ANALYTIXLABS

Machine Learning: Introduction

Introduction to Machine Learning

1. **What is Machine Learning & Use Cases**
2. **Major Classes of Learning Algorithms:**
   - Supervised, Unsupervised, Semi- Supervised, Reinforced
3. **Building a Machine Learning Model**
   - Data Pre-processing, Training & Test Split, Model Building, Validation, Prediction
4. **Performance Metrics & Evaluation – Concept of:**
   - Over/Under-fitting;
   - Bias, Variance, and Trade off
   - Regularization
   - Cross Validation

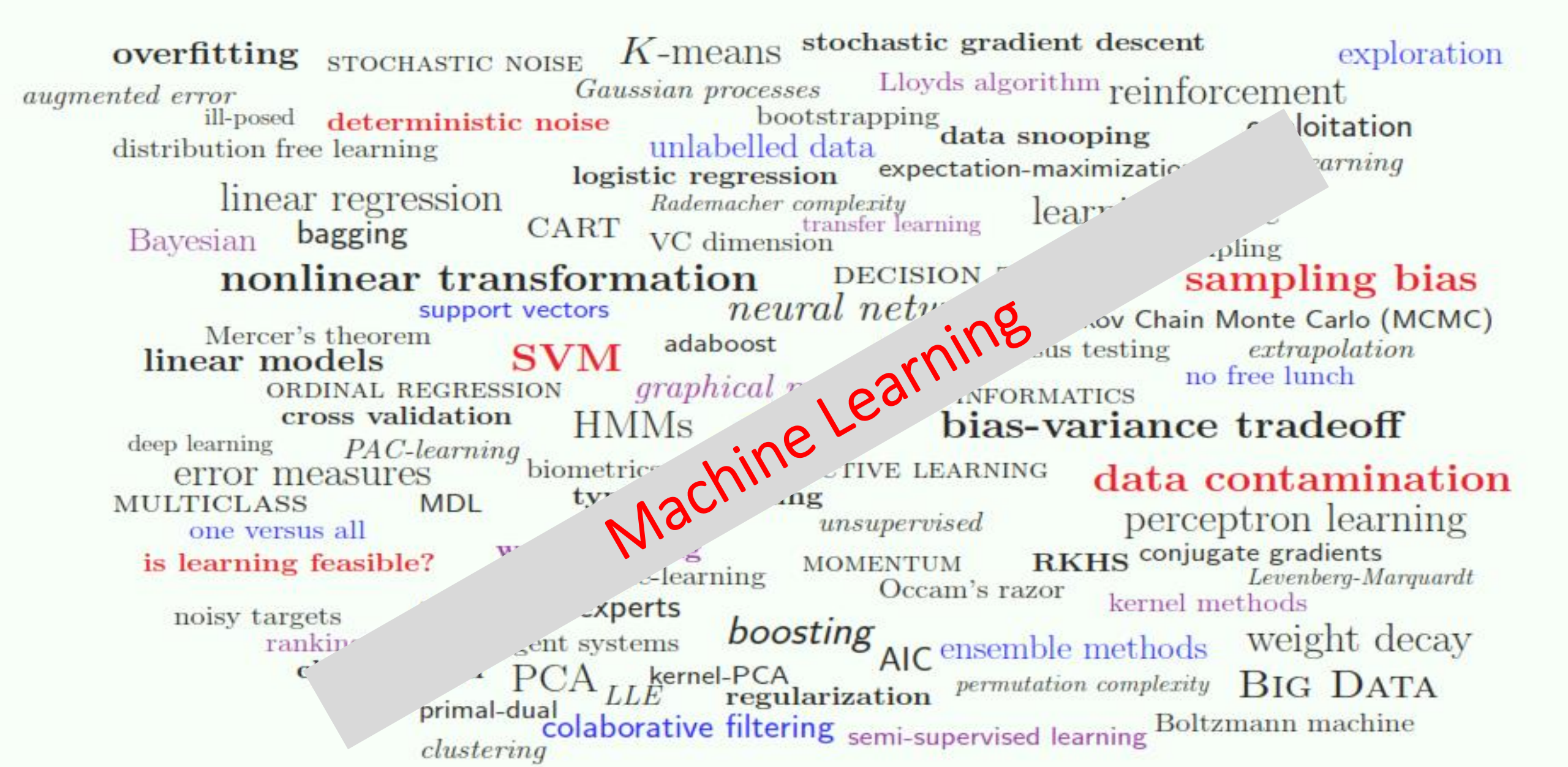
A. What is Machine Learning?

Context

Around 90% of WW data is generated within the last 2 years!!

What do we do with that?

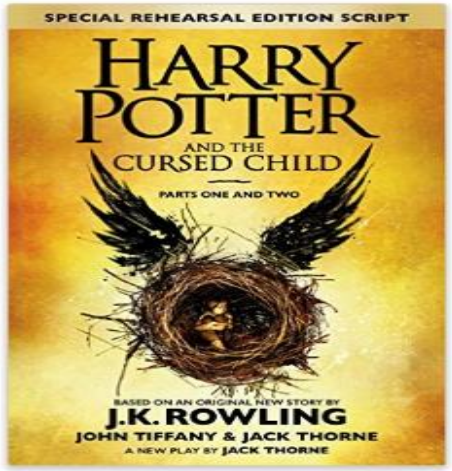
- Data is an asset for any organization to generate *actionable insights*
- Data is used for MIS reporting at the lowest level to building predictive & prescriptive models at the higher levels
- Machine Learning (ML) is an integral part of predictive & prescriptive modeling


Machine Learning
overfitting
STOCHASTIC NOISE
K-means
stochastic gradient descent
exploration
augmented error
Gaussian processes
Lloyds algorithm
reinforcement
ill-posed
deterministic noise
bootstrapping
distribution free learning
unlabelled data
data snooping
logistic regression
linear regression
Rademacher complexity
transfer learning
Bayesian
bagging
CART
VC dimension
nonlinear transformation
DECISION
sampling bias
support vectors
neural
Mercer's theorem
adaboost
linear models
SVM
extrapolation
ORDINAL REGRESSION
no free lunch
cross validation
HMMs
bias-variance tradeoff
deep learning
PAC-learning
error measures
data contamination
MULTICLASS
MDL
perceptron learning
one versus all
unsupervised
is learning feasible?
MOMENTUM
RKHS
conjugate gradients
Levenberg-Marquardt
Occam's razor
noisy targets
kernel methods
boosting
weight decay
AIC
ensemble methods
PCA
kernel-PCA
permutation complexity
BIG DATA
LLE
regularization
primal-dual
colaborative filtering
semi-supervised learning
Boltzmann machine
clustering

Amazon e-Marketplace

Harry Potter and the Cursed Child - Parts I & II and over 2 million other books are available for **Amazon Kindle** .

Books › Children's & Young Adult › Fantasy, Science Fiction & Horror



See all 2 images

Harry Potter and the Cursed Child - Parts I & II (Special Rehearsal Edition) Hardcover – 1 Aug 2016

by J.K. Rowling (Author), Jack Thorne (Author), John Tiffany (Author)

★★★★☆ 658 customer reviews

See all 3 formats and editions

Kindle Edition	Hardcover
₹ 881.75	₹ 534
Read with Our Free App	1 Used from ₹ 450.00 82 New from ₹ 515.00

Delivery to pincode **400001** - Mumbai within **2 - 4 business days**. Details

LIMITED QUANTITY

The order quantity for this product is limited to 2 units per customer

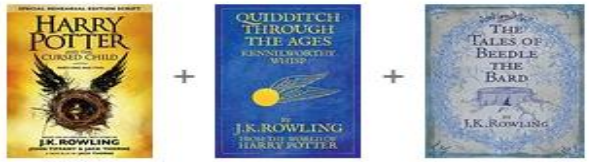
Please note that orders which exceed the quantity limit will be auto-canceled. This is applicable across sellers.

The Eighth Tale in the Harry Potter Saga

Being labelled as 'the boy who lived' for his whole life has not been easy for Harry Potter. In the official eighth instalment of the Harry Potter series penned in the form of a two-part stage production play, J. K. Rowling weaves yet another thrilling and magical yarn featuring the life of Harry Potter nineteen years later in the post-Voldemort wizarding world.

Read more

Frequently Bought Together



Total price: ₹1,224.21

Add all three to Cart

These items are dispatched from and sold by different sellers. Show details

- **This item:** Harry Potter and the Cursed Child - Parts I & II (Special Rehearsal Edition) by J.K. Rowling Hardcover ₹534.00
- Quidditch Through the Ages: From The World of Harry Potter by J.K. Rowling Paperback ₹279.00
- The Tales of Beedle the Bard by J.K. Rowling Hardcover ₹411.21

Google News

Google

News

India edition | Modern | हिन्दी | தமிழ் | മലയാളം | తెలుగు

Top Stories
- Mother Teresa
- Zika virus
- SpaceX
- Hillary Clinton
- Serena Williams
- China
- Sandeep Kumar
- Vladimir Putin
- Florida
- Narendra Modi

Bengaluru, Karnataka

India

World

Business

Technology

Entertainment

Sports

Science

Health

More Top Stories

Top Stories


Economic Times

Pope Francis declares Mother Teresa a saint

Economic Times - 1 hour ago

VATICAN CITY: Pope Francis declared Mother Teresa a saint on Sunday, honoring the tiny nun who cared for the world's most destitute and holding her up as a model for a Catholic Church that goes to the peripheries to find poor, wounded souls.

