DATA ANALYTICS & MACHINE LEARNING

PREMANANDA INDIC, PH.D.

DEPARTMENT OF ELECTRICAL ENGINEERING

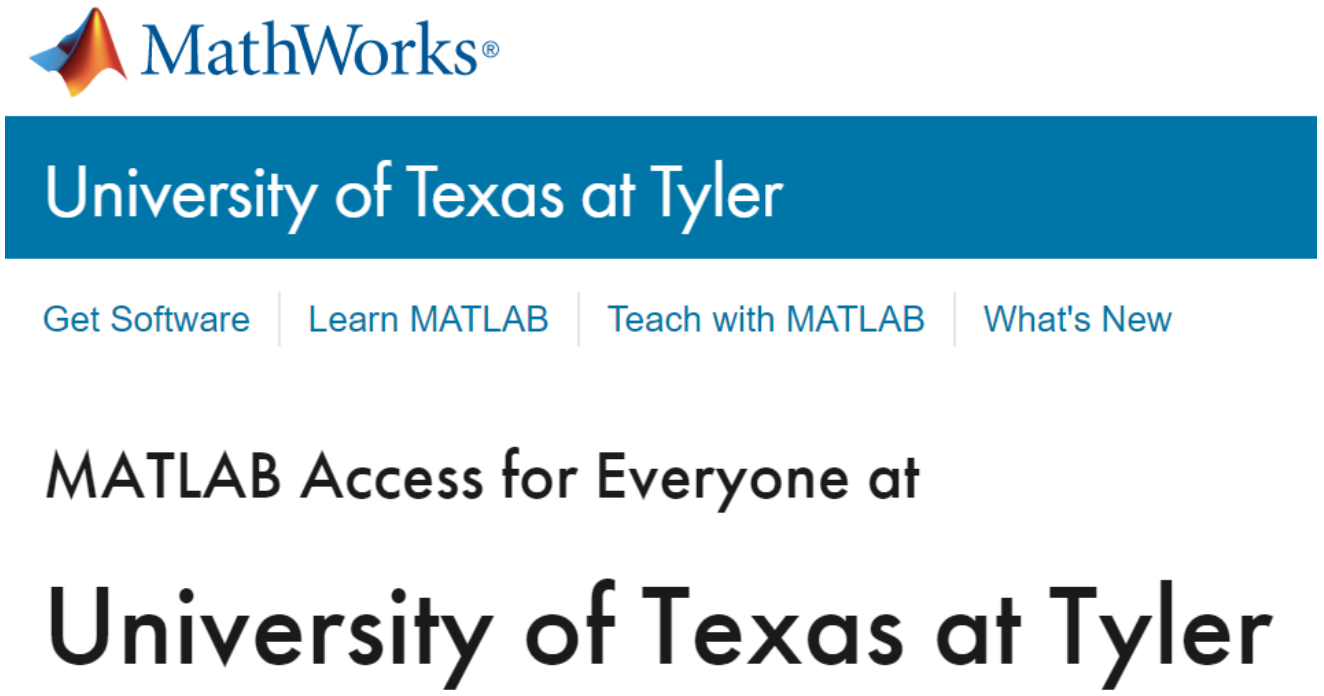
The University of Texas at TYLER
Center for Health Informatics & Analytics

ORS Research Design & Data Analysis Lab
Office of Research and Scholarship

WORKSHOP SCHEDULE

- WEEK1: DATA ANALYTICS
- WEEK2: FEATURE EXTRACTION
- WEEK3: MACHINE LEARNING

ANALYSIS PLATFORM



https://www.mathworks.com/academia/tah-portal/university-of-texas-at-tyler-1108545.html

