

FOUNDATIONS OF MACHINE LEARNING

ANUSH SANKARAN

Most of the slide content are borrowed from multiple source online. Thanks to the original authors

OVERVIEW OF THE COURSE

Week

Topics

Week 1

Intro to ML
Discovering ML Use Cases & ML in Business

Week 2

Python- Hands On
Supervised Learning & Regression

Week 3

Neural Network - 1
Neural Network -2 (Bias, Variance) & Hands ON

Week 4

Kernel Learning & SVM
Practical Advice for ML projects.

Week 5

Boosting
Decision Trees, Random Forest, & xgBoost

Week 6

Unsupervised Learning
Clustering & Dimensionality Reduction

Week 7

Time Series Data Analysis
Imputation & Prediction Systems

Week 8

ML Use Cases from Products & Research

COURSE OUTCOMES

- ▶ Understand the fundamental concepts of different machine learning models
 - ▶ Supervised learning
 - ▶ Unsupervised learning
- ▶ Ability to formulate a business problem as machine learning task. Identify machine learning opportunities in businesses.
- ▶ Appreciate the challenging involved in data driven machine learning problems
- ▶ Ability to manage the building of tools and products that involves different aspects of machine learning

EASY LOGISTICS: GITHUB

- ▶ Github Repo: <https://github.com/goodboyanush/isme-bangalore-Oct-Nov-2019>
- ▶ Lectures slides, Hands-on code, Assignment solutions
- ▶ Have any doubt in my lectures or assignments?
 - ▶ Go ahead and create an issue in the repo!
 - ▶ I will try to answer them asap!
 - ▶ Everyone will be benefitted by the questions asked by one

WEEK 1:

INTRODUCTION TO MACHINE LEARNING

BUILDING THE ML MINDSET IN BUSINESS

WHAT IS MACHINE LEARNING?

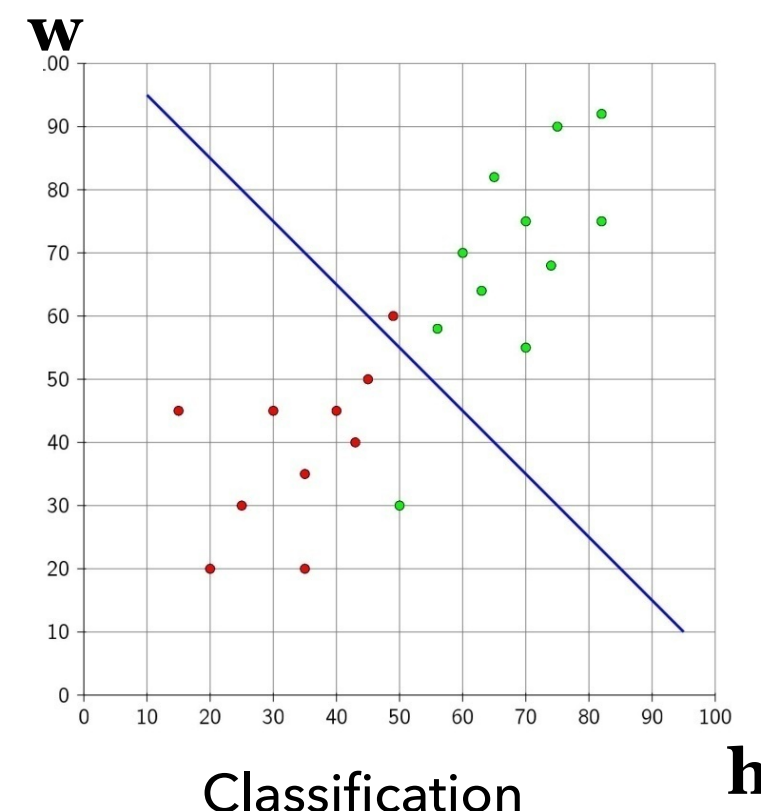
- Learn a classifier: learn a mapping function



Instances/ Input Data Points

$\left\{ \begin{array}{l} 1. \text{ M: } \langle h_1, w_1 \rangle \\ 2. \text{ F: } \langle h_2, w_2 \rangle \\ 3. \dots \\ \dots \\ N. \text{ M: } \langle h_n, w_n \rangle \end{array} \right\}$

Labelled Features



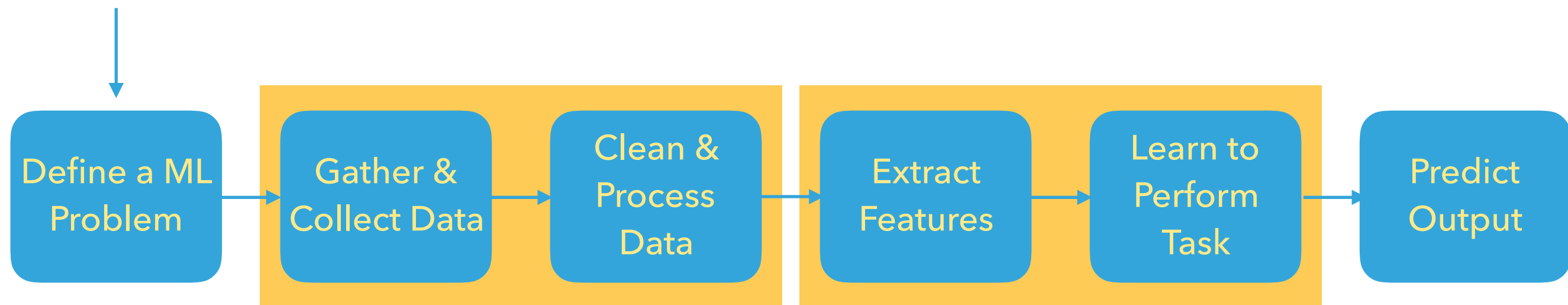
1. **Boundary**: Can be linear or non-linear boundary

2. **Method**: Can be a generative classifier or discriminative classifier

Examples: Naïve Bayes, Decision trees, Neural Network, Support Vector Machines etc.

MACHINE LEARNING PIPELINE

What are we focussing on today's lecture?



1. Articulate the problem (task)
2. Data Drive Strategy: Look for labelled data
3. Design your data for the task
4. Determine easily obtained inputs
5. Determine easily quantifiable outputs

