summary statistics for each field in the raw data set

2. Bivariate visualization — is performed to find the relationship between each variable in the dataset and the target variable of interest

3. Multivariate visualization — is performed to understand interactions between different fields in the dataset

4. Dimensionality reduction — helps to understand the fields in the data that account for the most variance between observations and allow for the processing of a reduced volume of data.

---

Q. Why skipping Exploratory Data Analysis is a bad idea

1. generating inaccurate models;

2. generating accurate models on the wrong data;

3. choosing the wrong variables for the model;

4. inefficient use of the resources, including the rebuilding of the model.

---

Q. How do you get better at formulating questions to ask your AI?

A. this is data exploratin, or exploratory data analysis. You can visualize, compute statistics, like mean, min/max, variance, covariance matrix, most frequent items, run some clustering algorithms. start from some kind of requirement, i.e. a "problem" as the target for your eventual solution. one thing that will make your life easier is if you reorganize your data in ways appropriate to your hypothesis and look for preliminary patterns using some very simple techniques. (https://www.reddit.com/r/MachineLearning/comments/15mh7q/question\_regression\_with\_many\_irrelevant\_variables/)

---

Q. What do you want to know from your data?

A. The Three Rules of Data analysis.

(https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/data-analysis/)

Using three basic rules of thumb can help you avoid incorrectly making claims about your data:

1. what it is you want to know. Do you want to prove that the Earth is round? Or do you want to prove that the Earth has a circumference? Framing this question is what we call stating the \*\*hypothesis\*\*.

2. Estimate a \*\*Central Tendency\*\* for your Data. Examples of measures of central tendency are the \*\*mean and median\*\*. Which one you use will depend on your hypothesis in Step 1.

3. Consider the exceptions to the central tendency. If you've measured the average, look at the figures that are not average. is the average misleading?

---

For hypothesis testing:

- A computer may be more capable than a human of finding subtle patterns in large databases, but it still needs a human to motivate the analysis and turn the result into meaningful action.

Machine Learning with R - Second Edition

By Brett Lantz

---

Q. Do you want to learn DataFrames right now?

## Big Questions

1. What is machine learning?

2. Name applications of machine learning.

3. Explain the following machine learning areas and tasks: supervised learning, unsupervised learning, reinforcement learning, classification, regression, clustering, feature selection, feature extraction, and topic modeling.

4. Explain the following machine learning approaches: decision tree learning, artificial neural networks, and Bayesian networks.

5. What are the advantages and disadvantages of the various approaches?

6. Which kinds of products can be used for machine learning in practice?

7. How to engineer a machine learning application?

8. Explain precision, recall, and F-measure

9. Who separatehe t training data set and test data set for validation?

---

Stephenen wolfram - READ

https://blog.stephenwolfram.com/2017/05/machine-learning-for-middle-schoolers/

https://www.wolfram.com/language/elementary-introduction/2nd-ed/

---

- How-do-you-explain-Machine-Learning-and-Data-Mining-to-non-Computer-Science-people - READ

https://www.quora.com/How-do-you-explain-Machine-Learning-and-Data-Mining-to-non-Computer-Science-people/answer/Pararth-Shah

---

- Top 5 Algos for Bioinformatics / Data Science

1. Classification - CART: 952,000 results

2. Statistical Learning - SVM, 744,000 results

3. Association Analysis - Apriori, 144,000 results

4. Clustering - K-Means, 730,000 results

5. Bagging and Boosting - AdaBoost, 65,900 results

[10 Machine Learning Algorithms & R Commands](https://vitalflux.com/cheat-sheet-10-machine-learning-algorithms-r-commands/)

8 Common Tasks:

https://vitalflux.com/7-common-machine-learning-tasks-related-methods/

---

y Interesting - Basic but very good for a sixteen yr old.

https://github.com/mjhendrickson/Learning-R/wiki

---

-

https://www.datasciencecentral.com/profiles/blogs/seven-techniques-for-data-dimensionality-reduction

---

- supervised-learning-for-the-common-human

https://towardsdatascience.com/an-involved-introduction-to-supervised-learning-for-the-common-human-6338d9559748

---

- Videos for learning ML

https://www.datasciencecentral.com/profiles/blogs/neural-networks-for-machine-learning

---

- "data mining"

The phrase "data mining" is also sometimes used as a pejorative

to describe the deceptive practice of cherry-picking data to

support a theory.

Data mining should be used tcherry-pickck data to support a theory but to first approach the data with a theorthatch can be tested.

---

A computer may be more capable than a human of finding subtle patterns in large databases, but it still needs a human to motivate the analysis and turn the result into meaningful action.

---

Remember: ML is good for classification of categorization, regression, dimensionality reduction, development of algorithmsthe , discovery of genetic sequences linked to diseases

---

- Supervised learning

''Supervised learning:'' Supervised learning is a form of evidence-based learning. The evidence is the known outcome for a given input and is in turn used to train the predictive model. Models are further classified into regression and classification, based on the outcome data type. In the former, the outcome is continuous, and in the latter the outcome is discrete. Stock trading and weather forecasting are some widely used applications of regression models, anspaman detection, speech recognition, and image classification are some widely used applications of classification models.

Some algorithms for regression are linear regression, Generalized Linear Models (GLM), Support Vector Regression (SVR), neural networks, decision trees, and so on; in classification, we have logistic regression, Support Vector Machines (SVM), Linear discriminant analysis (LDA), Naive Bayes, nearest neighbors, and so on.

''Semi-supervised learning:'' Semi-supervised learning is a class of supervised learning using unsupervised techniques. This technique is very useful in scenarios where the cost of labeling an entire dataset is highly impractical against the cost of acquiring and analyzing unlabeled data.

''Unsupervised learning:'' As the name suggests, learning from data with no outcome (or supervision) is called unsupervised learning. It is a form of inferential learning based on hidden patterns and intrinsic groups in the given data. Its applications include market pattern recognition, genetic clustering, and so on.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

- Model evaluation

is a key step in any machine learning process. It is different for supervised and unsupervised models. In supervised models, predictions play a major role; where,as in unsupervised models, homogeneity within clusters and heterogeneity across clusters play a major role.