Canonisation Over, Saint Terasa Of Calcutta Hailed By Thousands In Vatican NDTV

Live: St. Teresa spent her life bowing down before those left to die, says Pope Francis The Hindu

Opinion: Saint Teresa of Calcutta New York Daily News

Related
Mother Teresa »

See realtime coverage


euronews


Associated Press


Times of India

'Need to respect each other's aspirations,' Narendra Modi, Xi Jinping tell each other

Times of India - 1 hours ago

NEW DELHI: India and China should be sensitive to each other's aspirations and needs, Prime Minister Narendra Modi told Chinese President Xi Jinping when the two leaders met on the sidelines of the G20 Summit+ in Huangzhou on Sunday.


Financial E...

JKNPP rues all-party delegation meeting parties only in Kashmir

The Hindu - 9 hours ago

Jammu and Kashmir National Panthers Party (JKNPP) on Saturday condemned the Union and state governments "decision" to allow political parties to meet the all-party delegation only in Kashmir Valley, terming the move as an "insult" to political ...


The Hindu

Women's groups want Personal Law Board to apologise, withdraw talaq affidavit in SC

Times of India - 16 hours ago

ALIGARH: Women's groups have condemned the propositions put up by the All India Muslim Personal Law Board (AIMPLB) to the Supreme Court in their affidavit justifying triple talaq and polygamy, calling the arguments "patriarchal, inhuman and unjust".


The Indian ...

Javadekar rubbishes Congress allegations of Modi govt indulging in politics of vendetta

The Indian Express - 18 hours ago

Javadekar said that the Congress "appeared to be in panic" after an interview given by Prime Minister Narendra Modi and was making baseless allegations.

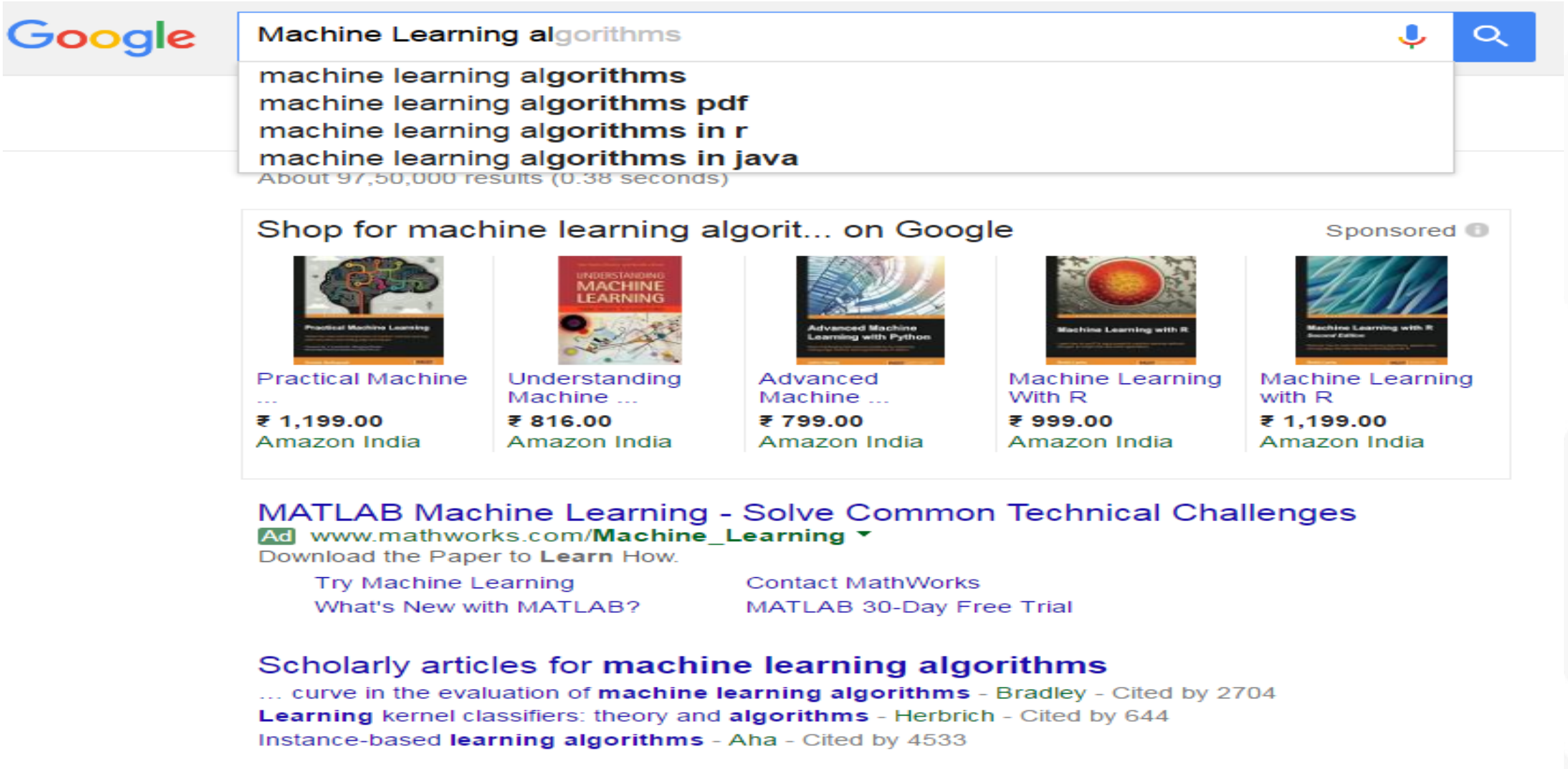

Times of India

Cong cautious on Sidhu, but confident of poll win

Times of India - 9 hours ago

CHANDIGARH: The Congress in Punjab on Saturday treaded cautiously on the new front floated by former ace cricketer Navjot Singh Sidhu amid fears that he can set up a formidable line-up of candidates in state's Majha region, where party hopes to encash ...

Google Search



Applications of Machine Learning

- Banking / Telecom / Retail
  - Identify:
    - Prospective customers
    - Dissatisfied customers
    - Good customers
    - Bad payers
  - Obtain:
    - More effective advertising
    - Less credit risk
    - Fewer fraud
    - Decreased churn rate
- Computer / Internet
  - Computer interfaces:
    - Troubleshooting wizards
    - Handwriting and speech
    - Brain waves
  - Internet
    - Hit ranking
    - Spam filtering
    - Text categorization
    - Text translation
    - Recommendation
- Biomedical / Biometrics
  - Medicine:
    - Screening
    - Diagnosis and prognosis
    - Drug discovery
  - Security:
    - Face recognition
    - Signature / fingerprint / iris verification
    - DNA fingerprinting

Example: Credit Card Approval

We would like to be able to predict customers those are likely to default

Application Information:

age	23 years
gender	Male
annual salary	USD 60,000
year in residence	1 year
year in job	0.5 year
current debt	USD 5,000

unknown pattern to be learned:

‘approve credit card good for bank?’

Example: Credit Card application approval problem

- input: $\mathbf{x} \in X$ (customer application)
- output: $y \in Y$ (good/bad after approving credit card)
- unknown pattern to be learned ⇔ target function:

$f : X \rightarrow Y$ (ideal credit approval formula)

- data ⇔ training examples: $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)\}$

(historical records in bank)

- hypothesis ⇔ skill with hopefully good performance:

$g : X \rightarrow Y$ ('learned' formula to be used)

$\{(\mathbf{x}_n, y_n)\}$ from f → ML → g

Use-Cases (Examples)

- Spam Email Detection – any mail providers
- Machine Translation (Language Translation) – e.g. google translator
- Image Search (Similarity): search engines
- Clustering (K-Means) : Amazon Recommendations
- Classification : Google News
- Text Summarization - Google News
- Rating a Review/Comment: Yelp
- Fraud detection : Credit card Providers
- Decision Making : e.g. Bank/Insurance sector
- Sentiment Analysis: e.g. product reviews (Visual analytics tools)
- Speech Understanding – iPhone with Siri
- Face Detection – Facebook's Photo tagging
- Image processing & Pattern recognition

Where we need Machine Learning?

Primarily Machine Learning can be applied where:

✓Explicit coding is not possible due to hidden patterns

✓There is issue related to large Complex data sets (*Scalability*)

Where Machine Learning cannot be applied:

- Optimization related problems where we have explicit mathematical relations

What is Machine Learning?

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed – (Arthur Samuel, 1959)

So how does computers learn?