COMMON LINGO

Input data/ Features

Output
Task Label

Instances

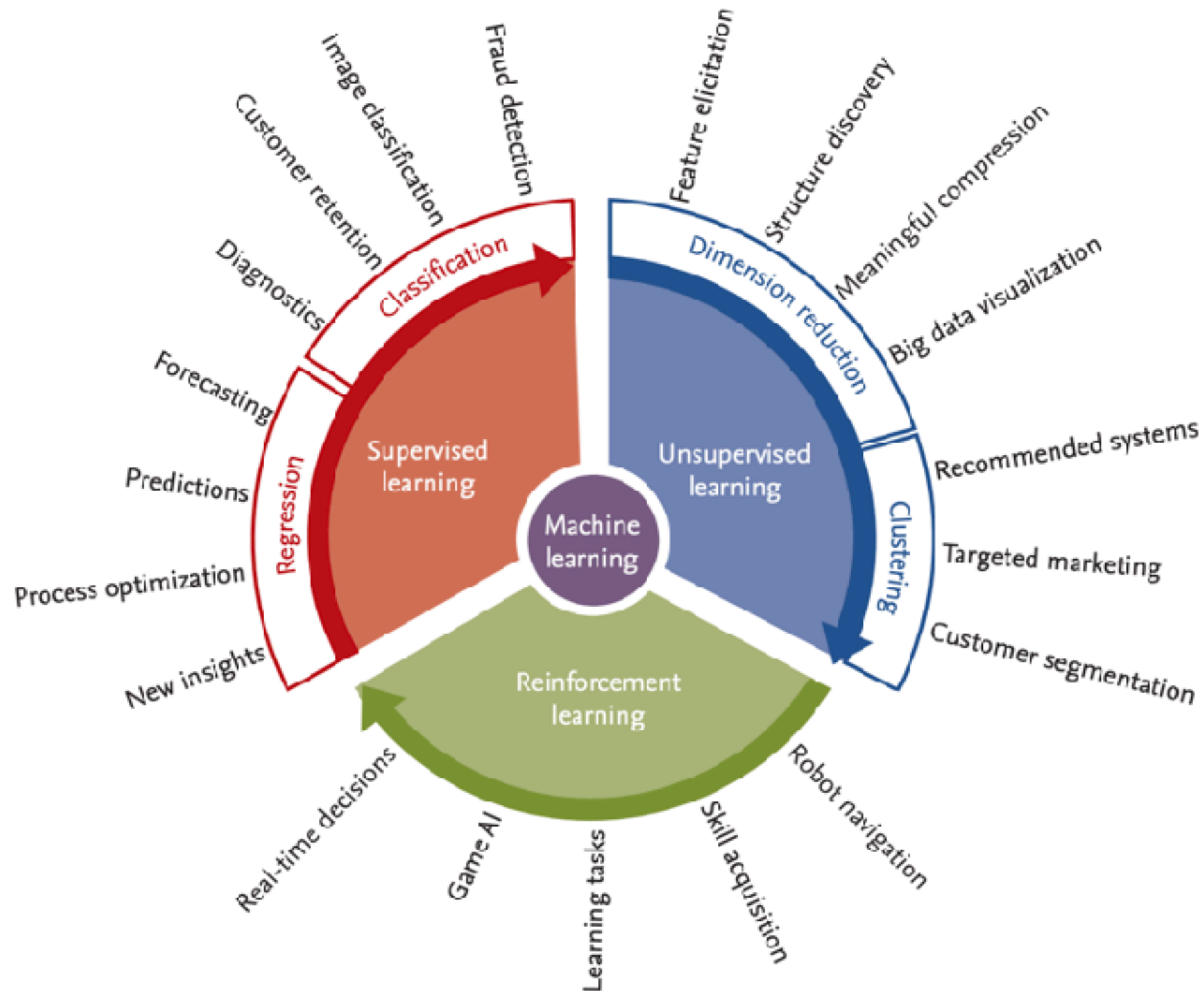
Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001003	Male	Yes		1 Graduate	No	4533	1508	128	300	1	Rural	N
LP001005	Male	Yes		0 Graduate	Yes	3000	0	66	350	1	Urban	Y
LP001006	Male	Yes		0 Not Graduate	No	2581	2758	120	350	1	Urban	Y
LP001008	Male	No		0 Graduate	No	6000	0	141	350	1	Urban	Y
LP001011	Male	Yes		2 Graduate	Yes	5417	4196	257	350	1	Urban	Y
LP001013	Male	Yes		0 Not Graduate	No	2333	1516	95	300	1	Urban	Y
LP001014	Male	Yes	31	Graduate	No	3036	2504	158	350	0	Semiurban	N
LP001016	Male	Yes		2 Graduate	No	4006	1526	168	350	1	Urban	Y
LP001020	Male	Yes		1 Graduate	No	12841	10968	349	350	1	Semiurban	N
LP001024	Male	Yes		2 Graduate	No	3200	700	70	350	1	Urban	Y
LP001028	Male	Yes		2 Graduate	No	3073	3106	200	300	1	Urban	Y
LP001029	Male	No		0 Graduate	No	1853	2840	114	350	1	Rural	N
LP001030	Male	Yes		2 Graduate	No	1200	1086	17	120	1	Urban	Y
LP001032	Male	No		0 Graduate	No	4950	0	125	350	1	Urban	Y
LP001036	Female	No		0 Graduate	No	3510	0	76	350	0	Urban	N

Seen data / Training data

Unseen data/ Test data

LP001038	Male	Yes		0 Not Graduate	No	4337	0	133	300	1	Rural	
LP001043	Male	Yes		0 Not Graduate	No	7550	0	104	350	0	Urban	
LP001046	Male	Yes		1 Graduate	No	5055	5625	115	350	1	Urban	
LP001047	Male	Yes		0 Not Graduate	No	2600	1911	116	350	0	Semiurban	

DIFFERENT TYPES OF ML ALGORITHMS



THE ML MINDSET

"Machine Learning changes the way you think about a problem.

The focus shifts from a mathematical science to a natural science, running experiments and using statistics, not logic, to analyse its results."

- **Peter Norvig**

THE ML MINDSET

Step	Example
1. Set the research goal.	I want to predict how heavy traffic will be on a given day.
2. Make a hypothesis.	I think the weather forecast is an informative signal.
3. Collect the data.	Collect historical traffic data and weather on each day.
4. Test your hypothesis.	Train a model using this data.
5. Analyze your results.	Is this model better than existing systems?
6. Reach a conclusion.	I should (not) use this model to make predictions, because of X, Y, and Z.
7. Refine hypothesis and repeat.	Time of year could be a helpful signal.

Get Comfortable with Some Uncertainty !

THANK YOU – NEXT WEEK

Week

Topics

Week 1

Intro to ML
Discovering ML Use Cases & ML in Business

Week 2

Python- Hands On
Supervised Learning & Regression



Week 3

Neural Network - 1
Neural Network -2 (Bias, Variance) & Hands ON

Week 4

Kernel Learning & SVM
Practical Advice for ML projects.