ANALYSIS PLATFORM

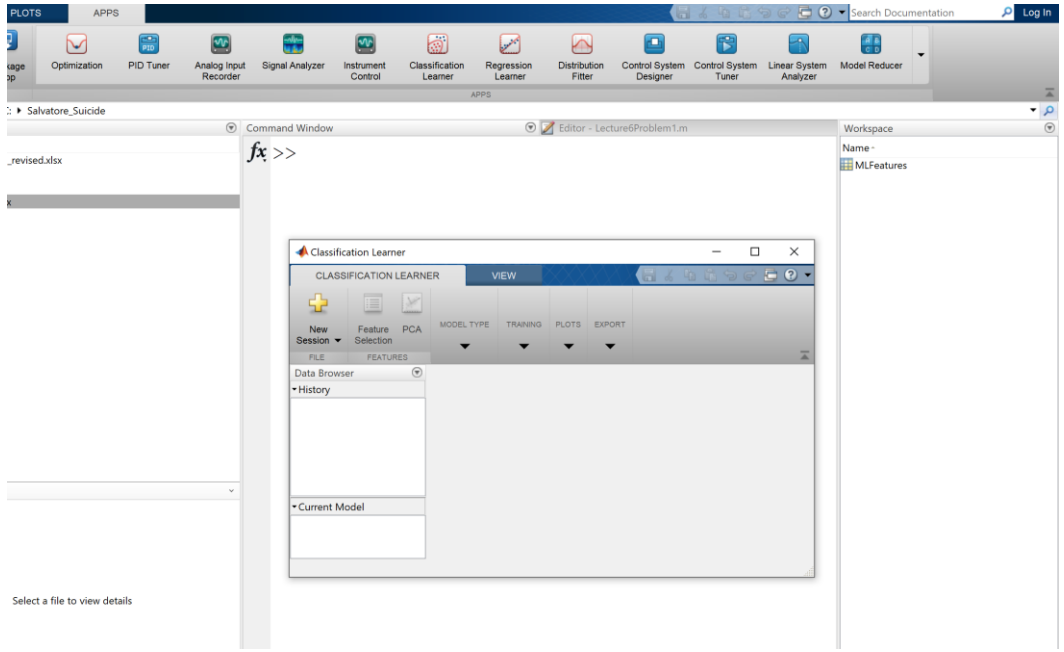
MathWorks®

University of Texas at Tyler

Get Software | Learn MATLAB | Teach with MATLAB | What's New

MATLAB Access for Everyone at

University of Texas at Tyler



https://www.mathworks.com/academia/tah-portal/university-of-texas-at-tyler-1108545.html

OUTLINE

- INTRODUCTION
- DIFFERENT MACHINE LEARNING APPROACHES
- PROJECTS

OUTLINE

- INTRODUCTION
- DIFFERENT MACHINE LEARNING APPROACHES
- PROJECTS

INTRODUCTION

➢What is Machine Learning ?

- Machine Learning is a field of study that gives computers the ability to “learn” without being explicitly programmed
  - Prediction
  - Classification

Samuel AL, IBM J. Research & Development, 1959, vol. 3 (3), 210-229

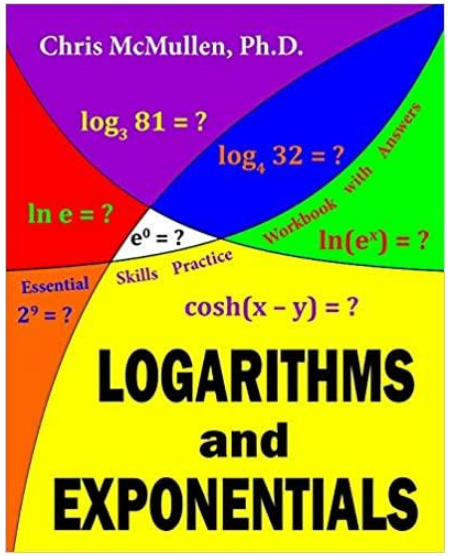
INTRODUCTION

- Too many books spoil the curiosity

- Start with Andrew Ng, Machine Learning, Stanford University available on YouTube

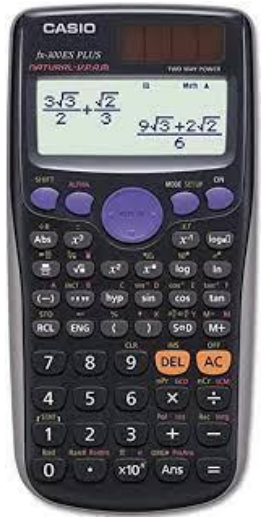
Some Statistics & Programming Knowledge Helps !

INTRODUCTION



Analytical Tools

Simple Calculator (Boolean Algebra)



Scientific Calculator (Series Expansion, Boolean Algebra)

Computer (Programming Language, Assembly Language, Series Expansion, Boolean Algebra)

Smart Devices (ML Models, Programming Language, Assembly Language, Series Expansion, Boolean Algebra)

INTRODUCTION

- Always there is a mathematical foundation

  Analytical Tools (Logarithm, Laplace Transform, Fourier Transform…….)

  Computational Tools (Boolean Algebra, Taylor Series Expansion,……..)

  Programming Languages (Basic, Fortran, C, C++, Java, …….)

  Assembly Languages (depending upon the computer processors)

  Machine Learning Models

  Artificial Intelligence

INTRODUCTION

- Examples of Smart Systems
  - Voice Recognition
  - Tumor Detection
  - Weather Forecast
  - Driverless Cars

WHAT IS NEEDED?