Some widely used model evaluation parameters for regression models (includincross-validationon) are as follows:

- Coefficient of determination

- Root mean squared error

- Mean absolute error

- Akaike or Bayesian information criterion

- Some widely used model evaluation parameters for classification models (includincross-validationon) are as follows:

- Confusion matrix (accuracy, precision, recall, and F1-score)

- Gain or lift charts

- Area under ROC (receiver operating characteristic) curve

- Concordant and discordant ratio

- Some of the widely used evaluation parameters of unsupervised models (clustering) are as follows:

- Contingency tables

- Sum of squared errors between clustering objects and cluster centers or centroids

- Silhouette value

- Rand index

- Matching index

- Pairwise and adjusted pairwise precision and recall (primarily used in NLP)

from: R Deep Learning Cookbook, PKS Prakash

---

- Bioinformatics Introduction

https://github.com/WilsonSayresLab/BioinformaticsIntroduction/blob/master/README.md

---

- Model assumptions

The assumptions underlying nonlinear regression models of the form specified

by Equation (1.3) are:

(1) correct mean function f

(2) variance homogeneity (homoscedasticity)

(3) normally distributed measurements errors

(4) mutually independent measurement errors ε i

---

- Most Commonly Used R Packages

Following is the list of 60 or so R packages which help take care of different aspects when working to create predictive models:

Predictive Modeling: Represents packages which help in working witvariousnt predictive models (linear/multivariate/logistic regression models, SVM, neural netwo,rk etc.)

caret: Stands for Classification And REgression TrainingIt provideses a set of functionthatch could be used to do some of the following when working with classification and regression problemsIt dependsds upothe n number of packages and loads these packages appropriately (on-demand) to achievthe e above objectives.

- Data processing (splitting)

- Feature selection

- Evaluate Model tuning parameters based on resampling

- Predictor variable importance estimation

- Estimate model performance from the training set

---

- randomForest:

Provides methods for working with classification and regression problems, based on random forests algorithm which instructs creation of large number of bootstrapped trees on random samples of variables, classifying a particular case using all of these trees in this forest, and deciding al outcome based on average or majority voting techniques depending upon whether regression or classification problem is dealt with.

- \*\*e1071\*\*: Provides methods to work with regression and classification problems. Algorithms such as following are included as part of functions:

1. Support Vector Machines (SVM)

2. naïve Bayes classifier

3. Bagged clustering

4Short-timeme Fourier transform

stats: This is the base packagthatch comes witthe h base installation of R.

---

- Visualization: Represent packages used for visualization.

ggplot2: One of the best tools for data visualization, ggplot2 could be used to create plots, layer-by-layer, using data from different data sources.

knitr: An alternative tool to Sweave, Knitr provides methods for dynamic report generation.

---

- Candidates for Most Useful Algos in D.S.

http://www.cs.uvm.edu/~icdm/algorithms/CandidateList.shtml

http://www.cs.uvm.edu/~icdm/algorithms/CandidateList.shtml

---

- Classification

- C4.5

Quinlan, J. R. 1993. C4.5: Programs for Machine Learning.

Morgan Kaufmann Publishers Inc.

Google Scholar Count in October 2006: 6907

- CART

L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and

Regression Trees. Wadsworth, Belmont, CA, 1984.

Google Scholar Count in October 2006: 6078

- 3. K Nearest Neighbors (kNN)

Hastie, T. and Tibshirani, R. 1996. Discriminant Adaptive Nearest

Neighbor Classification. IEEE Trans. Pattern

Anal. Mach. Intell. (TPAMI). 18, 6 (Jun. 1996), 607-616.

DOI= http://dx.doi.org/10.1109/34.506411

Google Scholar Count: 183

- Naive Bayes

Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All?

Internat. Statist. Rev. 69, 385-398.

Google Scholar Count in October 2006: 51

---

- Statistical Learning

- SVM

Vapnik, V. N. 1995. The Nature of Statistical Learning

Theory. Springer-Verlag New York, Inc.

Google Scholar Count in October 2006: 6441

- 6. EM

~McLachlan, G. and Peel, D. (2000). Finite Mixture Models.

J. Wiley, New York.

Google Scholar Count in October 2006: 848

---

- Association Analysis

- 7. Apriori

Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining

Association Rules. In Proc. of the 20th Int'l Conference on Very Large

Databases (VLDB '94), Santiago, Chile, September 1994.

http://citeseer.comp.nus.edu.sg/agrawal94fast.html

Google Scholar Count in October 2006: 3639

- 8. FP-Tree

Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without

candidate generation. In Proceedings of the 2000 ACM SIGMOInternationalal Conference on Management of Data (Dallas, Texas, United

States, May 15 - 18, 2000). SIGMOD '00. ACM Press, New York, NY, 1-12.

DOI= http://doi.acm.org/10.1145/342009.335372

Google Scholar Count in October 2006: 1258

---

- Link Mining

- 9. ~PageRank

Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual

Web search engine. In Proceedings of the Seventh international

Conference on World Wide Web (WWW-7) (Brisbane,

Australia). P. H. Enslow and A. Ellis, Eds. Elsevier Science

Publishers B. V., Amsterdam, The Netherlands, 107-117.

DOI= http://dx.doi.org/10.1016/S0169-7552(98)00110-X

GooglScholarar Count: 2558

- 10. HITS

Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked

environment. In Proceedings of the Ninth Annual ACM-SIAM Symposium on

Discrete Algorithms (San Francisco, California, United States, January

25 - 27, 1998). Symposium on Discrete Algorithms. Society for

Industrial and Applied Mathematics, Philadelphia, PA, 668-677.

GooglScholarar Count: 2240

---

- Clustering

- 11. K-Means

~MacQueen, J. B., Some methods for classification and analysis of

multivariate observations, in Proc. 5th Berkeley Symp. Mathematical

Statistics and Probability, 1967, pp. 281-297.

Google Scholar Count in October 2006: 1579

- 12. BIRCH

Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient

data clustering method for very large databases. In Proceedings of the

1996 ACM SIGMOInternationalal Conference on Management of Data

(Montreal, Quebec, Canada, June 04 - 06, 1996). J. Widom, Ed.