Week 5

Boosting
Decision Trees, Random Forest, & xgBoost

Week 6

Unsupervised Learning
Clustering & Dimensionality Reduction

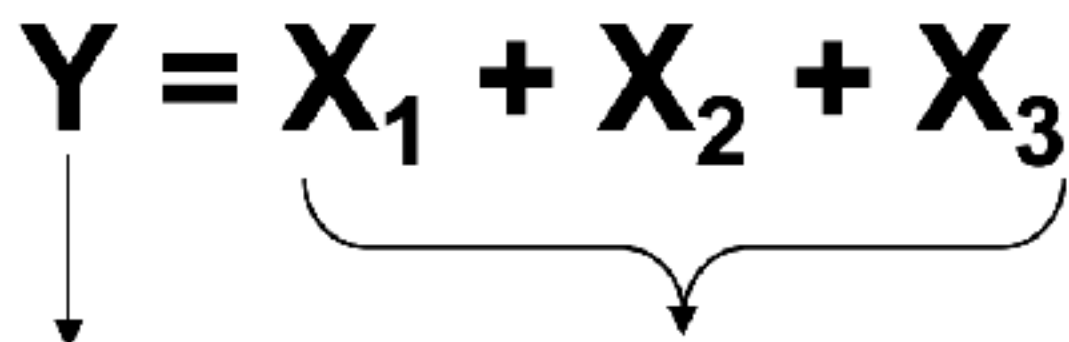
Week 7

Time Series Data Analysis
Imputation & Prediction Systems

Week 8

ML Use Cases from Products & Research

REGRESSION – LINGO

$$\mathbf{Y} = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3$$


Dependent Variable

Independent Variable

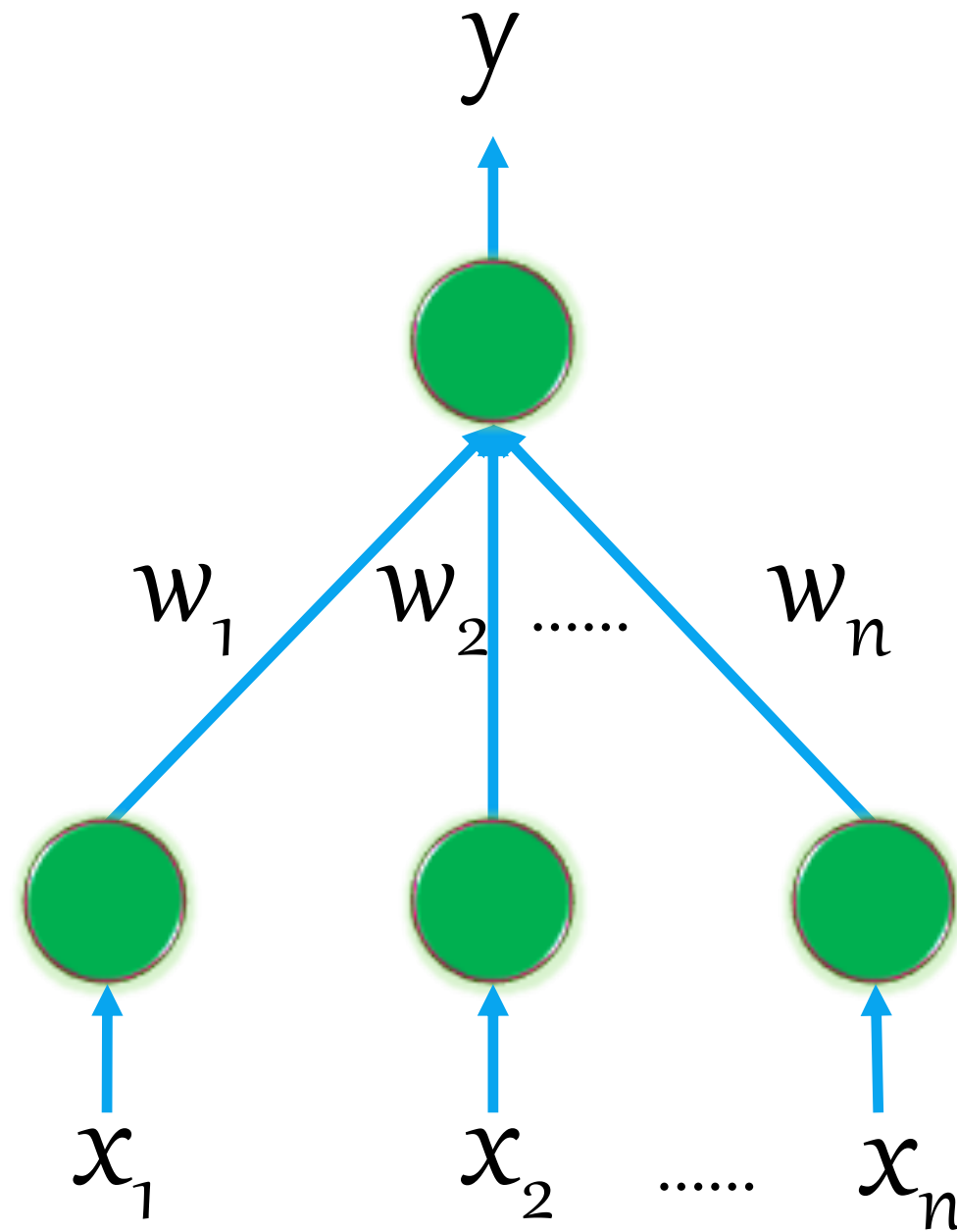
Outcome Variable

Predictor Variable

Response Variable

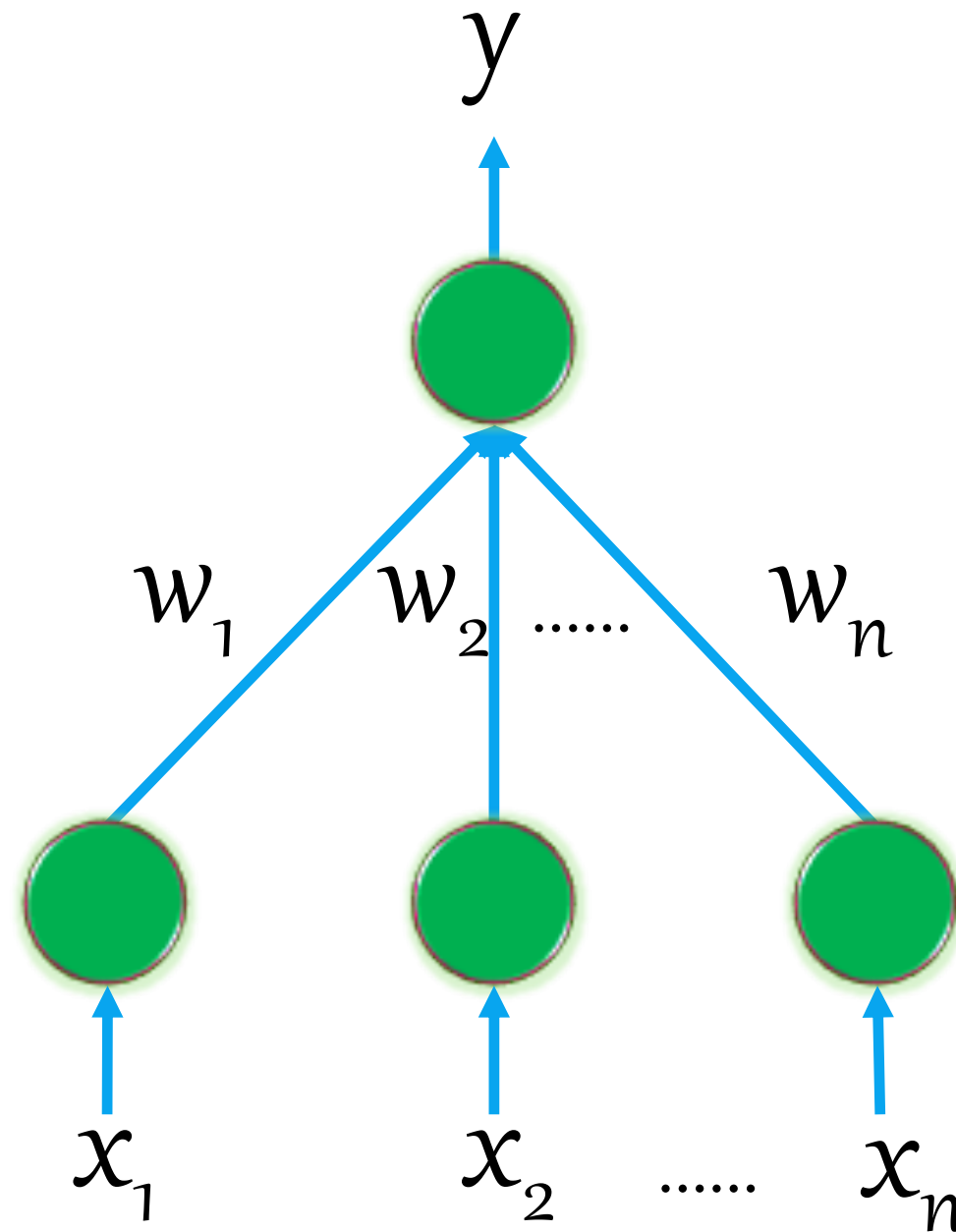