- Training Data
- Appropriate Model
- Procedure to Train (Make a machine to “learn”)

(Learning Algorithms, Online vs Batch Learning, Instance Based vs Model Based

- Test Data

OUTLINE

- INTRODUCTION
- DIFFERENT MACHINE LEARNING APPROACHES
- PROJECTS

APPROACHES

- SUPERVISED LEARNING
- UNSUPERVISED LEARNING

APPROACHES

- SUPERVISED LEARNING
- UNSUPERVISED LEARNING

APPROACHES

- SUPERVISED LEARNING (Classification / Prediction)

  Provide training set with features and solutions

APPROACHES

- SUPERVISED LEARNING (Classification / Prediction)

  Find the area of a rectangle

L	W	A	A1 (L+W)	A2 (L-W)	A3 (L*W)	A4 L/W
12.1	13.4	162.3	25.5	-1.3	162.14	0.90
8.6	9.7	83.4	18.3	-1.1	83.42	0.89
3.2	5.4	17.3	8.6	-2.2	17.28	0.59
6.1	10.2	62.25	16.3	-4.1	62.22	0.60
18.2	6.4	116.5	24.6	11.8	116.48	2.83
1.6	2.8	4.5	4.4	-1.2	4.48	0.57
7.7	0.6	4.7	8.3	7.1	4.62	12.83

APPROACHES

- SUPERVISED LEARNING (Classification / Prediction)

  Find the area of a rectangle

L	W	A	E1 |A-A1|	E2 |A-A2|	E3 |A-A3|	E4 |A-A4|
12.1	13.4	162.3	136.8	163.6	0.16	161.40
8.6	9.7	83.4	65.1	84.5	0.02	82.51
3.2	5.4	17.3	8.7	19.5	0.02	16.71
6.1	10.2	62.25	45.95	66.35	0.03	61.65
18.2	6.4	116.5	91.90	104.70	0.02	113.66
1.6	2.8	4.5	0.1	5.7	0.02	3.93
7.7	0.6	4.7	3.6	2.4	0.08	8.13

APPROACHES

- SUPERVISED LEARNING (Classification / Prediction)
  - Linear Regression
  - Logistic Regression
  - Support Vector Machines
  - k-Nearest Neighbors
  - Decision Trees and Random Forests
  - Neural Networks

APPROACHES

- SUPERVISED LEARNING (Classification / Prediction)
  - Linear Regression

Given m outcomes y^i where $i = 1,2, \ldots., m$ with each outcome depends on n features x_j where $j = 1,2, \ldots., n$. Find the best estimate of y^i as $\hat{y}^i$ using the n features with appropriate parameters θ_j such that $J = \left\langle \left(\hat{y}^i - y^i\right)^2 \right\rangle$

$$\hat{y}^i = \theta_0^i + \theta_1^i x_1^i + \theta_2^i x_2^i + \cdots \ldots \ldots + \theta_n^i x_n^i$$

APPROACHES

- SUPERVISED LEARNING (Classification / Prediction)
  - Linear Regression

$$\hat{y}^i = \theta_0^i + \theta_1^i x_1^i + \theta_2^i x_2^i + \cdots \ldots \ldots + \theta_n^i x_n^i$$

$$\hat{Y} = \Theta . X = h_\theta(X)$$

Cost Function to Minimize

$$J = \left\langle \left(\hat{y}^i - y^i\right)^2 \right\rangle = (\hat{Y} - Y)^T (\hat{Y} - Y)$$

APPROACHES

- SUPERVISED LEARNING (Classification / Prediction)
  - Linear Regression

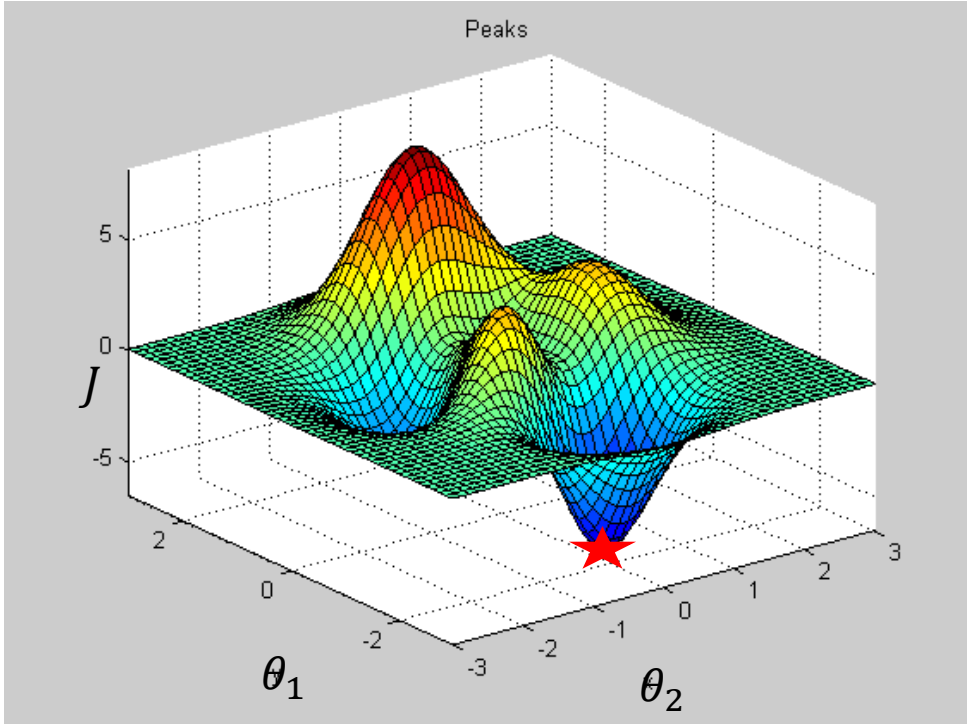
$$\hat{y}^i = \theta_0^i + \theta_1^i x_1^i + \theta_2^i x_2^i + \cdots \ldots \ldots + \theta_n^i x_n^i$$