SIGMOD '96. ACM Press, New York, NY, 103-114.

DOI= http://doi.acm.org/10.1145/233269.233324

Google Scholar Count in October 2006: 853

---

- Bagging and Boosting

- 13. ~AdaBoost

Freund, Y. and Schapire, R. E. 1997. A decision-theoretic

generalization of on-line learning and an application to

boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.

DOI= http://dx.doi.org/10.1006/jcss.1997.1504

Google Scholar Count in October 2006: 1576

---

- Sequential Patterns

- 14. GSP

Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns:

Generalizations and Performance Improvements. In Proceedings of the

5th international Conference on Extending Database Technology:

Advances in Database Technology (March 25 - 29, 1996). P. M. Apers,

M. Bouzeghoub, and G. Gardarin, Eds. Lecture NoteinIn Computer

Science, vol. 1057. Springer-Verlag, London, 3-17.

Google Scholar Count in October 2006: 596

- 15. ~PrefixSpan

J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and

M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by

Prefix-Projected Pattern Growth. In Proceedings of the 17tInternationalal Conference on Data Engineering (April 02 - 06,

2001). ICDE '01. IEEE Computer Society, Washington, DC.

Google Scholar Count in October 2006: 248

---

- Integrated Mining

- 16. CBA

Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and

association rule mining. KDD-98, 1998, pp. 80-86.

http://citeseer.comp.nus.edu.sg/liu98integrating.html

Google Scholar Count in October 2006: 436

---

- Rough Sets

- 17. Finding reduct

Zdzislaw Pawlak, Rough Sets: Theoretical Aspects of Reasoning about

Data, Kluwer Academic Publishers, Norwell, MA, 1992

Google Scholar Count in October 2006: 329

---

- Graph Mining

- 18. gSpan

Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern

Mining. In Proceedings of the 2002 IEEE International Conference on

Data Mining (ICDM '02) (December 09 - 12, 2002). IEEE Computer

Society, Washington, DC.

Google Scholar Count in October 2006: 155

---

- 35-free-online-books

https://vitalflux.com/machine-learning-list-of-35-free-online-books/

---

- Supervised learning problems vs Unsupervised

Following are two differenkindsnd of supervised learning problems which are later associated with different solution approaches:

Numerical related problems in which one predicts the quantity (represented using numbers). Algorithms such as regression, SVM, neural network, decision tre,es etc are used to solvthese kindsnd of problems.

Classification related problems in which one predicts classes such as yes/no, positive/negative, good/bad/ug,ly etc. Algorithms such as logistic regression, SVM, neural network, K-NN, decision tre,es etc are used to solvthese kindsnd of problems

https://vitalflux.com/top-10-solution-approaches-for-supervised-learning-problems/

---

- READ

https://medium.com/predict/three-popular-clustering-methods-and-when-to-use-

three-popular-clustering-methods-and-when-to-use-each

https://medium.com/predict/three-popular-clustering-methods-and-when-to-use-each-4227c80ba2b6?mkt\_tok=eyJpIjoiWmpKaU5qZzBNVEU0TWpReCIsInQiOiJ6aDhvXC9WeFpJaTJDSEZNXC83eEtIT1NKN25OOFNaXC9uVDRhcGRrcXVEQ2gzMXBscWlaUER4azJqYkN2dGRBaVdMRUtINmRwZ1huTFFTcis3Z05qSzdGYU5BSm01VWcyM3d5eDNsSGFkQVwvSHQ4aE9LdEkyaWFFWHFTOG1FNmFDVG8ifQ%3D%3D

---

See also: http://www.deeplearningbook.org/lecture\_slides.html

Deep Learning

An MIT Press book in preparation

Ian Goodfellow, Yoshua Bengio and Aaron Courville

---

## EDA book by Roger Peng,

Chapter 5.

## Outline

Formulate your question

Read in your data

Check the packaging

Run str()

Look at the top and the bottom of your data

Check your "n"s

Validate with at least one external data source

Try the easy solution first

Challenge your solution

Follow up

[EDA](https://en.wikipedia.org/wiki/Exploratory\_data\_analysis)

- Formulate your question

- Do I have the right data to answer this question?

- Read in your data

- https://en.wikibooks.org/wiki/R\_Programming/Importing\_and\_exporting\_data#Excel\_.28xls.2Cxlsx.29

- readr

- data.table

- csv

- .xls|.xlsx

- Check the packaging

- dim(df)

- Run str()

- structure(df)

- Look at the top and the bottom of your data

- head

- tail

- Check your "n"s, STOPPED AT PAGE 22

- Validate YOUR data with at least one external data source

- data are at least of the right order of magnitude

- same unis, are units included...

- the distribution is normal, skewedPoissonon, etc.

- Try the easy solution first

- What can you glean froa firstst glan?e.

- simple ideas, linear relationships, correlations or not

- define: Collinearity; (http://faculty.cas.usf.edu/mbrannick/regression/Collinearity.html)

- remember Stats 101 discussionon oIndependentmvariableses

- Challenge your solution

- take a larger view of your data, do trends hold over time, over ppl, over other variables...

- Follow up

1. Do you have the right data?

2. Do you need other data?

3. Do you have the right question?

- Chapter 6. Principles of Analytic Graphics

SeEdwardrd Tufte, 'Beautiful Evidence'

Principle 1 - Data irelativeea to what?

- Positive and Negative Controls

- Compared to what??

Principle 2 - Show Causality

Principle 3 - Show Multivariate data comparisons

- is there interaction (synergy) between variables

Principle 4 - can you integrate evidence in many ways tbolsterer your claim

- use words, graphs, tables, numbers, etc

Principle 5 - make sure to document and describe evidence with proper labels scales, sourcesPrinciplele 6 - Content is King

- do you have the right data for your question(s)

Chapter 7 - Exploratory Graphs

- make them quick and make lots for ideas

- for personal understanding first

- quick and dirty graphs

- using color to diff 2D from 3D and 4D

Simple Summaries: One Dimension

- Five-number summary

- Boxplots

- Barplot

- Histograms

- Density plot

Simple Summaries: Two Dimensions and Beyond

- Multiple or overlayed 1-D plots (Lattice/ggplot2)

- Scatterplots:

- Use color, size, shape to add dimensions

- interactive plots

- multiple 2-D plots

Chapter 8 - Plotting

- plot, hist, boxplot

- Lattice System p. 64

- ggplot2 System

- Output: PDF, PostScript, PNG, etc.