Explanatory Variable

LINEAR REGRESSION



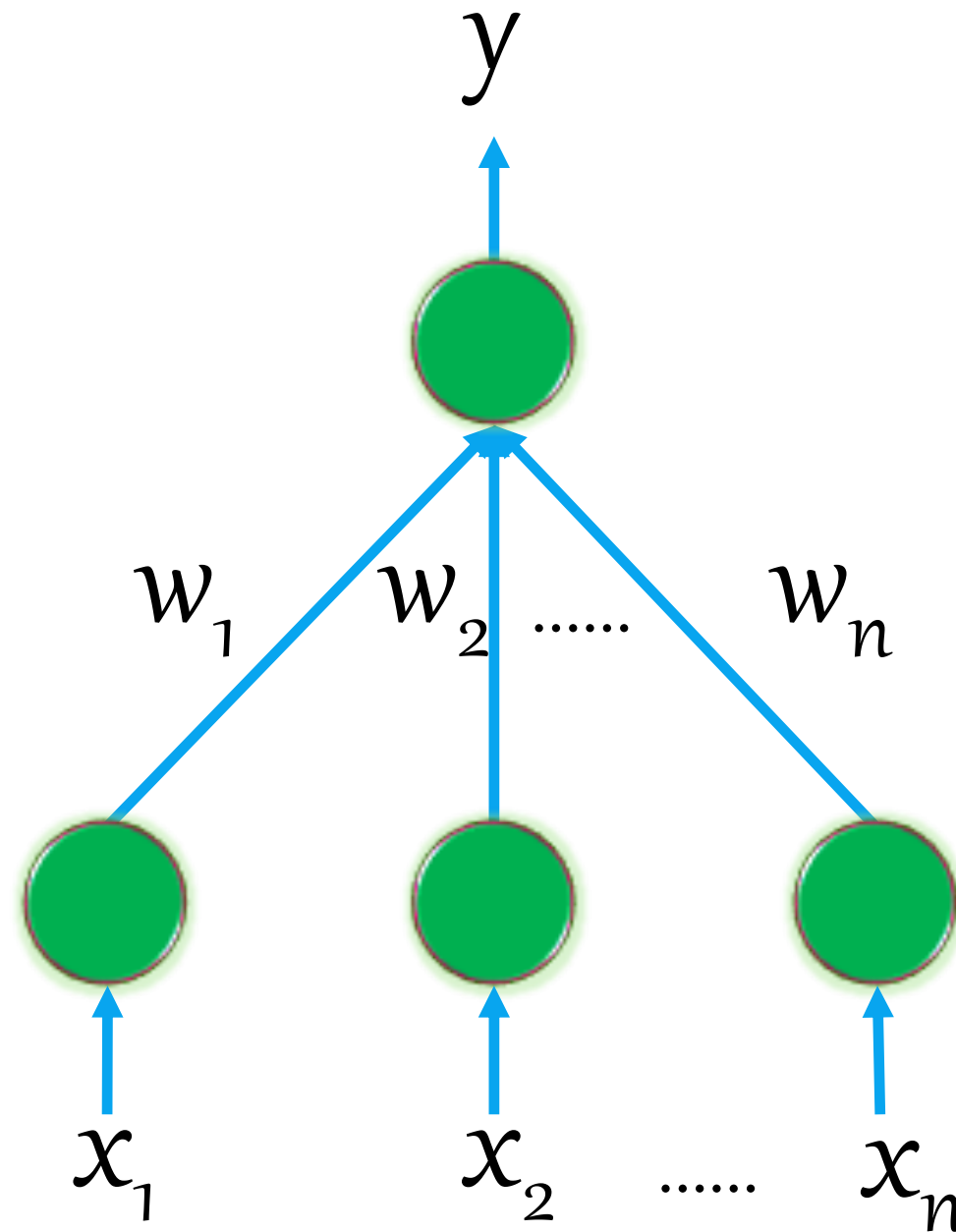
LINEAR REGRESSION – ACTIVATION FUNCTION

$$y = f(\sum_i w_i x_i)$$



LINEAR REGRESSION – ACTIVATION FUNCTION

$$y = f(\sum_i w_i x_i)$$



$$\text{Error} = (y - y')^2$$

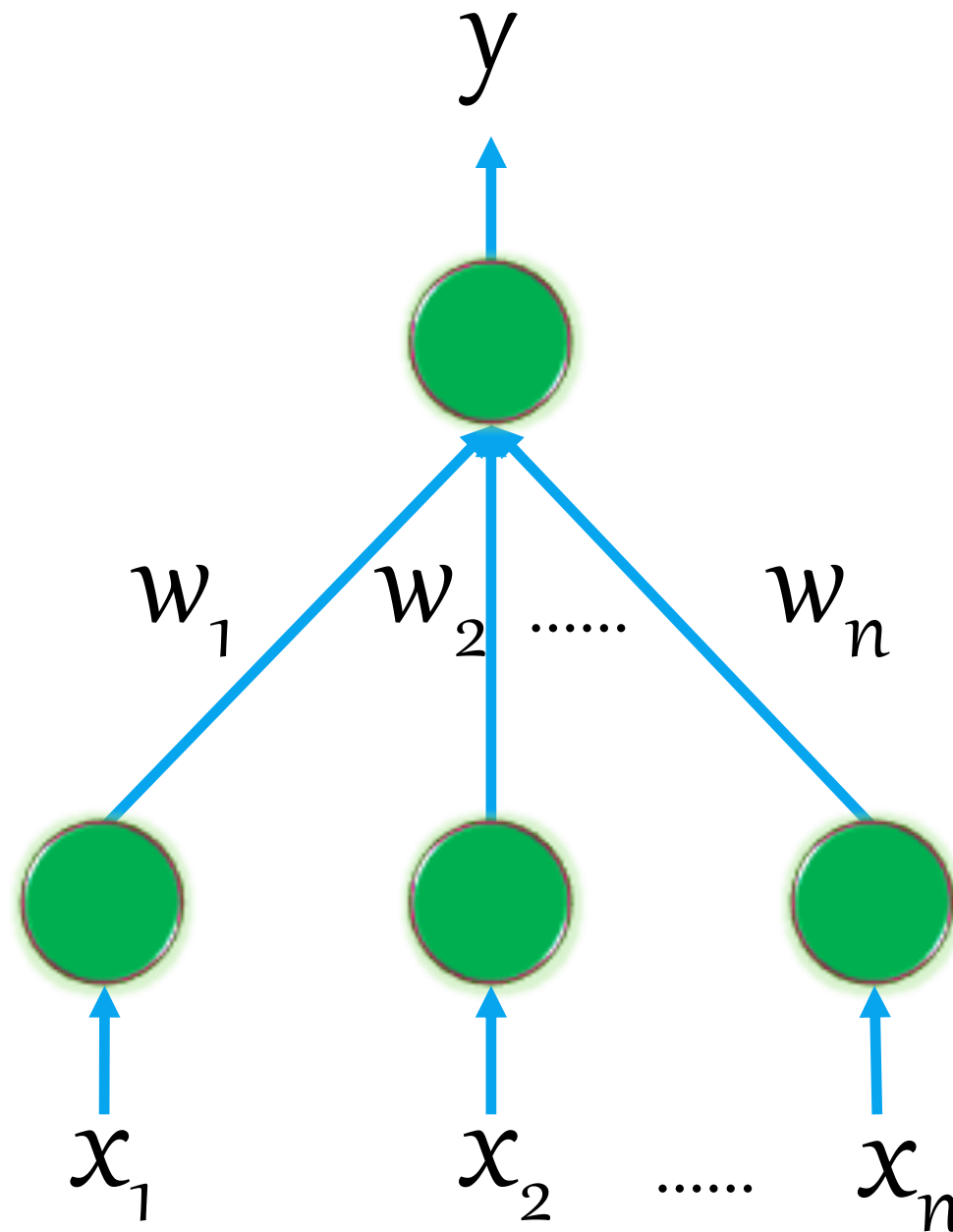
Residual Sum of Squares (RSS)

LINEAR REGRESSION – ACTIVATION FUNCTION

$$y = f(\sum_i w_i x_i)$$

$$\text{Error} = (y - y')^2$$

Residual Sum of Squares (RSS)

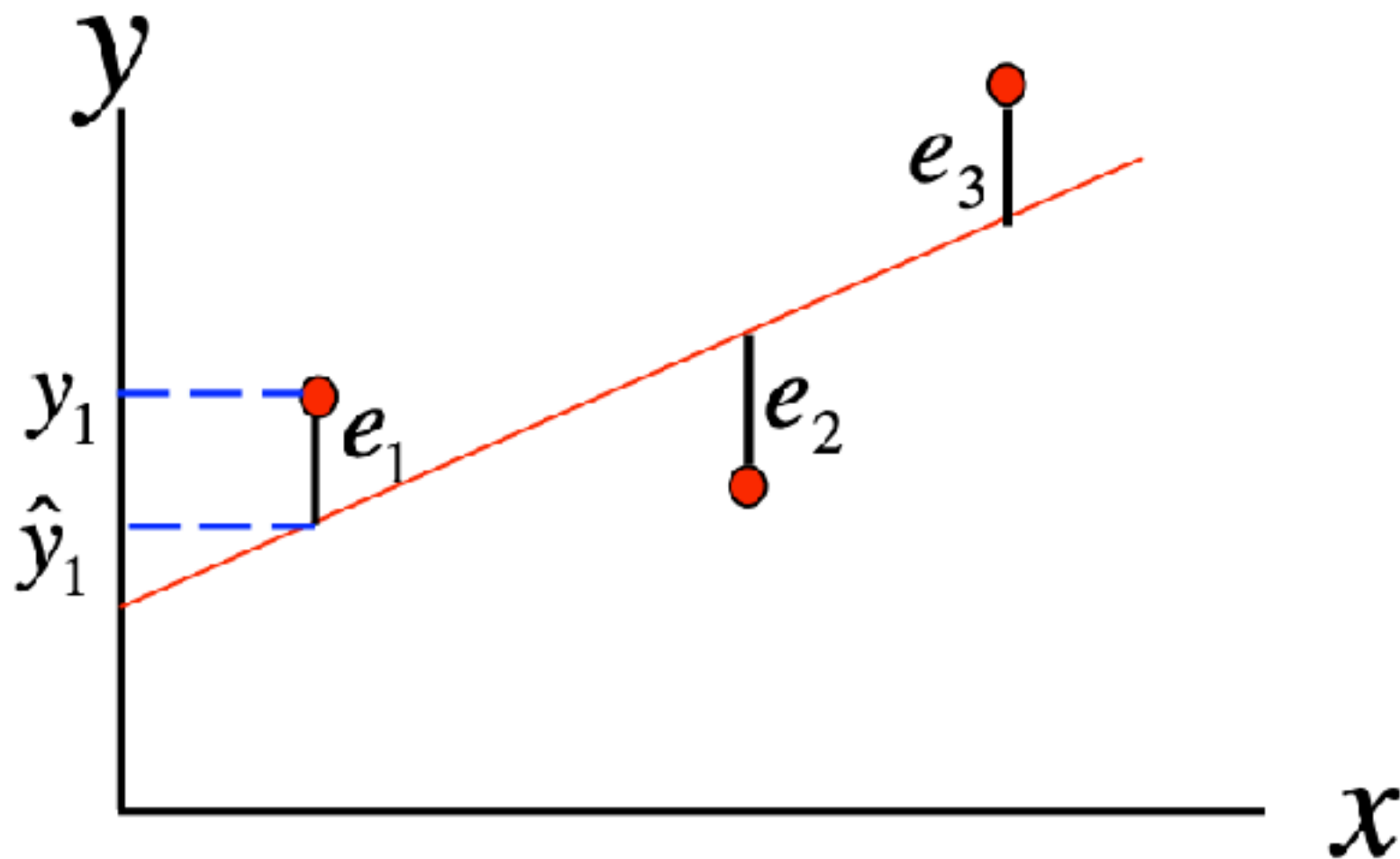


Update \mathbf{w} in such a way that the error is minimized !