- According to Tom Mutchell (1998): *"A computer learns effectively if its performance on certain **tasks** improves with **experience** - as measured by some **performance metrics**"*

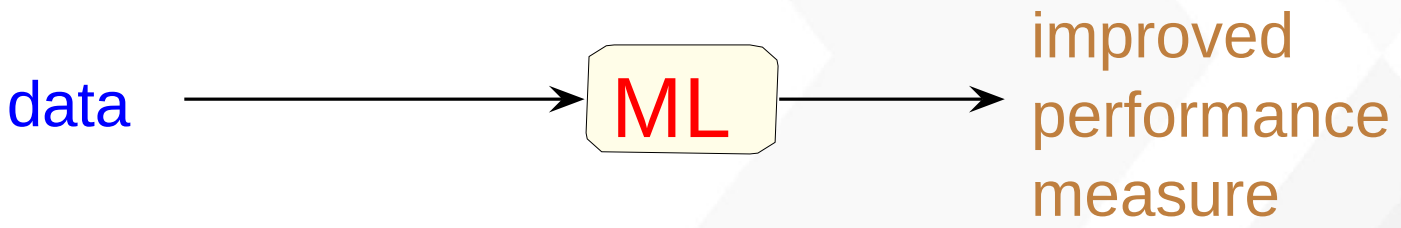
Human learning: Skills with experience accumulated from observations

machine learning: acquiring skill with experience accumulated/**computed** from data

A More Concrete Definition …

Skill ⇔ improve some performance measure (e.g. prediction accuracy)

machine learning: improving some performance measure(s) with experience **computed** from data



An Application in Computational Finance

Machine Learning vs. Statistics, Artificial Intelligence (AI), Data Mining

ML is closely associated with Statistics, AI, and Data Mining

Machine Learning Vs. Statistics

- Traditional Statistics focuses on provable results with math assumptions, and care less about computation
- “Statistics: A useful tool for Machine Learning"

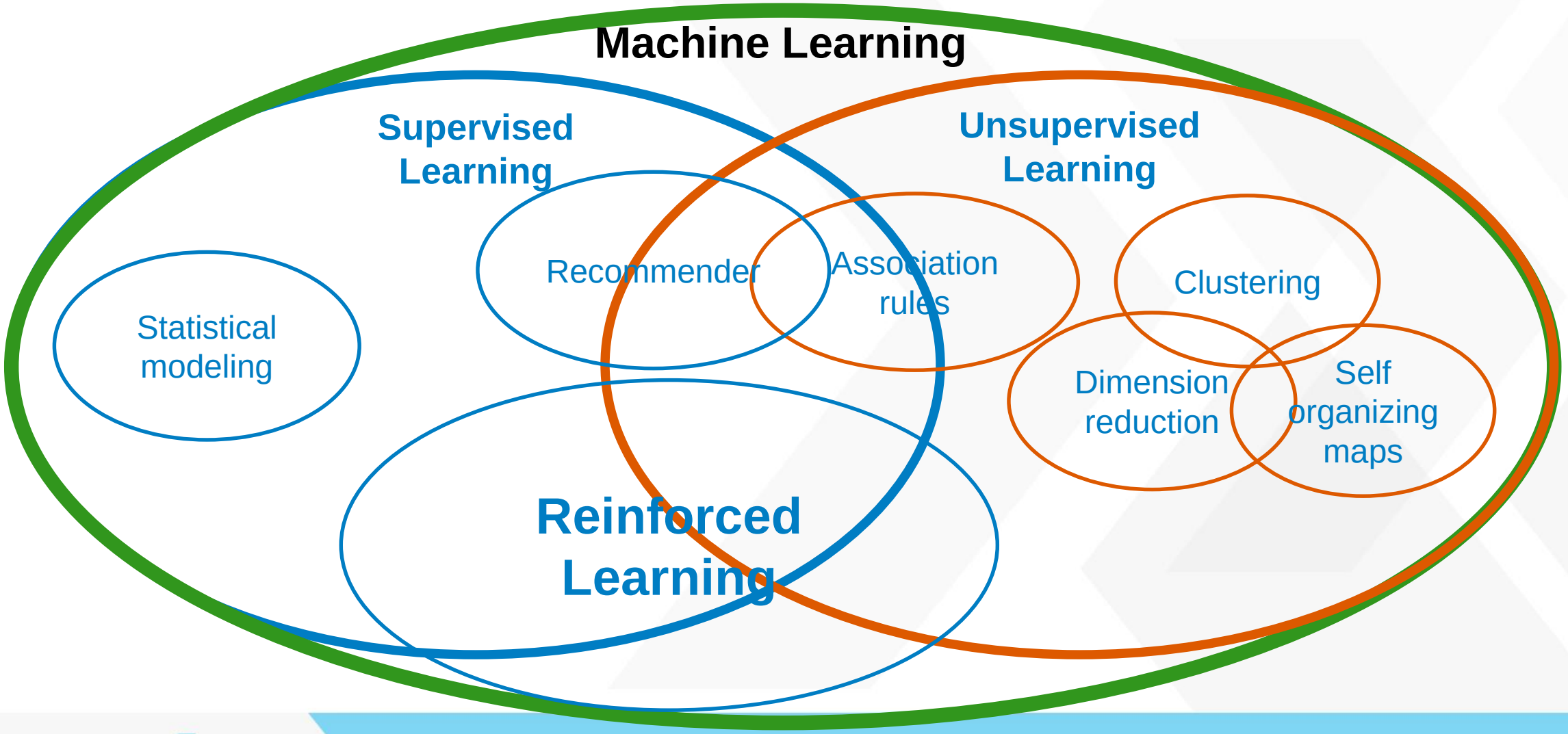
Machine Learning Vs. Artificial Intelligence

- " Machine Learning is one possible route to realize AI"

Machine Learning Vs. Data Mining (DM)

- Traditional DM focuses on provable results with math assumptions along with efficient computation in large database
- “Difficult to distinguish ML and DM in reality"

SO, HOW COMPUTERS LEARN IN MACHINE LEARNING?



Types of Learning

- **Supervised (*Inductive*) Learning**
  - Training data includes desired outputs
  - Labelled input data.
  - Creating classifiers to predict unseen inputs.
- **Unsupervised Learning**
  - Training data does not include desired outputs
  - Unlabelled input data.
  - Creating a function to predict the relation and output
- **Semi-supervised Learning**
  - Training data includes a few desired outputs
  - Combines Supervised and Unsupervised Learning methodology
- **Reinforcement Learning** (e.g., Markov Decision Process, Q Learning etc.)
  - No label is provided, but only indicates if a label is correct or not
  - Direct Rewards from sequence of actions
  - Reward-Punishment based agent

Supervised Learning

1. Supervised (inductive) Learning

Target function

Patient Age	Tumor Size	Clump Thickness	…	Malignant?
55	5	3		TRUE
70	4	7		TRUE
85	4	6		FALSE
35	2	1		FALSE
…	…	…	…	FALSE
				TRUE

Training Data

Labeled Data-Set

LEARNING

Model

Applications

- Classifier
- Spam Detection
- Information Retrieval
  - Personalisation based on ranks
- Speech Recognition

- Training data includes both predictors (Xi), and response (Yi)
- Labelled input data.
- Creating classifiers to predict unseen inputs.

Patient age	Tumor size	Clump	…	Malignant
72	3	3		?
66	4	4		?

Test Data

Supervised Learning - Algorithms

- Linear Regression
- Logistic Regression
- Decision Trees (CHAID, CART & Random Forest)
- k-Nearest Neighbours (KNN)
- Naive Bayes (Bayesian Learning)
- Discriminant Analysis (LDA/QDA) – Classification using linear regression
- Neural Networks
- SVM and Kernel estimation
- Perceptron and Multi-level Perceptrons (ANN – Deep Learning)
- Ensemble Models
- …

Un-supervised Learning

2. Un-Supervised Learning to detect natural patterns

Age	State	Annual Income	Marital status
25	CA	$80,000	M
45	NY	$150,000	D
55	WA	$100,500	M
18	TX	$85,000	S
…	…	…	…

Naturally occurring (hidden) structure

No Label

- Un-labelled input data.
- In this situation only the Xi's are observed.
- We need to use the Xi's to guess what Yi would have been and build a model from there.
- Finding hidden structure in data
- Creating a function to predict the relation and output