Good software:

https://www.youtube.com/watch?v=ma6-0PSNLHo&feature=youtu.be

---

# What is Machine Learning?

Define: Where does the name come from and what does it produce.

## Give some basics

Explain

- Three elements of a machine learning

model

Model = Representation + Evaluation + Optimization

Domingos, P. A few useful things to know about machine learning.

Commun. ACM. 55(10):7887 (2012).

---

## Big Questions

### What is machine learning?

### Name applications of machine learning.

https://vitalflux.com/machine-learning-slides-beginners/

Machine Learning Overview

Machine Learning: An Overview: The slides presenan t introduction to machine learning along with some of the following:

Different types of learning (supervised, unsupervised, reinforcement)

Dimensions of a learning system (different types of feedback, representation, use of knowledge)

Supervised learning algorithms such as Decision tree, neural network, support vector machines (SVM), Bayesian network learning, nearest neighbor models

Reinforcement learning

Unsupervised learning

###. Explain the following machine learning areas and tasks: supervised learning, unsupervised learning, reinforcement learning, classification, regression, clustering, feature selection, feature extraction, and topic modeling.

###. Explain the following machine learning approaches: decision tree learning, artificial neural networks, and Bayesian networks.

### What are the advantages and disadvantages of the various approaches?

### Which kinds of products can be used for machine learning in practice?

### How to engineer a machine learning application?

### Explain precision, recall, and F-measure

### Why to separatehe t training data set and test data set for validation?

---

StepheneW wolfram - READ

https://blog.stephenwolfram.com/2017/05/machine-learning-for-middle-schoolers/

https://www.wolfram.com/language/elementary-introduction/2nd-ed/

---

- How-do-you-explain-Machine-Learning-and-Data-Mining-to-non-Computer-Science-people - READ

https://www.quora.com/How-do-you-explain-Machine-Learning-and-Data-Mining-to-non-Computer-Science-people/answer/Pararth-Shah

---

- Top 5 Algos for Bioinformatics / Data Science

1. Classification - CART: 952,000 results

2. Statistical Learning - SVM, 744,000 results

3. Association Analysis - Apriori, 144,000 results

4. Clustering - K-Means, 730,000 results

5. Bagging and Boosting - AdaBoost, 65,900 results

---

[10 Machine Learning Algorithms & R Commands](https://vitalflux.com/cheat-sheet-10-machine-learning-algorithms-r-commands/)

---

- "data mining" The phrase "data mining" is also sometimes used as a pejorative

to describe the deceptive practice of cherry-picking data to

support a theory.

Data mining should be used tcherry-pickck data to support a theory but to first approach the data with a theorthatch can be tested.

---

Remember: ML is good for classification of categorization, regression, dimensionality reduction, development of algorithmsthe , discovery of genetic sequences linked to diseases

---

- Supervised learning

''Supervised learning:'' Supervised learning is a form of evidence-based learning. The evidence is the known outcome for a given input and is in turn used to train the predictive model. Models are further classified into regression and classification, based on the outcome data type. In the former, the outcome is continuous, and in the latter the outcome is discrete. Stock trading and weather forecasting are some widely used applications of regression models, anspaman detection, speech recognition, and image classification are some widely used applications of classification models.

Some algorithms for regression are linear regression, Generalized Linear Models (GLM), Support Vector Regression (SVR), neural networks, decision trees, and so on; in classification, we have logistic regression, Support Vector Machines (SVM), Linear discriminant analysis (LDA), Naive Bayes, nearest neighbors, and so on.

''Semi-supervised learning:'' Semi-supervised learning is a class of supervised learning using unsupervised techniques. This technique is very useful in scenarios where the cost of labeling an entire dataset is highly impractical against the cost of acquiring and analyzing unlabeled data.

''Unsupervised learning:'' As the name suggests, learning from data with no outcome (or supervision) is called unsupervised learning. It is a form of inferential learning based on hidden patterns and intrinsic groups in the given data. Its applications include market pattern recognition, genetic clustering, and so on.

---

- Model evaluation

is a key step in any machine learning process. It is different for supervised and unsupervised models. In supervised models, predictions play a major role; where,as in unsupervised models, homogeneity within clusters and heterogeneity across clusters play a major role.

Some widely used model evaluation parameters for regression models (includincross-validationon) are as follows:

- Coefficient of determination

- Root mean squared error

- Mean absolute error

- Akaike or Bayesian information criterion

- Some widely used model evaluation parameters for classification models (includincross-validationon) are as follows:

- Confusion matrix (accuracy, precision, recall, and F1-score)

- Gain or lift charts

- Area under ROC (receiver operating characteristic) curve

- Concordant and discordant ratio

- Some of the widely used evaluation parameters of unsupervised models (clustering) are as follows:

- Contingency tables

- Sum of squared errors between clustering objects and cluster centers or centroids

- Silhouette value

- Rand index

- Matching index

- Pairwise and adjusted pairwise precision and recall (primarily used in NLP)

from: R Deep Learning Cookbook, PKS Prakash

---

- \*\*randomForest\*\*:

Provides methods for working with classification and regression problems, based on random forests algorithm which instructs creation of large number of bootstrapped trees on random samples of variables, classifying a particular case using all of these trees in this forest, and deciding al outcome based on average or majority voting techniques depending upon whether regression or classification problem is dealt with.