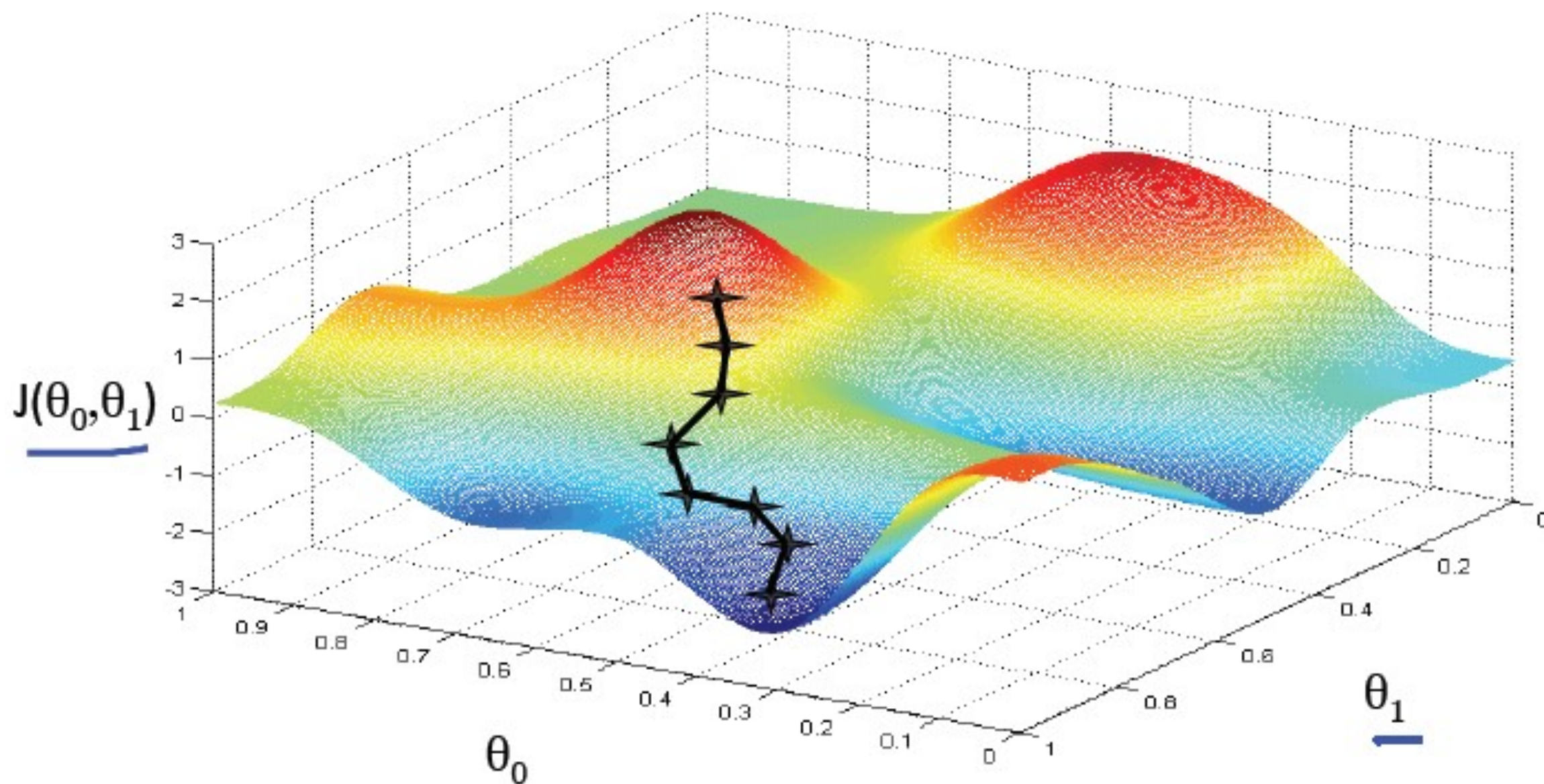
RESIDUAL SUM OF SQUARES

$$e_1 = y_1 - \hat{y}_1$$

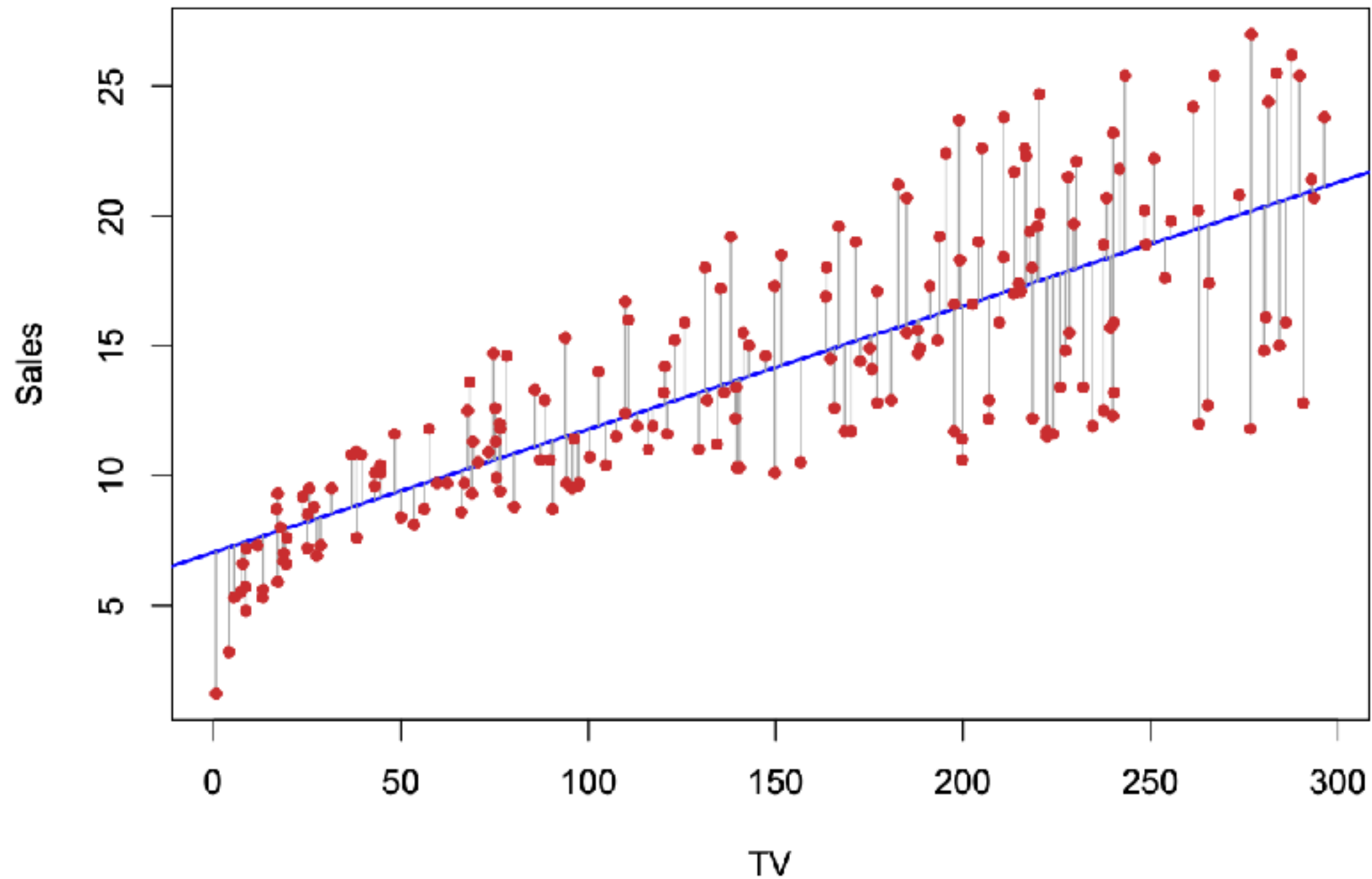
$$e_2 = y_2 - \hat{y}_2$$



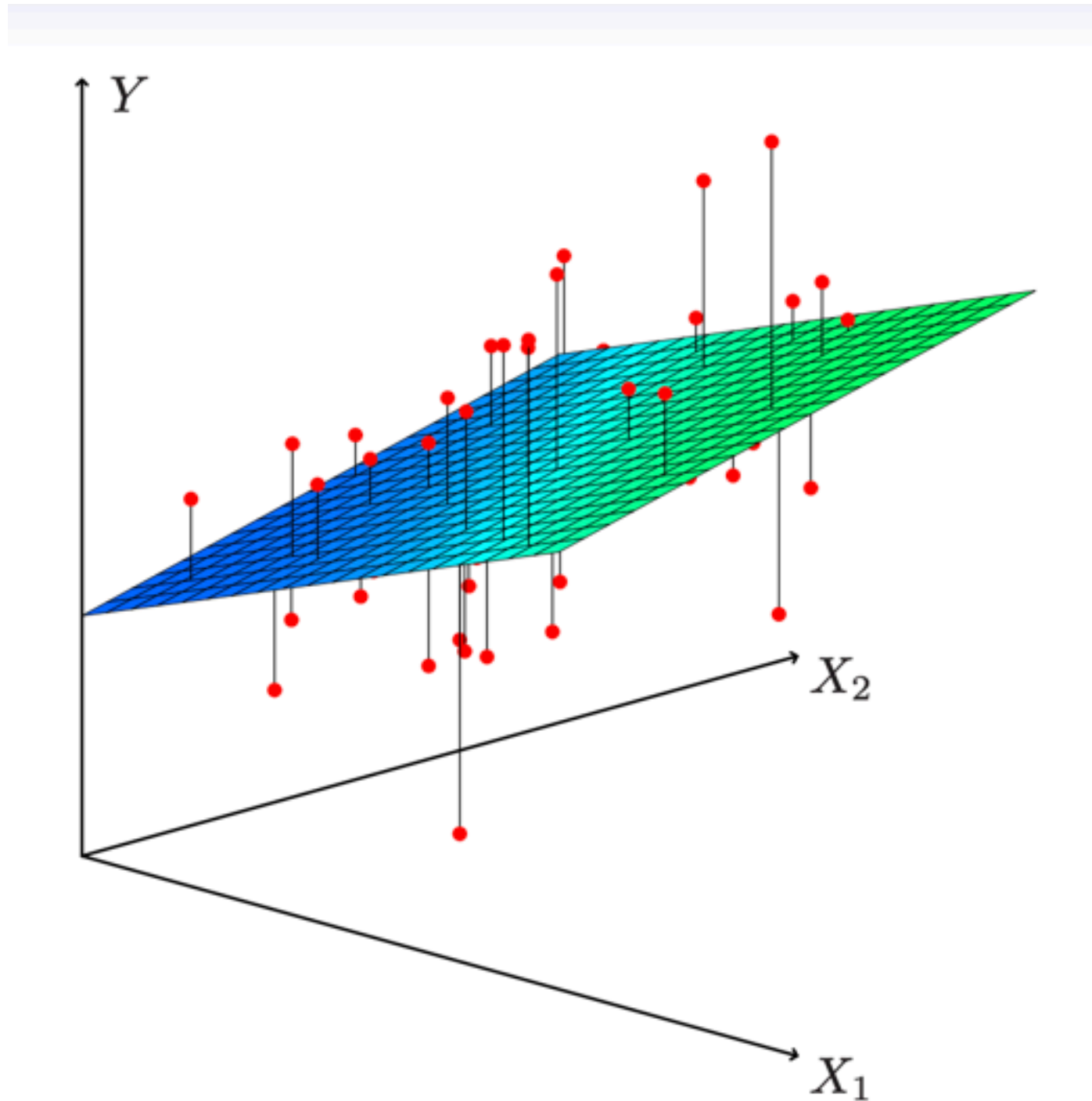
GRADIENT DESCENT ALGORITHM