$$\hat{Y} = \Theta . X = h_\theta(X)$$

- Gradient Descent by **Louis Augustin Cauchy** in 1847

Cost Function to Minimize

$$J = \left\langle \left(\hat{y}^i - y^i\right)^2 \right\rangle = (\hat{Y} - Y)^T (\hat{Y} - Y)$$



APPROACHES

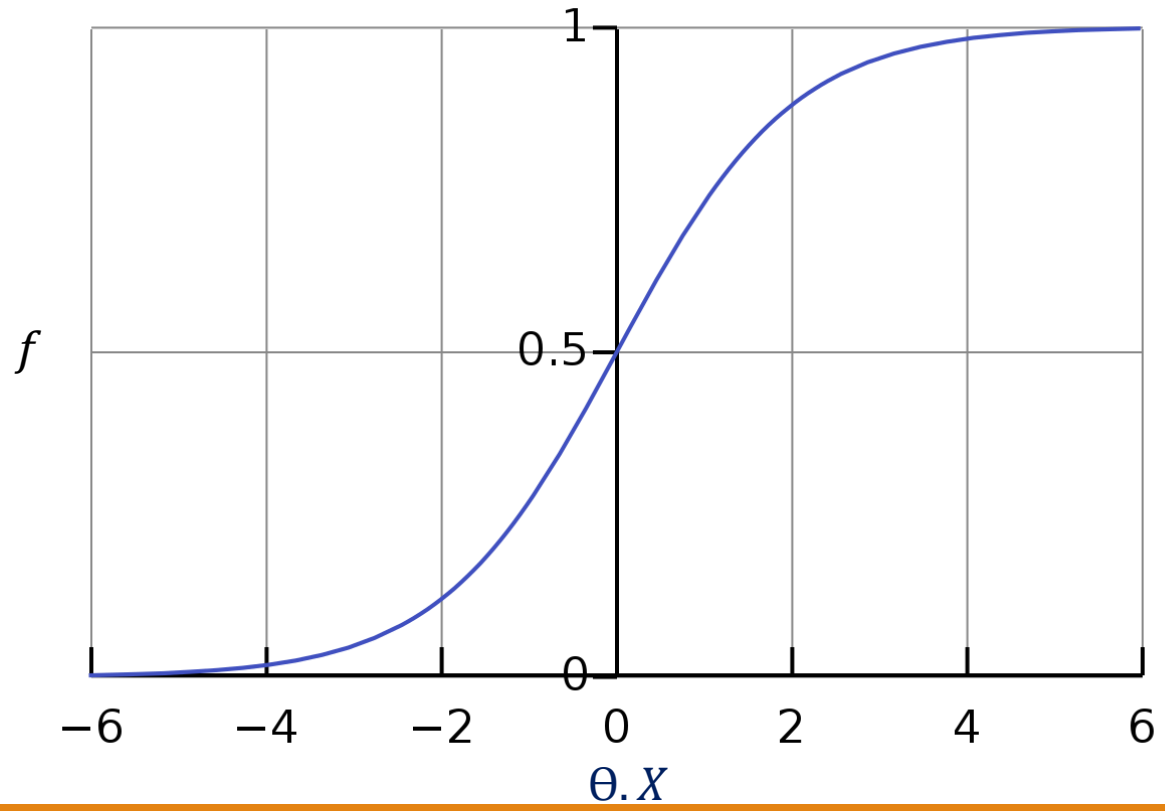
- SUPERVISED LEARNING (Classification / Prediction)
  - Logistic Regression

$$\hat{p} = f(\Theta.X) = h_\theta(X)$$

$$\hat{y} = 1 \; if \; \hat{p} < 0.5; \;\; \hat{y} = 0 \; if \; \hat{p} \geq 0.5$$