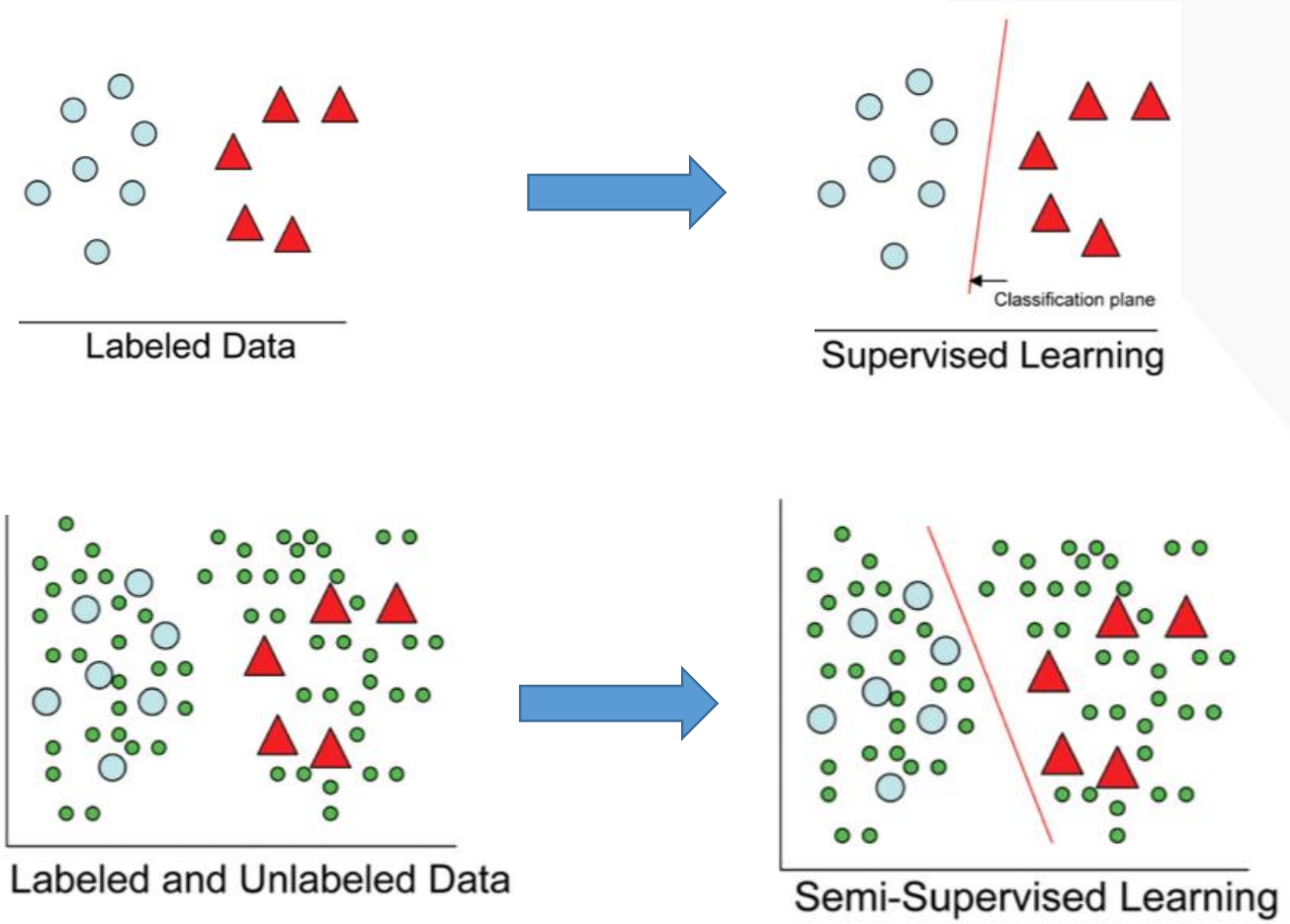
Applications

- Market segmentation - divide potential customers into groups based on their characteristics
- Pattern Recognition
- Groupings based on a distance measure
  - Group of People, Objects, …

Unsupervised Learning - Algorithms

- Clustering
  - k-Means, Hierarchical Clustering
- Hidden Markov Models (HMM)
- Dimension Reduction (Factor Analysis, PCA)
- Feature Extraction methods
- Self-organizing Maps (Neural Nets)
- …

Semi-Supervised Learning



Semi-Supervised Learning

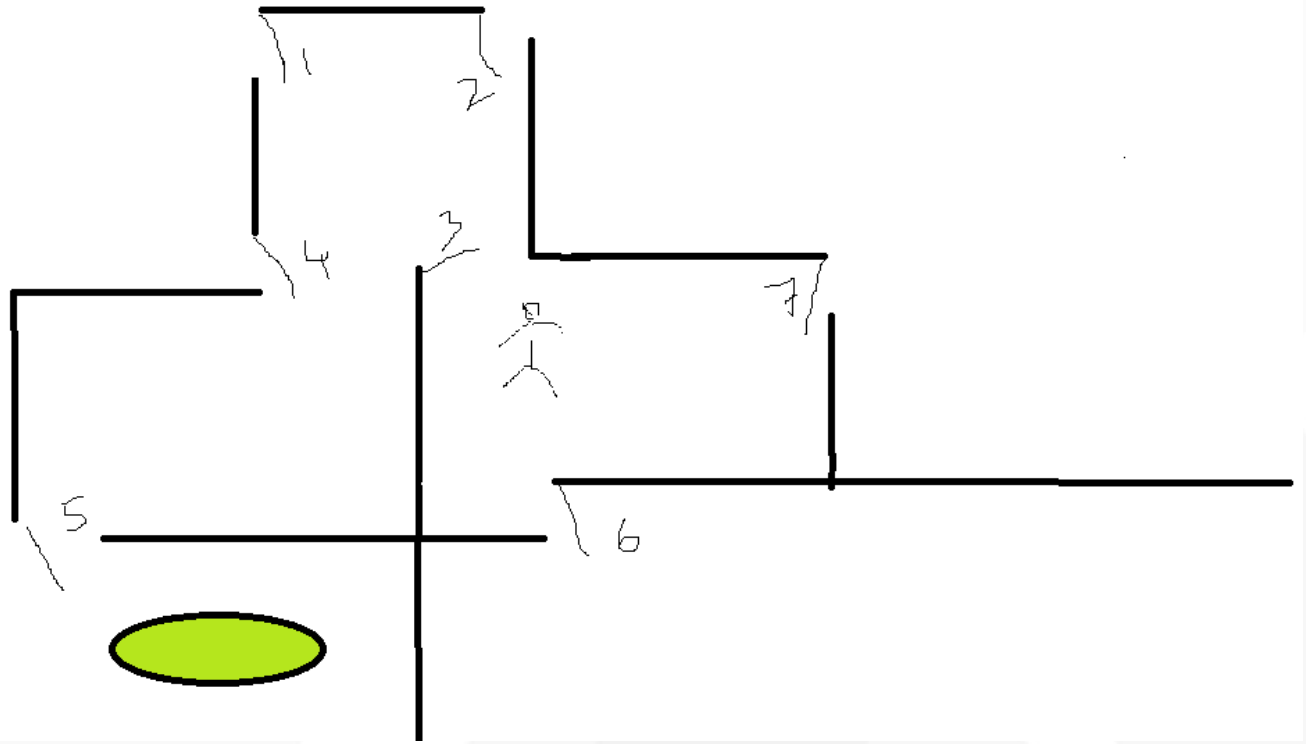
- Training data includes a few desired outputs
- Combines Supervised & Unsupervised Learning Methodologies

Application

- Webpage Classification:

Reinforced Learning

- No label is provided, but only indicates if a label is correct or not
- Direct Rewards from sequence of actions
- Reward-Punishment based agent
- e.g: Markov Decision Process, Q Learning etc.



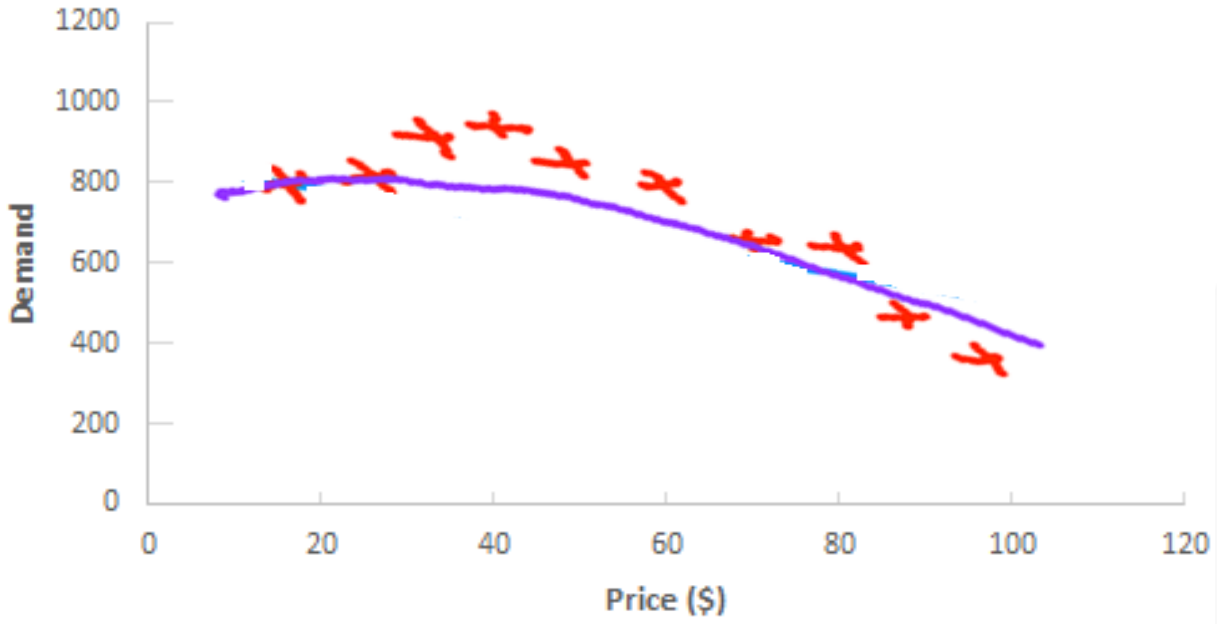
Machine Learning Problems

Broadly, the entire range of ML problems can be of TWO types:

Predicting a *Value* or an *Event/Class (Classifier)*

Regression: Predict a Continuous Value

- Predicting movie score
- Predicting price
- Predicting credit line approval
- Predicting stock prices
- Predicting temperature
- Predicting how many days to recover from sickness



Some techniques:

- Linear Regression / GLM
- Decision Trees
- Support vector regression
- Ensembles
- Etc…