LINEAR REGRESSION - EXAMPLE

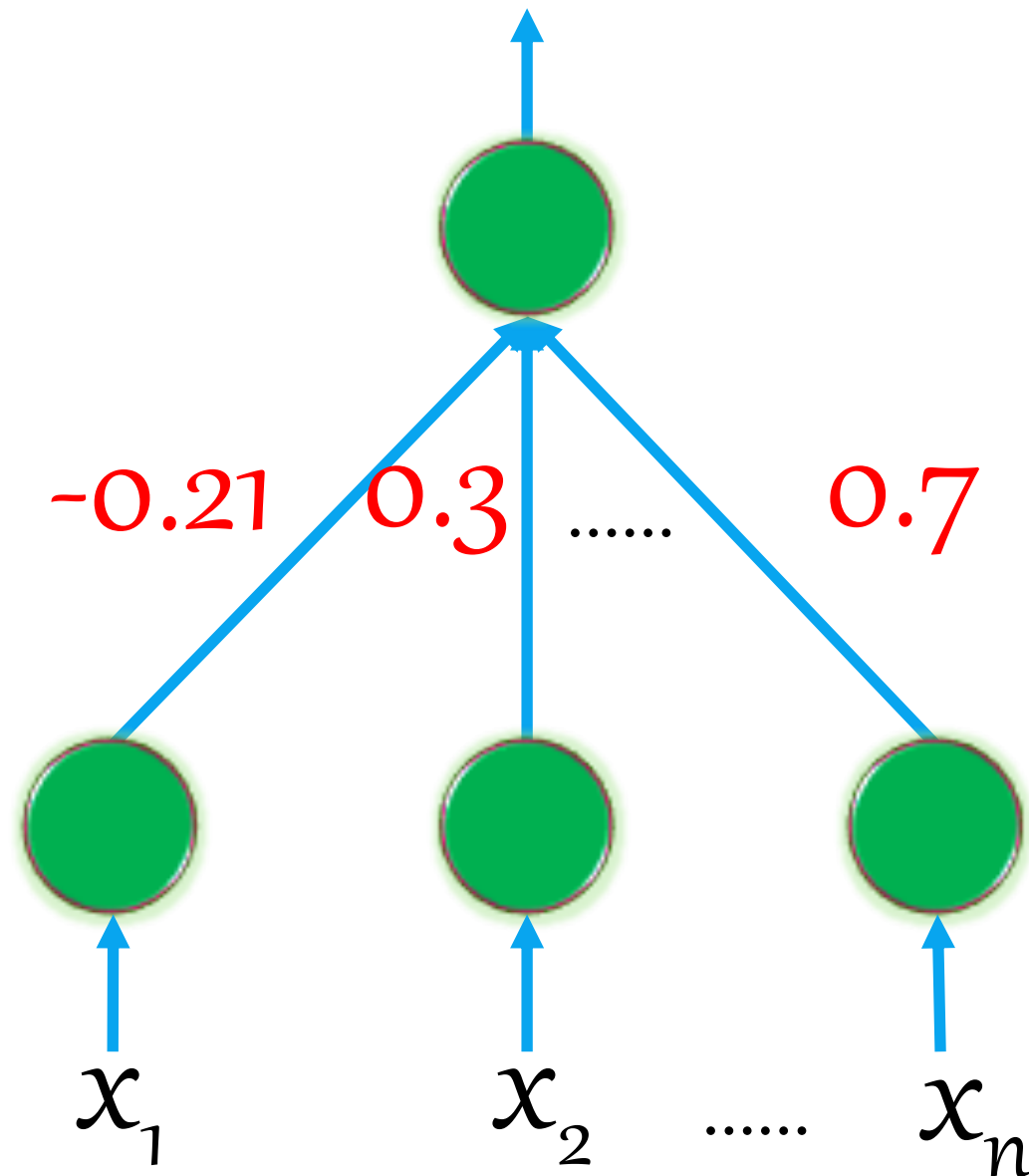


LINEAR REGRESSION – EXAMPLE



LINEAR REGRESSION - EXAMPLE

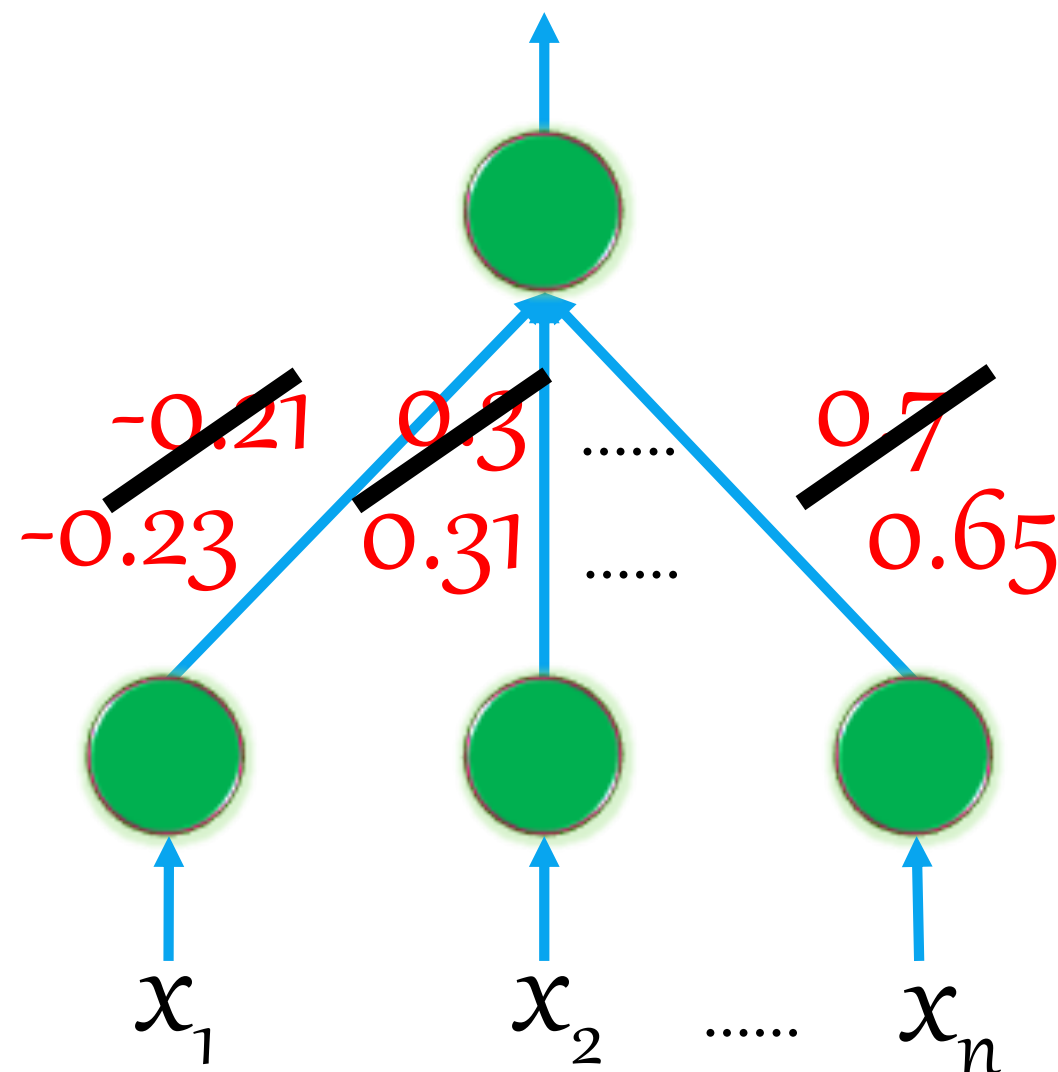
$$y = \max(0, -0.21 * x_1 + 0.3 * x_1 + 0.7 * x_1)$$