- \*\*e1071\*\*: Provides methods to work with regression and classification problems. Algorithms such as following are included as part of functions:

1. Support Vector Machines (SVM)

2. naïve Bayes classifier

3. Bagged clustering

4Short-timeme Fourier transform

---

- 35-free-online-books

https://vitalflux.com/machine-learning-list-of-35-free-online-books/

---

- Supervised learning problems vs Unsupervised

Following are two differenkindsnd of supervised learning problems which are later associated with different solution approaches:

Numerical related problems in which one predicts the quantity (represented using numbers). Algorithms such as regression, SVM, neural network, decision tre,es etc are used to solvthese kindsnd of problems.

Classification related problems in which one predicts classes such as yes/no, positive/negative, good/bad/ug,ly etc. Algorithms such as logistic regression, SVM, neural network, K-NN, decision tre,es etc are used to solvthese kindsnd of problems

https://vitalflux.com/top-10-solution-approaches-for-supervised-learning-problems/

---

- READ

https://medium.com/predict/three-popular-clustering-methods-and-when-to-use-

three-popular-clustering-methods-and-when-to-use-each

https://medium.com/predict/three-popular-clustering-methods-and-when-to-use-each-4227c80ba2b6?mkt\_tok=eyJpIjoiWmpKaU5qZzBNVEU0TWpReCIsInQiOiJ6aDhvXC9WeFpJaTJDSEZNXC83eEtIT1NKN25OOFNaXC9uVDRhcGRrcXVEQ2gzMXBscWlaUER4azJqYkN2dGRBaVdMRUtINmRwZ1huTFFTcis3Z05qSzdGYU5BSm01VWcyM3d5eDNsSGFkQVwvSHQ4aE9LdEkyaWFFWHFTOG1FNmFDVG8ifQ%3D%3D

---

See also: http://www.deeplearningbook.org/lecture\_slides.html

Deep Learning

An MIT Press book in preparation

Ian Goodfellow, Yoshua Bengio and Aaron Courville

---

y Interesting - Basic but very good for a sixteen yr old.

https://github.com/mjhendrickson/Learning-R/wiki

---

- https://www.datasciencecentral.com/profiles/blogs/seven-techniques-for-data-dimensionality-reduction

---

- supervised-learning-for-the-common-human

https://towardsdatascience.com/an-involved-introduction-to-supervised-learning-for-the-common-human-6338d9559748

---

- Videos for learning ML

https://www.datasciencecentral.com/profiles/blogs/neural-networks-for-machine-learning

---

A computer may be more capable than a human of finding subtle patterns in large databases, but it still needs a human to motivate the analysis and turn the result into meaningful action.

---

Q. How do you get better at formulating questions to ask your AI?

A. this is data exploratin, or exploratory data analysis. You can visualize, compute statistics, like mean, min/max, variance, covariance matrix, most frequent items, run some clustering algorithms. start from some kind of requirement, i.e. a "problem" as the target for your eventual solution. one thing that will make your life easier is if you reorganize your data in ways appropriate to your hypothesis and look for preliminary patterns using some very simple techniques. (https://www.reddit.com/r/MachineLearning/comments/15mh7q/question\_regression\_with\_many\_irrelevant\_variables/)

---

### USE IN ML

Predictive Modeling: Represents packages which help in working witvariousnt predictive models (linear/multivariate/logistic regression models, SVM, neural netwo,rk etc.)

caret: Stands for Classification And REgression TrainingIt provideses a set of functionthatch could be used to do some of the following when working with classification and regression problemsIt dependsds upothe n number of packages and loads these packages appropriately (on-demand) to achievthe e above objectives.

- Data processing (splitting)

- Feature selection

- Evaluate Model tuning parameters based on resampling

- Predictor variable importance estimation

- Estimate model performance from the training set

---

# What is Machine Learning? START 040

Define: Where does the name come from and what does it produce.

## Give some basics

Explain

- Three elements of a machine learning

model

Model = Representation + Evaluation + Optimization

Domingos, P. A few useful things to know about machine learning.

Commun. ACM. 55(10):7887 (2012).

---

## Big Questions

### What is machine learning?

### Name applications of machine learning.

https://vitalflux.com/machine-learning-slides-beginners/

Machine Learning Overview

Machine Learning: An Overview: The slides presenan t introduction to machine learning along with some of the following:

- Different types of learning (supervised, unsupervised, reinforcement)

- Dimensions of a learning system (different types of feedback, representation, use of knowledge)

- Supervised learning algorithms such as Decision tree, neural network, support vector machines (SVM), Bayesian network learning, nearest neighbor models

- Reinforcement learning

- Unsupervised learning

###. Explain the following machine learning areas and tasks: supervised learning, unsupervised learning, reinforcement learning, classification, regression, clustering, feature selection, feature extraction, and topic modeling.

###. Explain the following machine learning approaches: decision tree learning, artificial neural networks, and Bayesian networks.

### What are the advantages and disadvantages of the various approaches?

### Which kinds of products can be used for machine learning in practice?

### How to engineer a machine learning application?

### Explain precision, recall, and F-measure

### Why to separatehe t training data set and test data set for validation?

---

Stephenen wolfram - READ

https://blog.stephenwolfram.com/2017/05/machine-learning-for-middle-schoolers/

https://www.wolfram.com/language/elementary-introduction/2nd-ed/

---

- How-do-you-explain-Machine-Learning-and-Data-Mining-to-non-Computer-Science-people - READ

https://www.quora.com/How-do-you-explain-Machine-Learning-and-Data-Mining-to-non-Computer-Science-people/answer/Pararth-Shah

---

- Top 5 Algos for Bioinformatics / Data Science

1. Classification - CART: 952,000 results

2. Statistical Learning - SVM, 744,000 results

3. Association Analysis - Apriori, 144,000 results

4. Clustering - K-Means, 730,000 results

5. Bagging and Boosting - AdaBoost, 65,900 results

---

[10 Machine Learning Algorithms & R Commands](https://vitalflux.com/cheat-sheet-10-machine-learning-algorithms-r-commands/)

---

- "data mining" The phrase "data mining" is also sometimes used as a pejorative

to describe the deceptive practice of cherry-picking data to

support a theory.

Data mining should be used tcherry-pickck data to support a theory but to first approach the data with a theorthatch can be tested.