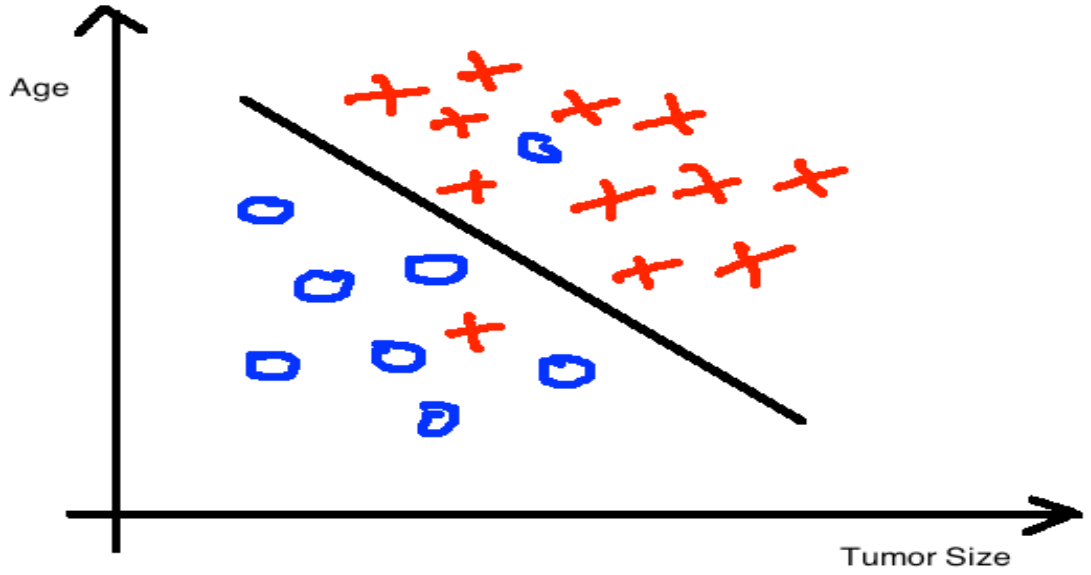
Classification: Predicting a Category

Binary Problems

- credit approve/disapprove
- email spam/non-spam
- patient sick/not sick
- ad profitable/not profitable
- answer correct/incorrect

Multi Class Problems

- written digits ⇒ 0, 1, . . . , 9
- pictures ⇒ apple, orange, strawberry
- emails ⇒ spam, primary, social, promotion, update (Google)



Some techniques:

- Naïve Bayes
- Decision Tree
- Logistic Regression/GLM
- Support Vector Machines
- Neural Network
- Ensembles
- LDA, QDA
- Etc…

Product Affinity & Recommendation

A. Product-to-Product Affinity

	PL1	PL2	...	...	...	...	...
PL1							
PL2							
...							
...							
...							
...							
...							

C. Customer-to-Product Propensity

	PL1	PL2	PL3	...	...	...	...
C1							
C2							
C3							
...							
...							
...							
...							

B. Identifying frequent item sets

	Item 1	Item 2	Item 3	Item 4	Item 5
Tx 1	Y	N	N	Y	N
Tx 2	Y	N	N	Y	N
Tx 3	Y	Y	N	Y	N
Tx 4	N	N	Y	Y	Y
Tx 5					
...					

	Item 1	Item 2	Item 3	Item 4	Item 5
Tx 1	Y	N	N	Y	N
Tx 2	Y	N	N	Y	N
Tx 3	Y	Y	N	Y	N
Tx 4	N	N	Y	Y	Y
Tx 5					
...					

Some techniques:

- Market Basket Analysis
- FP Growth
- A-priori Algorithm
- Collaborative Filtering
- Etc…

B. Building a Machine Learning Application

Building a Machine Learning Application

1. Formulating the problem
2. Data Tidying & Pre-processing
3. Training-Test Split thru' Sampling
4. Model Building
5. Validation & Model Accuracy
6. Using the chosen model for prediction

1. Formulating the problem

- Define a clear problem statement as per business need
- Is this a ‘Value’ or ‘Event/Class’ prediction?

2. Data Tidying & Pre-processing

2. Data Tidying & Pre-processing Data

1. Around 80% of data analysis time is spent on the process of cleaning & preparing the data
2. Data cleaning starts with data tidying where **each variable is a column**, **each observation is a row**, and each type of observational unit is a table
3. Fixed variables should come first, followed by measured variables, each ordered so that related variables are contiguous. Rows can then be ordered by the first variable, breaking ties with the second and subsequent (fixed) variables

Table-1

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Table-2

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Table-3

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Tidy

Source: *Tidy Data | Wickham H | Journal of Statistical Software*

2. Data Tidying & Pre-processing Data (Contd.)

Common Errors…

1. Column headers are values, not variable names…

TS Data

Product	1-Jan-15	2-Jan-15	3-Jan-15	4-Jan-15	5-Jan-15	6-Jan-15
XAZ256	27	34	60	81	76	137
XAZ256	12	27	37	52	35	70
XAZ256	27	21	30	34	33	58
XAZ256	418	617	732	670	638	1116
XAZ256	1	9	7	9	11	34
XAZ256	20	27	24	24	21	30
XAZ256	19	19	25	25	30	95

Modeling & Analysis becomes easy with this structure…

Product	DATE	Qty
XAZ256	1/1/2015	27
XAZ256	1/2/2015	34
XAZ256	1/3/2015	60
XAZ256	1/4/2015	81
XAZ256	1/5/2015	76
XAZ256	1/6/2015	137
XAZ256	1/1/2015	12
XAZ256	1/2/2015	27

***Source**: Tidy Data | Wickham H | Journal of Statistical Software*

2. Data Tidying & Pre-processing Data (Contd.)

Common Errors...

2. Multiple variables stored in one column...

WHO Data on some disease as extracted

Country	Year	Candidate	Cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

Modeling & Analysis becomes easy with this structure...

Country	Year	Gender	Age	Cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

***Source**: Tidy Data | Wickham H | Journal of Statistical Software*

2. Data Tidying & Pre-processing Data (Contd.)

Common Errors...

3. Variables are stored in both rows and columns…

Weather data from Global Climatology Network

ID	Date	Element	Value
MX17004	1/30/2010	tmax	27.8
MX17004	1/30/2010	tmin	14.5
MX17004	2/2/2010	tmax	27.3
MX17004	2/2/2010	tmin	14.4
MX17004	2/3/2010	tmax	24.1
MX17004	2/3/2010	tmin	14.4
MX17004	2/11/2010	tmax	29.7
MX17004	2/11/2010	tmin	13.4
MX17004	2/23/2010	tmax	29.9
MX17004	2/23/2010	tmin	10.7

Modeling & Analysis becomes easy with this structure...

ID	Date	Tmax	Tmin
MX17004	1/30/2010	27.8	14.5
MX17004	2/2/2010	27.3	14.4
MX17004	2/3/2010	24.1	14.4
MX17004	2/11/2010	29.7	13.4
MX17004	2/23/2010	29.9	10.7
MX17004	3/5/2010	32.1	14.2
MX17004	3/10/2010	34.5	16.8
MX17004	3/16/2010	31.1	17.6
MX17004	4/27/2010	36.3	16.7
MX17004	5/27/2010	33.2	18.2

***Source**: Tidy Data | Wickham H | Journal of Statistical Software*

After data tidying, it's time for Data Pre-Processing

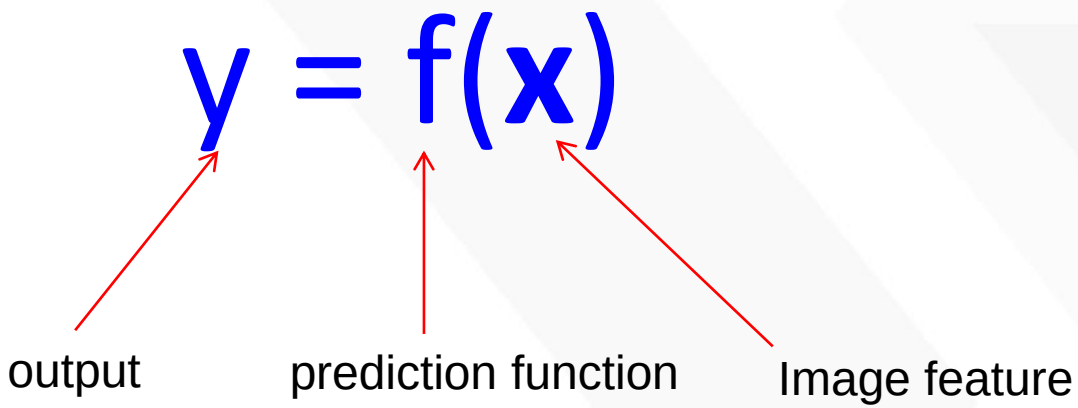
1. **Filter:** sub-setting or removing observations based on some condition.
2. **Aggregate:** collapsing multiple values into a single value (e.g., by summing or taking means).
3. **Missing Value Treatment**
   - Zero, Series average, Neighborhood Average/Median/Mode, Regression, Business Logic, etc.
4. **Outlier Treatment**
   - Std. Deviation Method:
     - ❖ Mean ±2.5 Std. Deviation (sample size < 80), and Mean ±3.0 Std. Deviation (sample size > 80)
   - Interquartile Range IQR: Median ± 2 * IQR
   - Replacement by steps similar to MVT
5. **Data Modification**
   - Standardize/Normalize Data
6. **Data Reduction**
   - Dimensionality Reduction (e.g. PCA)
7. **Feature Creation/Transformation**
   - Transformation (log, polynomial, multiply features)
   - Kernel

3. Training-Test Split thru' Sampling

- A random Sampling is done to split the data into training & test data
- A time series data should not be sampled randomly, it should be contiguous in nature
  - One time period might depending all time periods previously
- Well distributed balanced sample
  - Over & Under-Sampling need to Carried out for an imbalanced data!!