Derive Cost Function to Minimize

J



APPROACHES

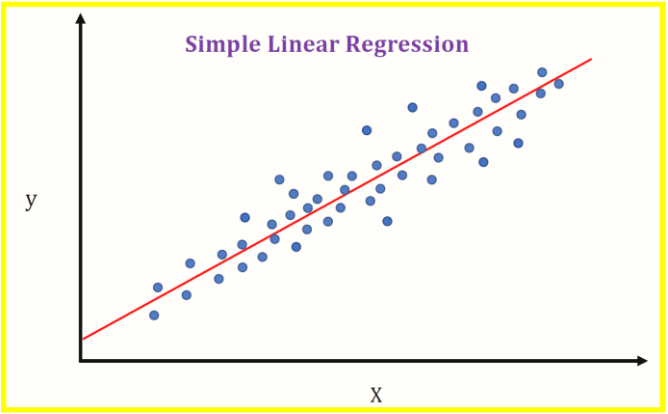
- SUPERVISED LEARNING (Classification / Prediction)
  - Linear Regression

Mainly for regression (predicting an outcome)

- Logistic Regression

Mainly for classification (0 or 1)

High Risk vs. Low Risk



https://medium.datadriveninvestor.com/machine-learning-101-part-1-24835333d38a

APPROACHES

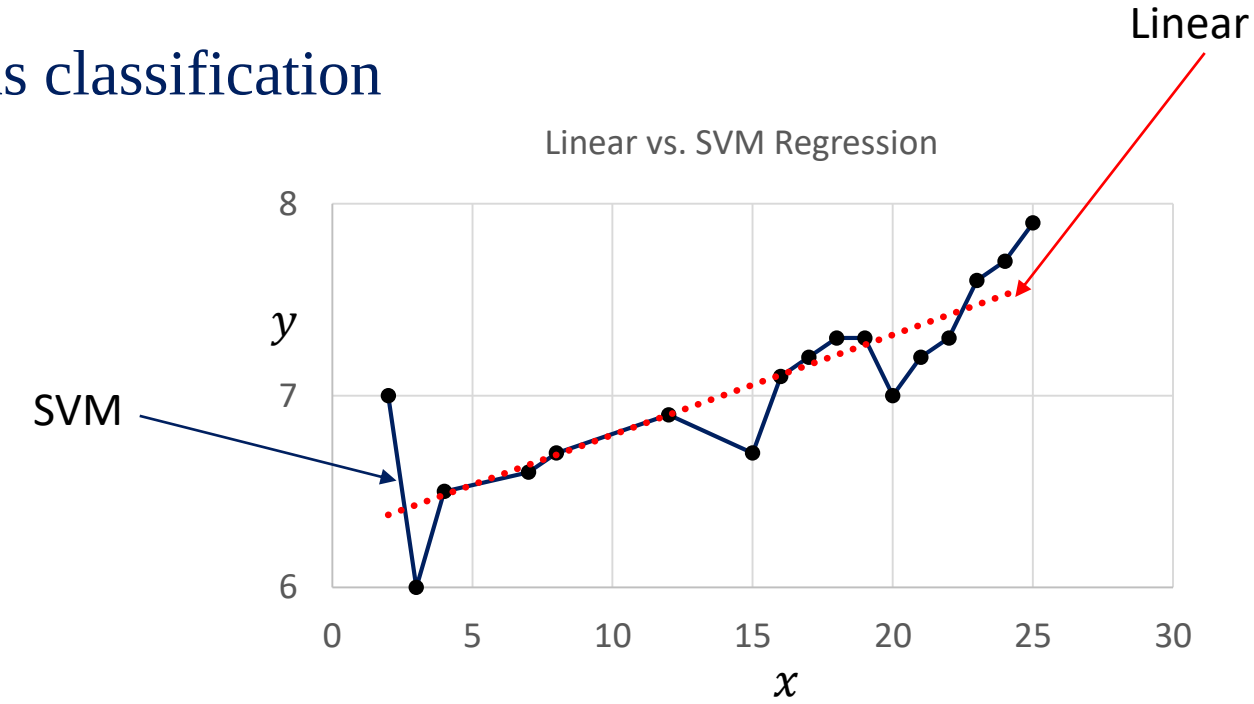
- SUPERVISED LEARNING (Classification / Prediction)
  - Support Vector Machine

Used for regression as well as classification

APPROACHES

- SUPERVISED LEARNING (Classification / Prediction)
  - Support Vector Machine (SVM)