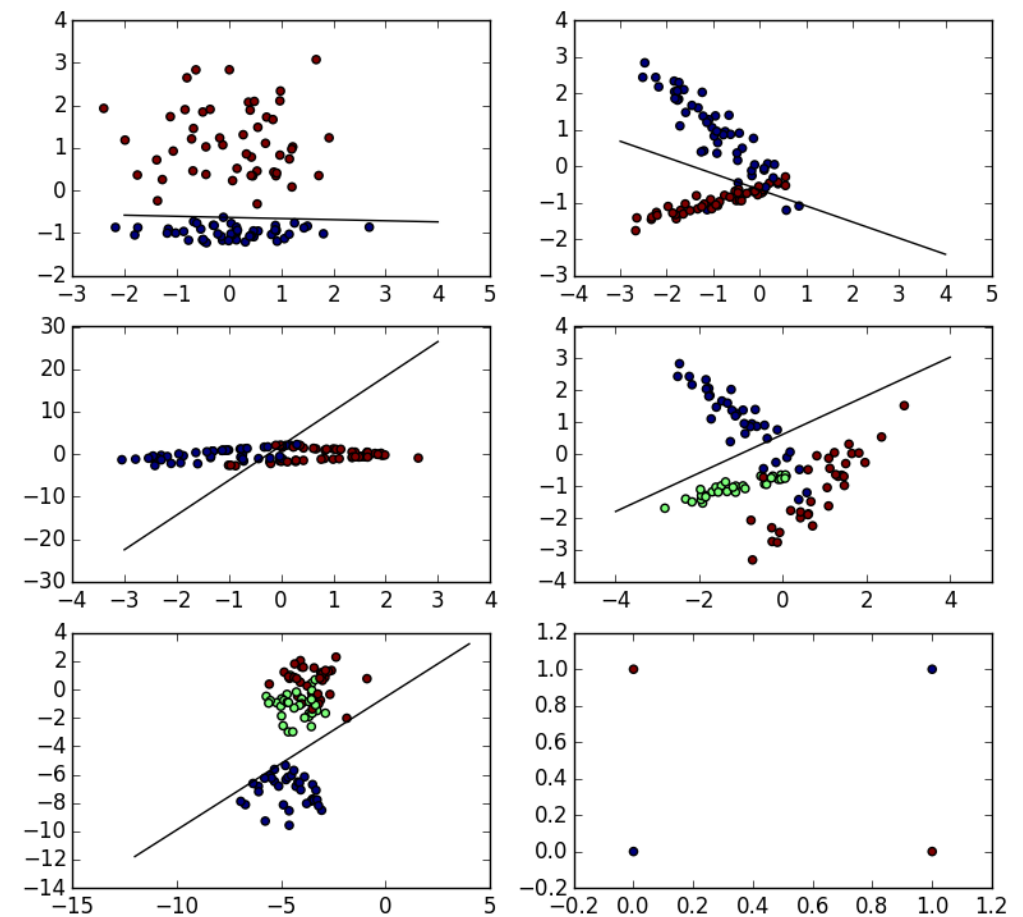
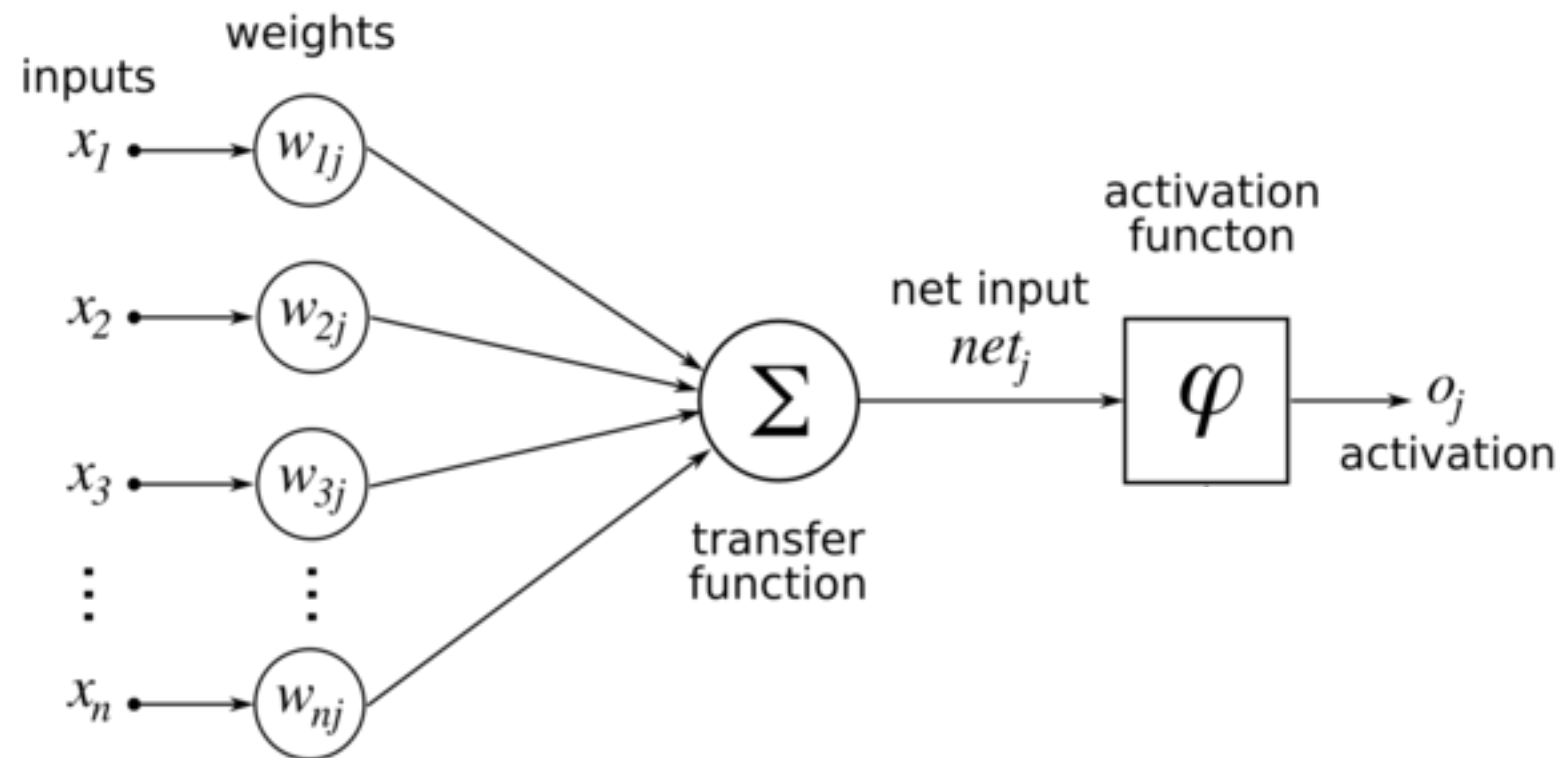
LINEAR REGRESSION - EXAMPLE