4. Model Building & Training

The Machine Learning framework



- **Training:** given a *training set* of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set
- **Testing:** apply f to a never before seen *test example* $\mathbf{x}$ and output the predicted value $y = f(\mathbf{x})$

Methods includes Parametric Methods and Non-parametric Methods

Machine Learning framework: *Parametric vs Non-parametric*

- **Parametric:** Algorithms that assumes a known form of function are called parametric machine learning algorithms.
- The algorithms involve two steps:
  - Select a form for the function.
  - Learn the coefficients for the function from the training data.
  - E.g., Linear and Logistic Regression.
- **Non-Parametric:** Algorithms that do not make strong assumptions about the form of the mapping function are called nonparametric machine learning algorithms.
  - By not making assumptions, they are free to learn any functional form from the training data.
  - Non-parametric methods are often more flexible, achieve better accuracy but require a lot more data and training time.
  - E.g., Support Vector Machines, Neural Networks and Decision Trees.

5. Performance Metrics & Validation

In-Sample Vs. Out of Sample Errors

- ✓ **In sample error:** Error resulted from applying your prediction algorithm to the dataset you built it with
  - Also known as *resubstitution error*
  - Often optimistic (less than on a new sample) as the model may be tuned to error of the sample

- ✓ **Out of sample error:** Error resulted from applying your prediction algorithm to a new data set
  - Also known as *generalization error*
  - Out of sample error most important as it better evaluates how the model should perform

- ✓ **In sample error < out of sample error**
  - Will explain the reasons in a minute….

Regression Setting: Measuring Quality of Fit

- ✓Suppose we have a regression problem.
- ✓One common measure of accuracy is the mean squared error (MSE) i.e.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- ✓Where $\hat{y}_i$ is the prediction our method gives for the observation in our training data.

The Problem

- The training method has been designed to make MSE small on the training data we are looking at (e.g. with linear regression we choose the line such that MSE is minimized.)

- What we really care about is how well the method works on new data. We call this new data "**Test Data**".

- There is no guarantee that the method with the smallest training MSE will have the smallest test (i.e. new data) MSE.

Training vs. Test MSE's

- In general the more flexible a method is the lower its training MSE will be i.e. it will "fit" or explain the training data very well.

- However, the test MSE may in fact be higher for a more flexible method than for a simple approach like linear regression.

The Trade-off

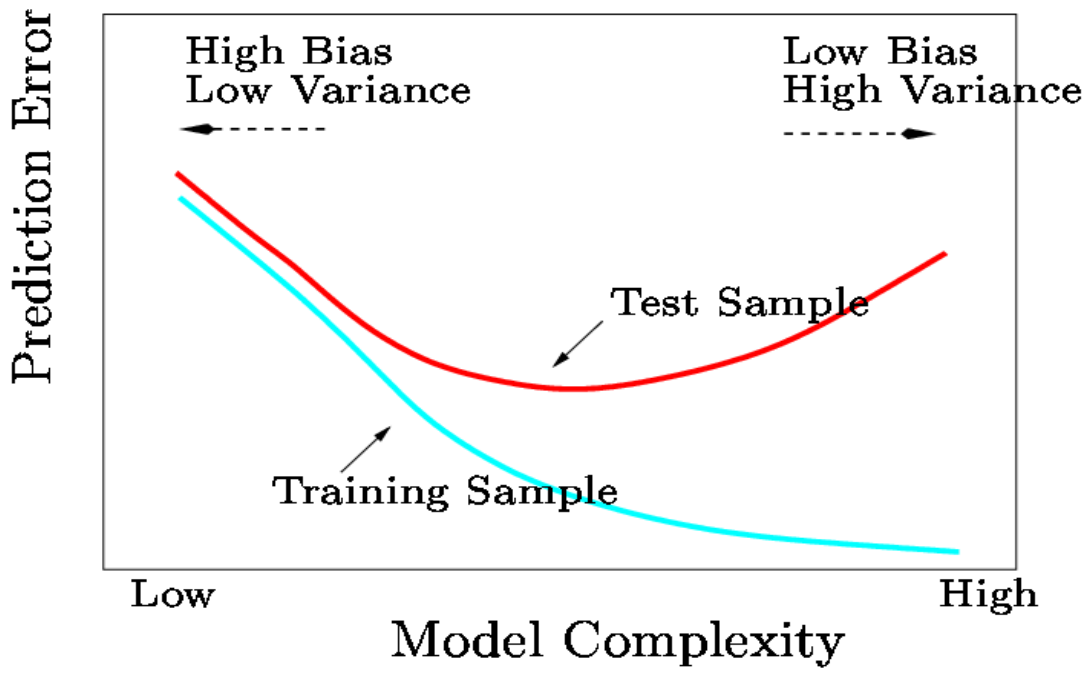
- ✓ It can be shown that for any given, $X=x_0$, the expected test MSE for a new Y at x_0 will be equal to

$$Expected\,Test\,MSE = E\left(Y - f(x_0)\right)^2 = Bias^2 + Var + \underbrace{s^2}_{\text{Irreducible Error}}$$

- ✓ As a model gets more complex, the bias will decrease and the variance will increase but expected test MSE may go up or down!

A Fundamental Picture

- ✓ Variance refers to how much your estimate for f would change by if you had a different training data set.
- ✓ Generally, the more flexible a method is the more variance it has.
- ➢ In general training errors will always decline if model complexity increases
- ➢ However, test errors will decline at first (as reductions in bias dominate) but will then start to increase again (as increases in variance dominate).



We must always keep this picture in mind when choosing a learning method.

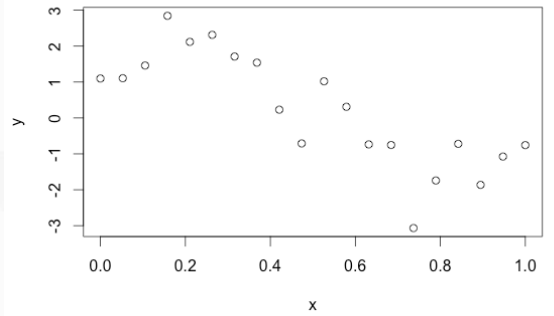
More flexible/complicated is not always better!

Bias/ Variance Tradeoff

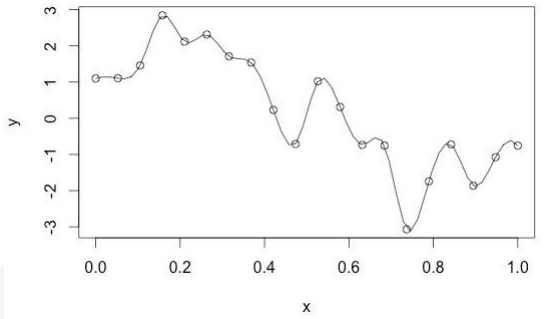
- The previous graphs of test versus training MSE's illustrates a very important tradeoff that governs the choice of statistical learning methods.

- There are always two competing forces that govern the choice of learning method i.e. bias and variance.

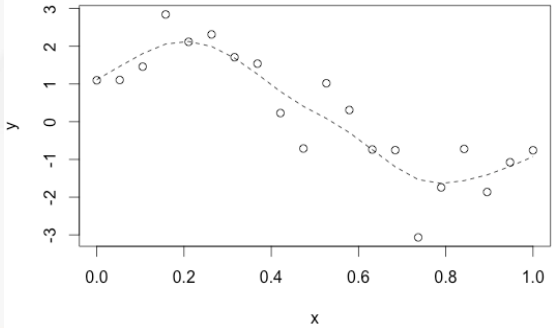
- High Variance – Over-fitting
- High Bias – Under-fitting



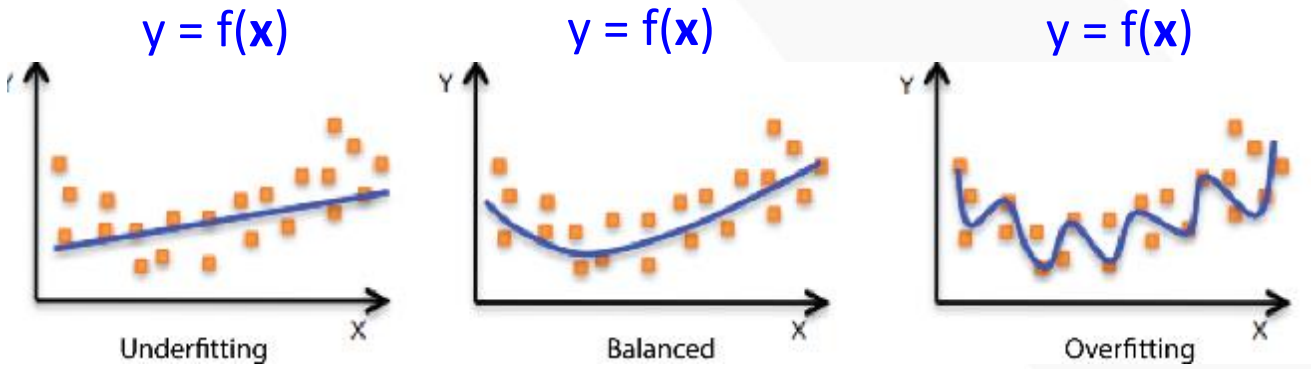
Low Bias, High Variance



High Bias, Low Variance



Over-fitting/ Under-fitting



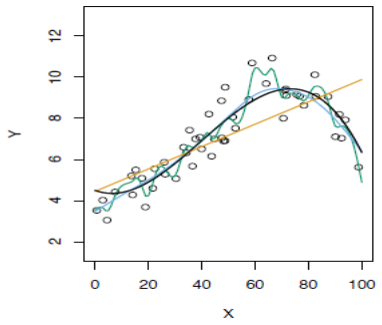
- Due to erroneous model specification(s), the model fits the noise.
- Frequently occurs when model is excessively complex
- Poor performance when deployed