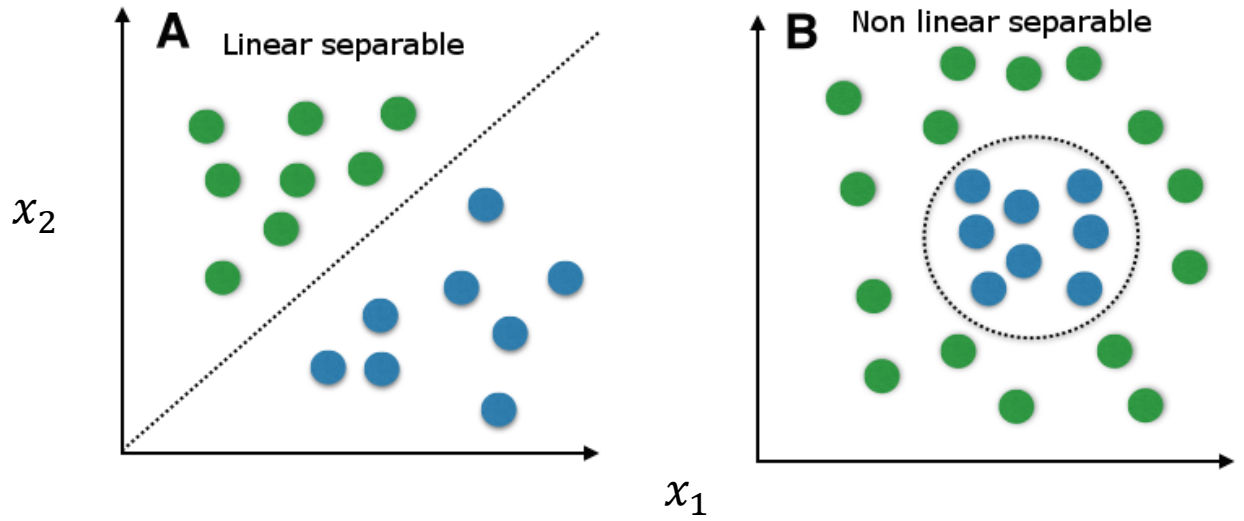
Used for regression as well as classification



APPROACHES

- SUPERVISED LEARNING (Classification / Prediction)
  - Support Vector Machine (SVM)

Used for regression as well as classification

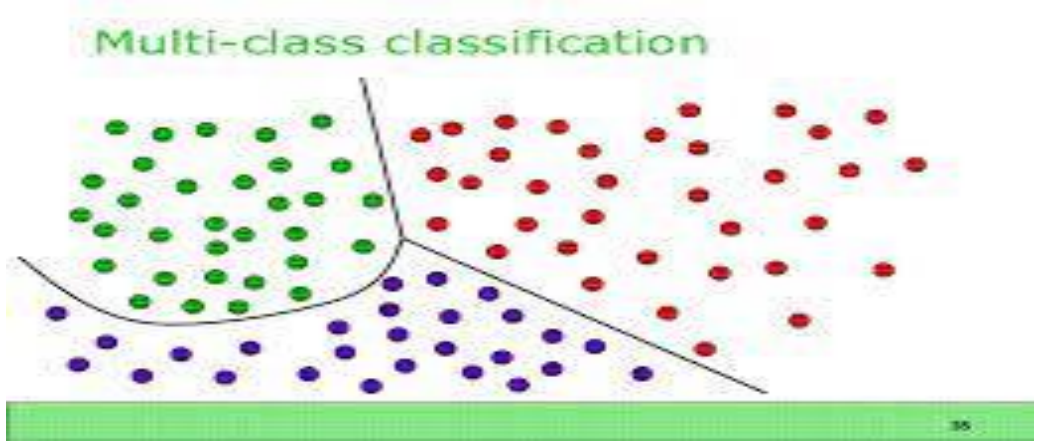


https://medium.com/@LSchultebraucks/introduction-to support-vector-machines-9f8161ae2fcb

APPROACHES

- SUPERVISED LEARNING (Classification / Prediction)
  - Support Vector Machine (SVM)

Used for regression as well as classification



https://www.mathworks.com/matlabcentral/fileexchange/62061-multi-class-svm

APPROACHES

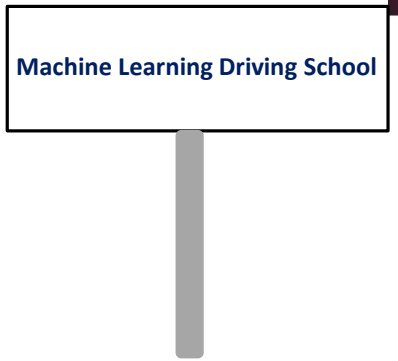
- SUPERVISED LEARNING (Classification / Prediction)
  - Linear Regression
  - Logistic Regression
  - Support Vector Machines
  - k-Nearest Neighbors
  - Decision Trees and Random Forests
  - Neural Networks