---

Remember: ML is good for classification of categorization, regression, dimensionality reduction, development of algorithmsthe , discovery of genetic sequences linked to diseases

---

- Supervised learning

''Supervised learning:'' Supervised learning is a form of evidence-based learning. The evidence is the known outcome for a given input and is in turn used to train the predictive model. Models are further classified into regression and classification, based on the outcome data type. In the former, the outcome is continuous, and in the latter the outcome is discrete. Stock trading and weather forecasting are some widely used applications of regression models, anspaman detection, speech recognition, and image classification are some widely used applications of classification models.

Some algorithms for regression are linear regression, Generalized Linear Models (GLM), Support Vector Regression (SVR), neural networks, decision trees, and so on; in classification, we have logistic regression, Support Vector Machines (SVM), Linear discriminant analysis (LDA), Naive Bayes, nearest neighbors, and so on.

''Semi-supervised learning:'' Semi-supervised learning is a class of supervised learning using unsupervised techniques. This technique is very useful in scenarios where the cost of labeling an entire dataset is highly impractical against the cost of acquiring and analyzing unlabeled data.

''Unsupervised learning:'' As the name suggests, learning from data with no outcome (or supervision) is called unsupervised learning. It is a form of inferential learning based on hidden patterns and intrinsic groups in the given data. Its applications include market pattern recognition, genetic clustering, and so on.

---

- Model evaluation

is a key step in any machine learning process. It is different for supervised and unsupervised models. In supervised models, predictions play a major role; where,as in unsupervised models, homogeneity within clusters and heterogeneity across clusters play a major role.

Some widely used model evaluation parameters for regression models (includincross-validationon) are as follows:

- Coefficient of determination

- Root mean squared error

- Mean absolute error

- Akaike or Bayesian information criterion

- Some widely used model evaluation parameters for classification models (includincross-validationon) are as follows:

- Confusion matrix (accuracy, precision, recall, and F1-score)

- Gain or lift charts

- Area under ROC (receiver operating characteristic) curve

- Concordant and discordant ratio

- Some of the widely used evaluation parameters of unsupervised models (clustering) are as follows:

- Contingency tables

- Sum of squared errors between clustering objects and cluster centers or centroids

- Silhouette value

- Rand index

- Matching index

- Pairwise and adjusted pairwise precision and recall (primarily used in NLP)

from: R Deep Learning Cookbook, PKS Prakash

---

- \*\*randomForest\*\*:

Provides methods for working with classification and regression problems, based on random forests algorithm which instructs creation of large number of bootstrapped trees on random samples of variables, classifying a particular case using all of these trees in this forest, and deciding al outcome based on average or majority voting techniques depending upon whether regression or classification problem is dealt with.

- \*\*e1071\*\*: Provides methods to work with regression and classification problems. Algorithms such as following are included as part of functions:

1. Support Vector Machines (SVM)

2. naïve Bayes classifier

3. Bagged clustering

4Short-timeme Fourier transform

---

- 35-free-online-books

https://vitalflux.com/machine-learning-list-of-35-free-online-books/

---

- Supervised learning problems vs Unsupervised

Following are two differenkindsnd of supervised learning problems which are later associated with different solution approaches:

Numerical related problems in which one predicts the quantity (represented using numbers). Algorithms such as regression, SVM, neural network, decision tre,es etc are used to solvthese kindsnd of problems.

Classification related problems in which one predicts classes such as yes/no, positive/negative, good/bad/ug,ly etc. Algorithms such as logistic regression, SVM, neural network, K-NN, decision tre,es etc are used to solvthese kindsnd of problems

https://vitalflux.com/top-10-solution-approaches-for-supervised-learning-problems/

---

- READ

https://medium.com/predict/three-popular-clustering-methods-and-when-to-use-

three-popular-clustering-methods-and-when-to-use-each

https://medium.com/predict/three-popular-clustering-methods-and-when-to-use-each-4227c80ba2b6?mkt\_tok=eyJpIjoiWmpKaU5qZzBNVEU0TWpReCIsInQiOiJ6aDhvXC9WeFpJaTJDSEZNXC83eEtIT1NKN25OOFNaXC9uVDRhcGRrcXVEQ2gzMXBscWlaUER4azJqYkN2dGRBaVdMRUtINmRwZ1huTFFTcis3Z05qSzdGYU5BSm01VWcyM3d5eDNsSGFkQVwvSHQ4aE9LdEkyaWFFWHFTOG1FNmFDVG8ifQ%3D%3D

---

See also: http://www.deeplearningbook.org/lecture\_slides.html

Deep Learning

An MIT Press book in preparation

Ian Goodfellow, Yoshua Bengio and Aaron Courville

---

y Interesting - Basic but very good for a sixteen yr old.

https://github.com/mjhendrickson/Learning-R/wiki

---

- https://www.datasciencecentral.com/profiles/blogs/seven-techniques-for-data-dimensionality-reduction

---

- supervised-learning-for-the-common-human

https://towardsdatascience.com/an-involved-introduction-to-supervised-learning-for-the-common-human-6338d9559748

---

- Videos for learning ML

https://www.datasciencecentral.com/profiles/blogs/neural-networks-for-machine-learning

---

A computer may be more capable than a human of finding subtle patterns in large databases, but it still needs a human to motivate the analysis and turn the result into meaningful action.

---

Q. How do you get better at formulating questions to ask your AI?

A. this is data exploratin, or exploratory data analysis. You can visualize, compute statistics, like mean, min/max, variance, covariance matrix, most frequent items, run some clustering algorithms. start from some kind of requirement, i.e. a "problem" as the target for your eventual solution. one thing that will make your life easier is if you reorganize your data in ways appropriate to your hypothesis and look for preliminary patterns using some very simple techniques. (https://www.reddit.com/r/MachineLearning/comments/15mh7q/question\_regression\_with\_many\_irrelevant\_variables/)

---

### USE IN ML

Predictive Modeling: Represents packages which help in working witvariousnt predictive models (linear/multivariate/logistic regression models, SVM, neural netwo,rk etc.)

caret: Stands for Classification And REgression TrainingIt provideses a set of functionthatch could be used to do some of the following when working with classification and regression problemsIt dependsds upothe n number of packages and loads these packages appropriately (on-demand) to achievthe e above objectives.