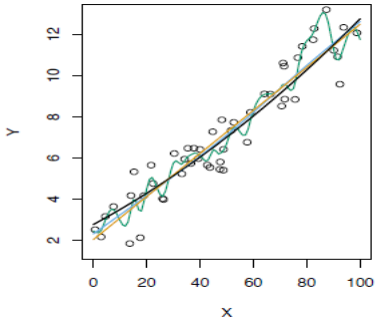
To avoid overfitting

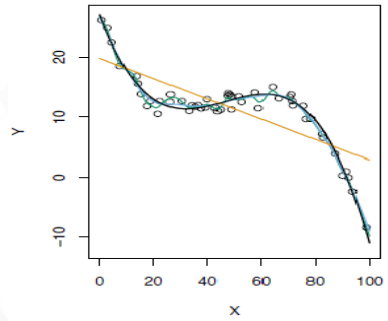
1) Regularization
2) Cross Validation

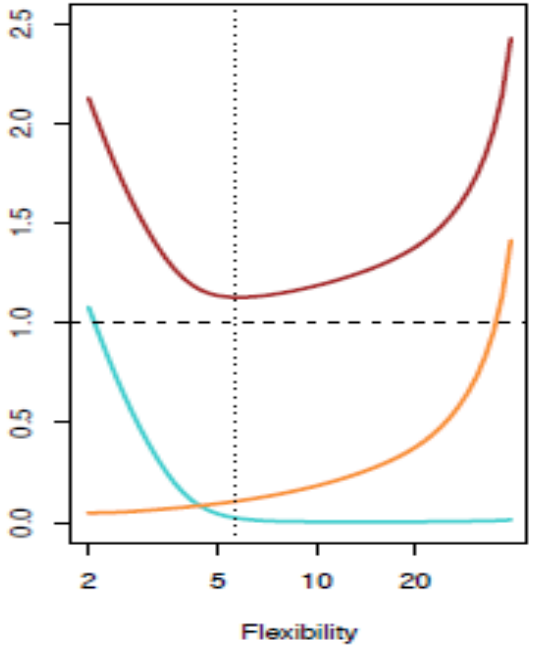
Image source: pingax.com

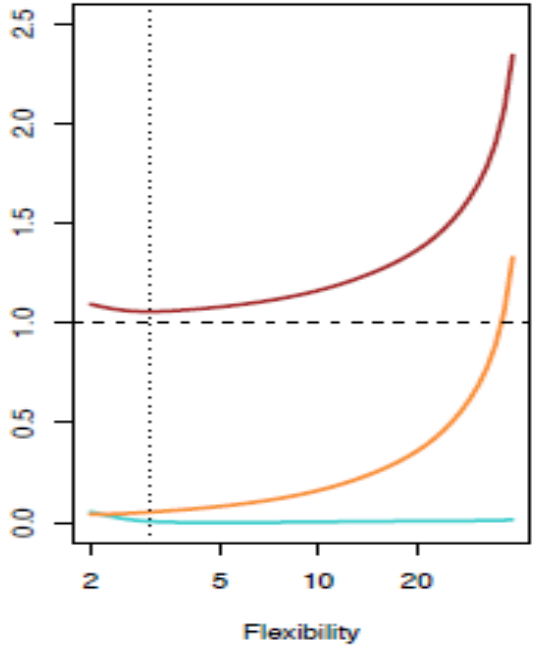
Test MSE, Bias and Variance

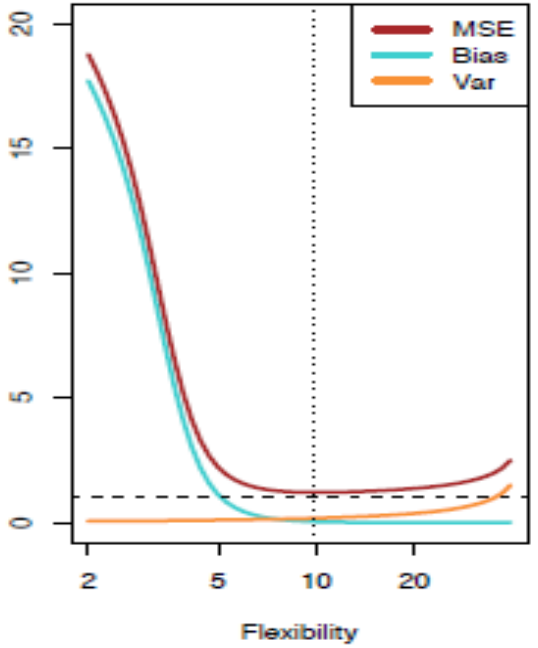












Bias – Variance Trade Off – Few Tips

- High number of features and less examples (observations)
- Reduce number of features (But that is information lost)
- Regularization
- If your predictions are seeing large error -
  - Get more training data
  - Try a smaller set a features
  - Try getting additional features
  - Adding polynomial features
  - Building your own, new, better features based on your knowledge of the problem
    - Can be risky if you accidentally over-fit your data by creating new features which are inherently specific/relevant to your training data

Regularization

Constrain the weights.
Impose penalty for complexity

Model=argmin ∑L (actual, predicted(Model))+λR (Model)

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

For Linear/Logistic Regression

✓If λ_2 is 0 , the regularization is called LASSO. Advantage: It does feature selection too.

✓If λ_1 is 0, the regularization is called LARS/ Ridge Regression

✓$\lambda_1 + \lambda_2$ is always 1. If both are present in the objective function, it is called elastic net regularization

Cross Validation

Procedures: Split training set into sub-training/test sets → Build model on sub-training set → Evaluate on sub-test set →
Repeat and average estimated errors

Result:

- we are able to fit/test various different models with different variables included to find the best one on the cross-validated test sets
- we are able to test out different types of prediction algorithms to use and pick the best performing one
- we are able to choose the parameters in prediction function and estimate their values
- ***Note**: original test set completely untouched, so when final prediction algorithm is applied, the result will be an unbiased measurement of the **out of sample accuracy** of the model*

Approaches:

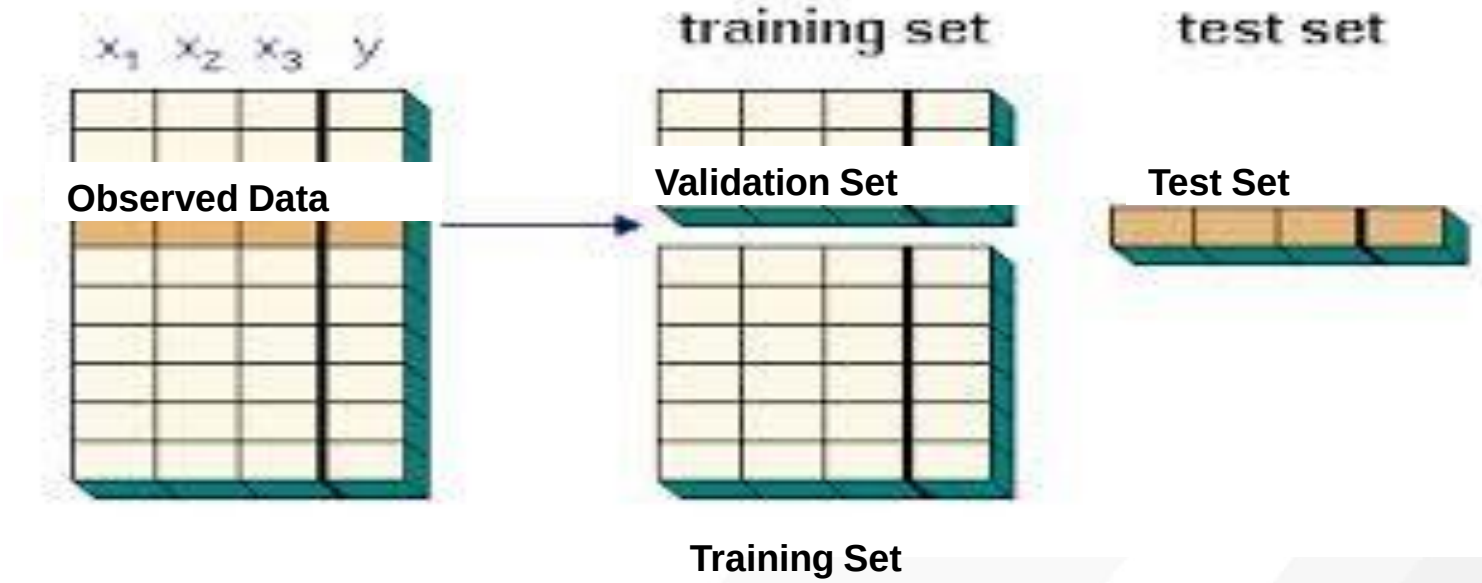
- Random subsampling/ Holdout Method
- K-fold
- Leave one out