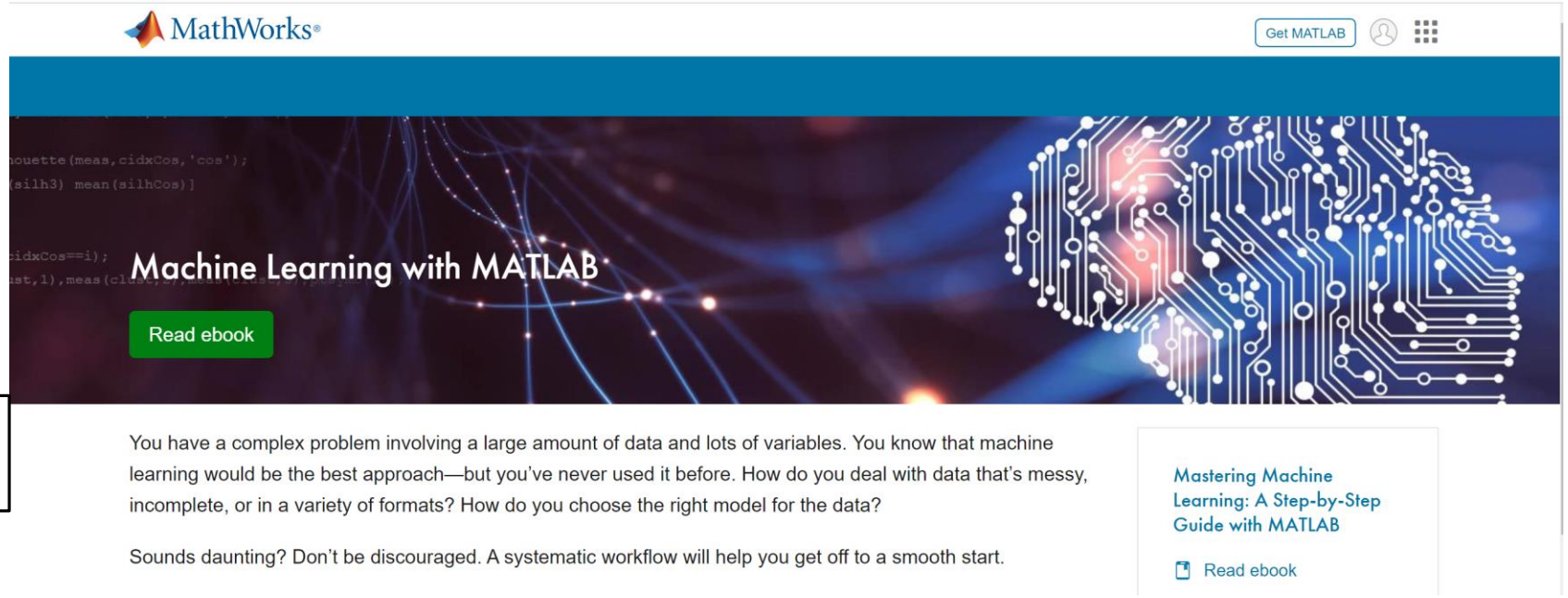
APPROACHES

- Machine Learning with MATLAB

https://commons.wikimedia.org/wiki/File:Man_Driving_Car_Cartoon_Vector.svg

http://clipart-library.com/mechanic-cliparts.html





Project 1

- Prediction of House Price (housing.csv) Regression Problem

longitude
latitude
housing_median_age
total_rooms
total_bedrooms
population
households median_income
median_house_value
ocean_proximity

Project 2

- Prediction of House Price (housing.csv) Classification Problem

longitude
latitude
housing_median_age
total_rooms
total_bedrooms
population
households median_income
median_house_value (High/Low) Threshold= 257500
ocean_proximity

Project 2

- Prediction of House Price (housing.csv) Classification Problem

Confusion Matrix

True Class	1	0
1	True Positive	False Negative → Total Positive
0	False Positive	True Negative → Total Negative
	1	0