- Data processing (splitting)

- Feature selection

- Evaluate Model tuning parameters based on resampling

- Predictor variable importance estimation

- Estimate model performance from the training set

---

#---

#title: "Just Enough Machine Learning"

#author: "Matthew Curcio"

#date: "`r gsub(' 0', ' ', format(Sys.Date(), '%B %d, %Y'))`"

#site: bookdown::bookdown\_site

#output:

# bookdown::gitbook: default

# bookdown::pdf\_book: default

#github-repo: rstudio/bookdown-demo

#description: "Work based on 'Top 10 Algorithms In Data Mining' by X.Wu etal, Knowl Inf Syst (2008) 14:1-37 and 'Applied Predictive Modeling', Max Kuhn etal, Springer, 2018."

#---

```{r, include=FALSE}

library(knitr)

knitr::opts\_chunk$set(cache = TRUE,

warning = FALSE,

message = FALSE,

cache.lazy = FALSE)

```

# Preface {-}

#### Welcome {-}

This is booklet had startff as a hands-on Linux command line guide for new biology students. As I stepped back-and-forth, back-and-forth from that project I realized I was more interested in teaching myself machine learning using R and RStudio.

However, while I was reading on the topic of Linux, I found a quote that describes the philosophy of Unix.

>The Unix philosophy is one othe f simple tools, each doing one job well...

>Anonymous

This is \*\*NOT\*\* the philosophy of R nor RStudio. The R library repository, as of April 29, 2019, has more than 14,096 packages. I am still looking for an R package that fits the Unix motto.

#### Who This Book Is For {-}

This book is for me. It is very informal as is my conversational style. It is my record of what I believed would help me in the future. I plan on using this booklet as a template for other data science projects that I will dshortlyre.

#### Acknowledgments {-}

I would like to thank Doctors Dmitry Korkine and Elizabeth Ryder for their patience and generosity.

#### Dedication {-}

>To The Socratic Paradox,

>

>\*I know that I know nothing\*

#### About the Author {-}

Matthew C. Curcio is a Chemist, Biochemist, Teach,er and Student who loves science.

<div style="page-break-after: always;"></div>

#### Abstract ???????

Predictive modeling methos, supervised and unsupervised, have been introduced in 'Top 10 Algorithms In Data Mining' by X. Wet alal, 2008.[^i1] Unfortunately, there are no easy rules to follow for which models to use in all situations. The comically named 'No Free Lunch Theorems for Optimization'[REF] "demonstrates the danger of comparing [general-purpose] algorithms by their performance on a small sample of problems."[^i2] Therefore the author felt it is necessary to carry out an empirical study of 9 of the 10 machine learning algorithms which were deemed influential in the field of data mining by International Conference on Data Mining, which was polling source for the Wu paper. These algorithms include C5.0, k-Means, SVM, Prior, EM, AdaBoost, kNN, Naive Bayes, and CART. PageRank was not considered due to its in-applicability. These methods can be carried out using onall-encompassingng program known as \*caret\* written by Max Kuhet alal.[^i3] \*caret\* is an R language packagthatch allows one to use the same syntax for machine learning libraries written by over 200 different researchers also in the R language. These nine machine learning methods will be used on the classification of Oxygen binding proteins, eg. Hemoglobin, into six groupings as well as a negative control supplementing work carried out by S. Muthukrishnaet alal, which only investigates the utility of SVM to differentiate the 6 groups.[^i4] The machine learning tests were wilcompareed primarily using two measuresCPUpu-run time, and Cohen's kappa.

[^i1]:\*Top 10 algorithms in data mining\*, X. Wu, Knowl Inf Syst (2008) 14:1–37,

DOI 10.1007/s10115-007-0114-2

[REF]'No Free Lunch Theorems for Optimization'[REF]

[^i2]:\*No Free Lunch Theorems for Optimization\*, IEEE Transactions On Evolutionary Computation, Vol. 1,NO. 1, April 1997, D. Wolperet alal

[^i3]:\*The caret Package\*, M. Kuhn, 2019-03-27, https://topepo.github.io/caret/index.html

[^i4]:\*Oxypred: Prediction and Classification of Oxygen-Binding Proteins\*, S. Muthukrishn,aet alal, Geno. Prot. Bioinfo., Vol. 5 No. 3–4, 2007

---

# Prerequisites {#section-Prerequisites}

As a Bioinformatician, I also consider myself a data scientist. If you are not aware of the ubiquitous Venn diagram of Bioinformatics and Data Science, let's look at this for a moment.

```{r, echo=FALSE, fig.align='center', fig.asp=.75, fig.cap='Venn diagram of Bioinformatics and Data Science', message=FALSE, warning=FALSE, out.width='75%'}

knitr::include\_graphics('images/Bio-DS-Venn-diagram-1024x576.png')

```

To Biologists and Biochemists (which I also consider myself), it is my task to become more familiar with computer languages, miscellaneous concep,ts and statistics.

#### Regarding the Computing Side {-}

As of the writing of this booklet, Data Science and Bioinformatics are now centered around the computer languages [R](https://cran.r-project.org/) and [Python](https://www.anaconda.com/).

To many researchers in data science and bioinformatics the field now includes such languages as,

\*in no particular order\*,

- [R](https://cran.r-project.org/)

- [Python](https://www.anaconda.com/).

- [Bash shell scripting](https://www.datascienceatthecommandline.com/),

- [SQL](https://en.wikipedia.org/wiki/SQL),

- [Julia](https://julialang.org/),

- [Perl](https://www.perl.org/),

- [C](https://en.wikipedia.org/wiki/C\_%28programming\_language%29),

- [Javascript](https://www.javascript.com/),

- [RMarkdown](https://rmarkdown.rstudio.com/), (this one is fun and easy)

#### Just start with one language\*\* {-}

- Stick with it for a time and learn it. Learn the ins-and-outs of one language first.