Considerations:

- **For time series data must be used in "chunks"**
  - one time period might depending all time periods previously (should not take random samples)

Cross validation Approach-1: Train, Test, Validate Datasets

Validation set used as a proxy to estimate out of sample error



Sample Design Guidelines for Prediction Study

- ✓ For large sample sizes: 60% training - 20% test - 20% validation
- ✓ For medium sample sizes: 60% training - 40% test – no validation set to refine model (to ensure test set is of sufficient size)
- ✓ For small sample sizes:
  - Carefully consider if there are enough sample to build a prediction algorithm
  - Report caveat of small sample size and highlight the fact that the prediction algorithm has never been tested for out of sample error
- ✓ There should always be a test/validation set that is held away and should ***NOT*** be looked at when building model
  - When complete, apply the model to the held-out set only one time
- ✓ ***Randomly sample*** training and test sets
  - For data collected over time, build training set in chunks of times
- ✓ Datasets must reflect structure of problem
  - If prediction evolves with time, split train/test sets in time chunks (known as *back testing* in finance)
- ✓ **Subsets of data should reflect as much diversity as possible**

Cross-Validation Approach-2: K-fold validation

- ✓ Break training set into *K* subsets
- ✓ Build the model/predictor on the remaining training data in each subset and applied to the test subset
- ✓ Rebuild the data *K* times with the training and test subsets
- ✓ Average the findings

- ✓ ***Considerations:***

  larger k = less bias, more variance

  smaller k = more bias, less variance

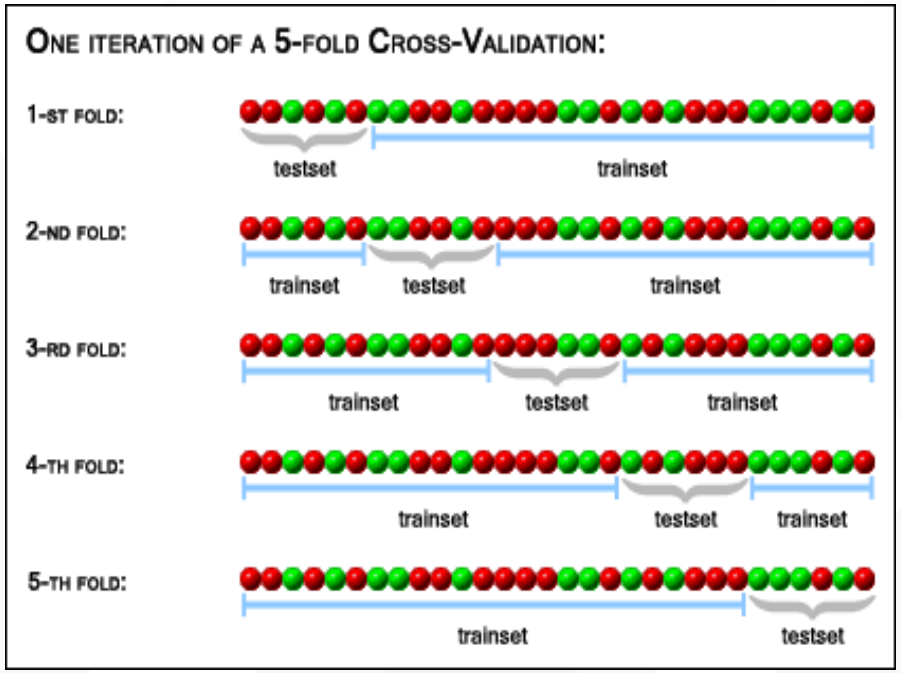


Image Source: www.kaggle.com/c/chess

source: wikipedia

Cross-Validation Approach-3: Leave one out (LOO)

- ✓ leave out exactly one sample and build predictor on the rest of training data
- ✓ predict value for the left out sample
- ✓ repeat for each sample

source: wikipedia

Let’s Check Our Understanding…

Quiz-1

Which of the following is best suited for machine learning?

A. Predicting whether the next cry of the baby girl happens at an even-numbered minute or not

B. Determining whether a given graph contains a cycle

C. Deciding whether to approve credit card to some customer

D. Guessing whether the earth will be destroyed by the misuse of nuclear power in the next ten years

A. no pattern

B. programmable definition

C. pattern: customer behavior; definition: not easily programmable; data: history of bank operation

D. arguably no (or not enough) data yet

Quiz-1

Which of the following is best suited for machine learning?

A. Predicting whether the next cry of the baby girl happens at an even-numbered minute or not

B. Determining whether a given graph contains a cycle

C. Deciding whether to approve credit card to some customer

D. Guessing whether the earth will be destroyed by the misuse of nuclear power in the next ten years

Answer: C

A. no pattern

B. programmable definition

C. pattern: customer behavior; definition: not easily programmable; data: history of bank operation

D. arguably no (or not enough) data yet

Quiz-2

Which of the following claim is not totally true?

A. machine learning is a route to realize artificial intelligence
B. machine learning, data mining and statistics all need data
C. data mining is just another name for machine learning
D. statistics can be used for data mining

Note: While data mining and machine learning do share a huge overlap, they are arguably not equivalent because of the difference of focus.

Quiz-2

Which of the following claim is not totally true?

A. machine learning is a route to realize artificial intelligence
B. machine learning, data mining and statistics all need data
C. data mining is just another name for machine learning
D. statistics can be used for data mining

Answer: C

Note: While data mining and machine learning do share a huge overlap, they are arguably not equivalent because of the difference of focus.

Quiz-3

Of the following examples, which would you address using an unsupervised learning algorithm?

A) Given email labeled as spam/not spam, learn a spam filter.
B) Given a set of news articles found on the web, group them into set of articles about the same story.
C) Given a database of customer data, automatically discover market segments and group customers into different market segments.
D) Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as havi diabetes or not.

Quiz-3

Of the following examples, which would you address using an unsupervised learning algorithm?

A) Given email labeled as spam/not spam, learn a spam filter.
B) Given a set of news articles found on the web, group them into set of articles about the same story.
C) Given a database of customer data, automatically discover market segments and group customers into different market segments.
D) Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as havi diabetes or not.

The Answer: B & C

Quiz-4

You're running a company, and you want to develop learning algorithms to address each of two problems.

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

Should you treat these as classification or as regression problems?

A) Treat both as classification problems.

B) Treat problem 1 as a classification problem, problem 2 as a regression problem.

C) Treat problem 1 as a regression problem, problem 2 as a classification problem.

D) Treat both as regression problems.

Quiz-4

You're running a company, and you want to develop learning algorithms to address each of two problems.

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

Should you treat these as classification or as regression problems?

A) Treat both as classification problems.

B) Treat problem 1 as a classification problem, problem 2 as a regression problem.

C) Treat problem 1 as a regression problem, problem 2 as a classification problem.

D) Treat both as regression problems.

The Answer: C

Quiz-5

The entrance system of the school gym, which does automatic face recognition based on machine Learning, is built to charge four different groups of users differently: staff, student, professor, other. What type of learning problem best fits the need of the system

A. Binary Classification
B. Multi Class classification
C. Regression
D. None of the above

Quiz-5

The entrance system of the school gym, which does automatic face recognition based on machine Learning, is built to charge four different groups of users differently: staff, student, professor, other. What type of learning problem best fits the need of the system

A. Binary Classification
B. Multi Class classification
C. Regression
D. None of the above

The Answer: B

Puzzle

A huntsman can hit a target with probability of 0.2.

He sees a flock of birds (*150 birds*) atop a banyan tree. He takes aim and fires three continuous shots. He hits exactly one of the birds.

Question: How many birds remain on the tree?

And then, there were none

Domain knowledge is very important

Don’t lose the big picture

What we have 'learnt' so far…

- ✓ **What is Machine Learning & Use Cases**
- ✓ **Major Classes of Learning Algorithms:**
  - ✓ Supervised, Unsupervised, Semi- Supervised, Reinforced
- ✓ **Building a Machine Learning Model**
  - ✓ Data Pre-processing, Training & Test Split, Model Building, Validation, Prediction
- ✓ **Performance Metrics & Evaluation – Concept of:**
  - ✓ Over/Under-fitting;
  - ✓ Bias, Variance, and Trade off
  - ✓ Regularization
  - ✓ Cross Validation

Next Session: Supervised ML | Linear Regression

Contact Us

Visit us on: http://www.analytixlabs.in/

For more information, please contact us: http://www.analytixlabs.co.in/contact-us/

Or email: info@analytixlabs.co.in

Call us we would love to speak with you: (+91) 88021-73069

Join us on:

Twitter - http://twitter.com/#!/AnalytixLabs

Facebook - http://www.facebook.com/analytixlabs

LinkedIn - http://www.linkedin.com/in/analytixlabs

Blog - http://www.analytixlabs.co.in/category/blog/