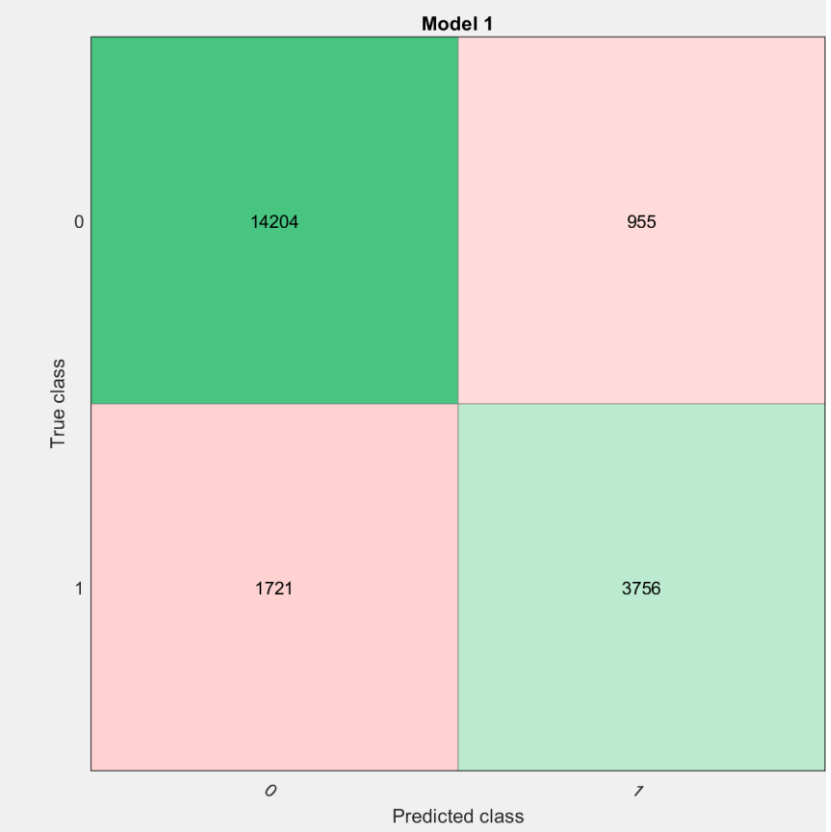
Predicted Class

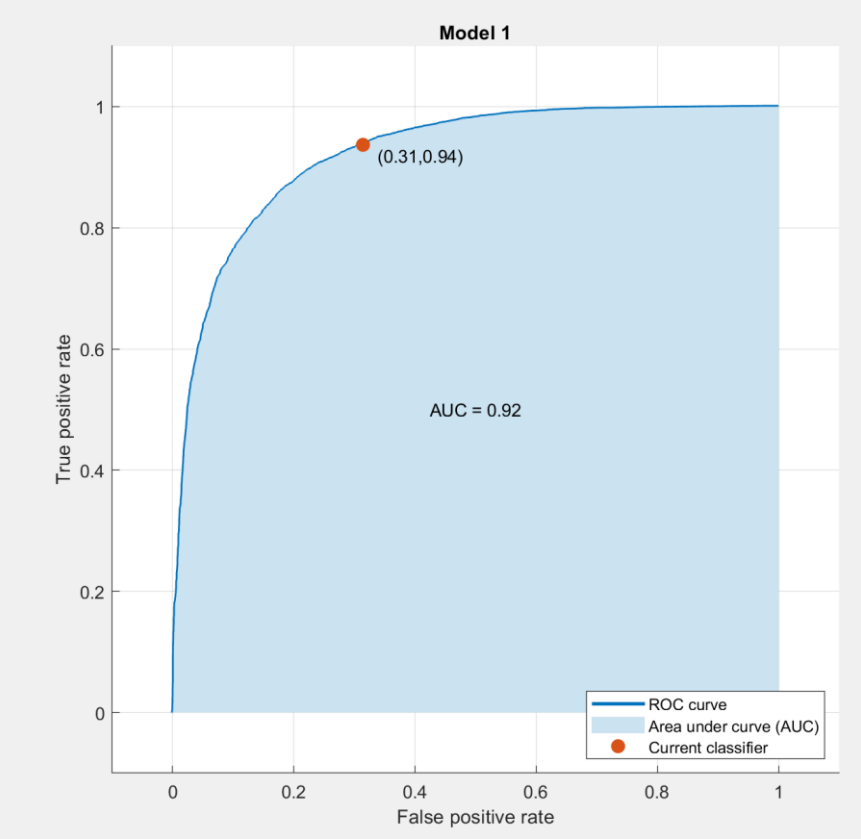
True Positive Rate = True Positive / Total Positive

True Negative Rate = True Negative / Total Negative = 1 – False Positive Rate

Project 2

- Prediction of House Price (housing.csv) Classification Problem





Project 3

- To test the hypothesis that the accelerometer data can detect stress

Project 4

- To test the hypothesis that the features of SpO2 can detect smoker from non-smoker

Project 5

- To test the hypothesis that the features of saliva can detect COPD from other conditions

- Data set : http://archive.ics.uci.edu/ml/datasets/Exasens

APPROACHES

- SUPERVISED LEARNING
- UNSUPERVISED LEARNING

APPROACHES

- UNSUPERVISED LEARNING

Clustering

- k-means
- Principal Component Analysis
- Independent Component Analysis
- Singular Value Decomposition
- …….
- ……

Project 6

- Demonstration of Clustering

Still More Methods…

- Deep Learning Methods

THANK YOU

Sloke Shrestha, UG

Mohammed Alenazi, Graduate

Pravitha Ramanand, PhD, Postdoc

ORS Research Design & Data Analysis Lab

Office of Research and Scholarship