#### Is R or Python better?\*\* {-}

1. R & Python are both FREE,

1. Both have great integrated development environments (IDEs),

- [RStudio](https://www.rstudio.com/) is great & FREE,

- [Spyder](https://www.spyder-ide.org/) is also great & FREE,

1. Both languages have been around for > 20 years, therefore both have tons of FREE information & tutorials on YouTube,

Ultimately, I chose to start with R because I liked the \*community\* around it.

## What you will need for this project?

1. An Internet connection,

2. A computer with between 1 & 3 Gigabytes of free hard drive storage,

- R-Packages & libraries \*can take a lot of space\*,

3. The program [R](https://cran.r-project.org/),

4. The program [RStudio](https://www.rstudio.com/).

## Download/Install R & RStudio

\*\*NOTE\*\*: I use [Ubuntu](https://www.ubuntu.com/).

1. Go to: https://cran.r-project.org/

```{r, echo=FALSE, fig.align='center', fig.asp=.75, fig.cap='https://cran.r-project.org', message=FALSE, warning=FALSE, out.width='65%'}

knitr::include\_graphics('images/R-cran-homepage.png')

```

1. Choose R for your operating system.

1. I recommend that you download/install 3 R files.

1. r-base-core\_3.#.# \* .deb (approx. 30 MB)

1. r-base-dev\_3.#.# \* .deb (approx. 45 KB)

1. r-base\_3.#.#-\*.deb (approx. 90 KB)

1. r-base-html-3.#.#-\*.deb (approx. 90 KB)

1. If you do have Linux you may try this video: [How to install R.](https://www.youtube.com/watch?v=LeB15G14N7g)

1. Go to: https://www.rstudio.com/

```{r, echo=FALSE, fig.align='center', fig.asp=.75, fig.cap='RStudio Home Page', message=FALSE, warning=FALSE, out.width='60%'}

knitr::include\_graphics('images/rstudio-homepage\_circles.png')

```

1. From RStudio's homepage click, \*Products\* then click \*RStudio\* from the dropdown menu.

1. Click the Download of the FREE version of RStudio Desktop

1. Click 'RStudio #.#' to download a version for your machine

Q. Have Linux? Try this - [How to install RStudio.](https://www.youtube.com/watch?v=LeB15G14N7g)

1. If you are looking for instructions for Mac & Windows machines try:

- [FreeCodeCamp](https://guide.freecodecamp.org/r/)

## Help! Where to find.

1. [Cheat Sheets](https://resources.rstudio.com/rstudio-cheatsheets)

1. https://community.rstudio.com

1. https://www.reddit.com/r/RStudio/

1. https://stackoverflow.com/

1. https://R-bloggers.com/

1. https://resources.rstudio.com/

1. [Rpubs.com](https://rpubs.com/)

- \*\*Rpubs.com\*\* contains R/RStudio notebooks and Markdown pages, VERY HELPFUL work from othepeople'ses online R documents. It is a way to learn from others and share your work. - Sign up, it is [FREE](https://rpubs.com/)! then press: \*Get Started\*

\*\*NOTE\*\*: If you are interested in seeing what others have published search Google, Rpubs.com does not have iwn search function. In Google, Search: `site:rpubs.com eda`

Other sites:

1. [Coursera]()

1. https://www.quora.com/What-is-data-science

1. [Roger Peng's EDA](https://youtu.be/ZXrHkIz-krE?t=15)

1. [Bookdown](https://bookdown.org/).

By the way, LEARN GIT!

## Resources for learning Data Science

The Lean Publishing (https://leanpub.com) company contains a library in the form of FREdownloadablele books/pdfs. I recommend;

1. How to be a modern scientist[^i5] by Jeffrey Leek[^i6]

2. R Programming for Data Science[^i7] by Roger Peng[^i8]

3. Exploratory Data Analysis with R[^i9] by Roger Peng

4. Data Analysis for the Life Sciences[^i10] by Rafael Irizarry & Michael Love

[^i5]:https://leanpub.com/modernscientist

[^i6]:http://jtleek.com

[^i7]:https://leanpub.com/rprogramming

[^i8]:https://simplystatistics.org

[^i9]:https://leanpub.com/exdata

[^i10]:https://leanpub.com/dataanalysisforthelifesciences

## Load all libraries used in this project

If you copy the text below intan a RMarkdown file (.rmd) you will be able to load all the libraries you will need without the bother of re-installing libraries.

---

Copy from here:

```{r eval=FALSE}

install.packages("rlang")

library(rlang)

knitr::opts\_chunk$set(cache = TRUE,

warning = FALSE,

message = FALSE,

cache.lazy = FALSE)

# If you are using Ubuntu Linux may need these Linux libraries first.

# Install: libcurl4-openssl-dev, libssl-dev, libxml2-dev

load\_or\_install <- function(package\_names) {

for(package\_name in package\_names) {

if(!is\_installed(package\_name)) {

install.packages(package\_name, repos = "http://lib.stat.cmu.edu/R/CRAN",

dependencies = TRUE)

}

# Uncomment line below to load libraries.

# library(package\_name,character.only=TRUE,quietly=TRUE,verbose=FALSE)

print("OK")

}

}

load\_or\_install(c("readr", "doMC", "corrplot", "knitr", "caret", "tidyverse"))

```

```{r eval=FALSE}

load\_or\_install(c("ggplot2", "C50", "rmarkdown", "bookdown", "blogdown", "kernlab", "mclust"))

```

```{r eval=FALSE}

load\_or\_install(c("mvtnorm", "fastAdaboost", "e1071", "plyr", "RColorBrewer"))

```

```{r eval=FALSE}

load\_or\_install(c("rpart", "Cubist", "kknn", "class", "klaR", "MASS"))

```

```{r eval=FALSE}

load\_or\_install(c("LogicReg", "naivebayes", "bnclassify", "randomForest", "foreach"))

```

```{r eval=FALSE}

load\_or\_install(c("import", "ipred", "dplyr", "stringr", "stringi", "cluster", "factoextra"))

```

```{r eval=FALSE}

load\_or\_install(c("tidyr", "lubridate", "ggplot2", "htmlwidgets", "zoo", "xtable"))

```

End copy here.