$$y = \max(0, -0.23 * x_1 + 0.31 * x_1 + 0.65 * x_1)$$

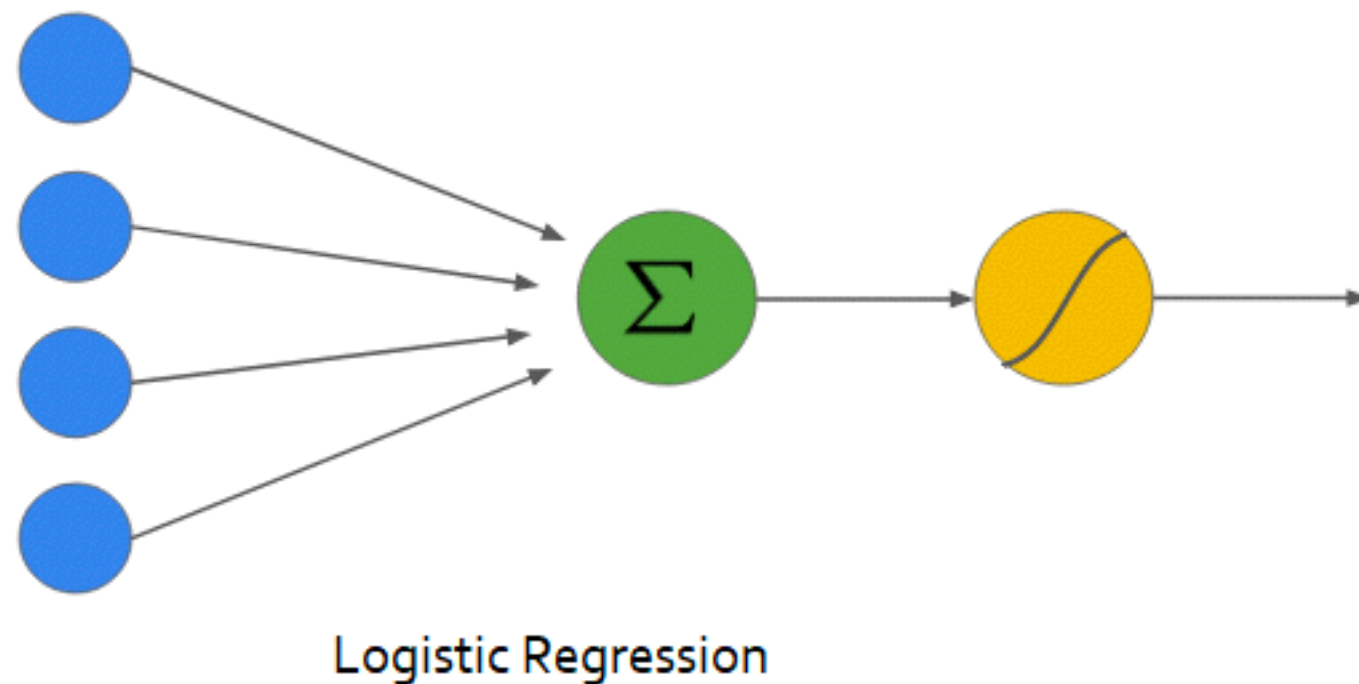
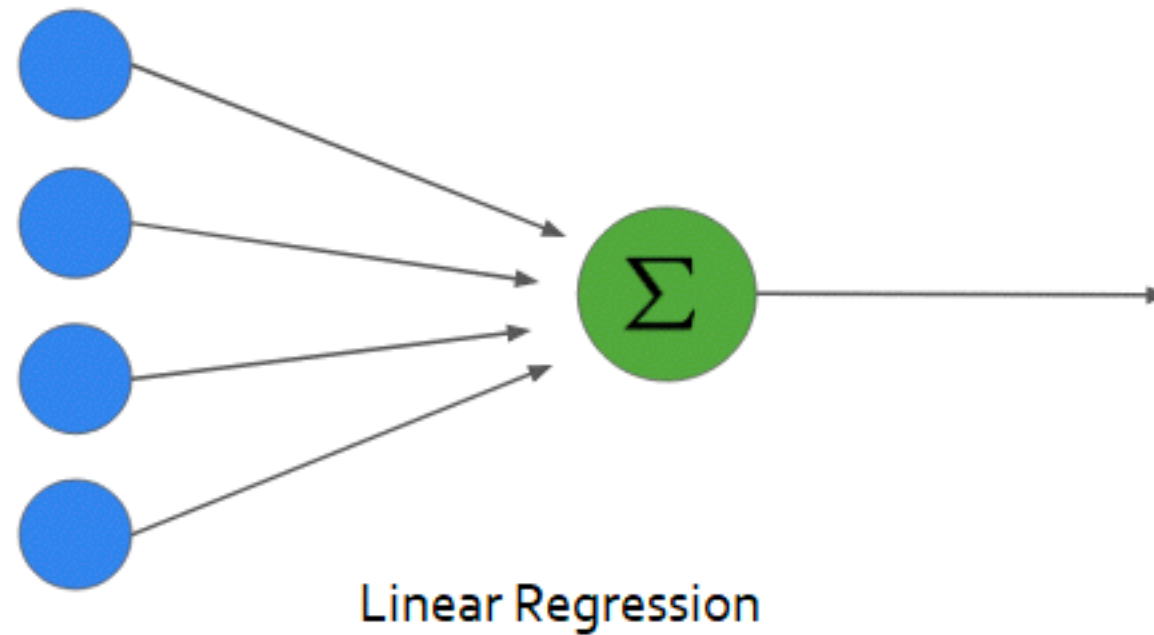
~~$$y = \max(0, -0.21 * x_1 + 0.3 * x_1 + 0.7 * x_1)$$~~



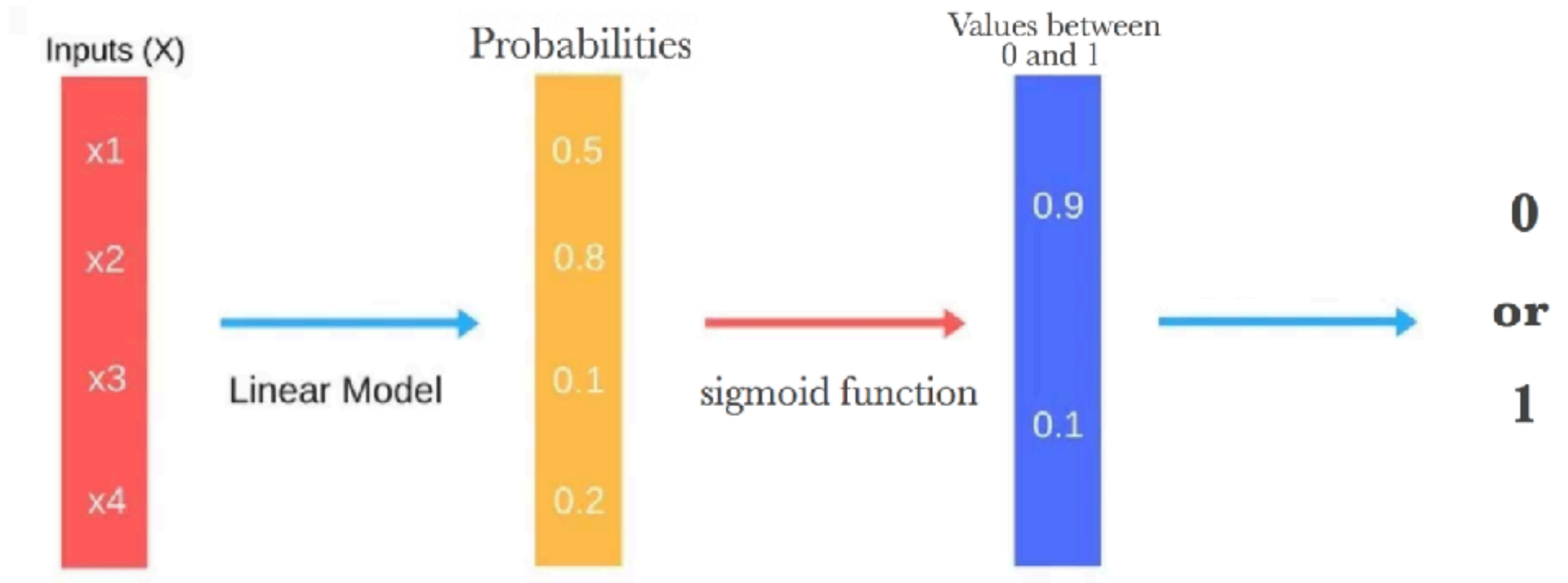
LINEAR REGRESSION – PIPELINE



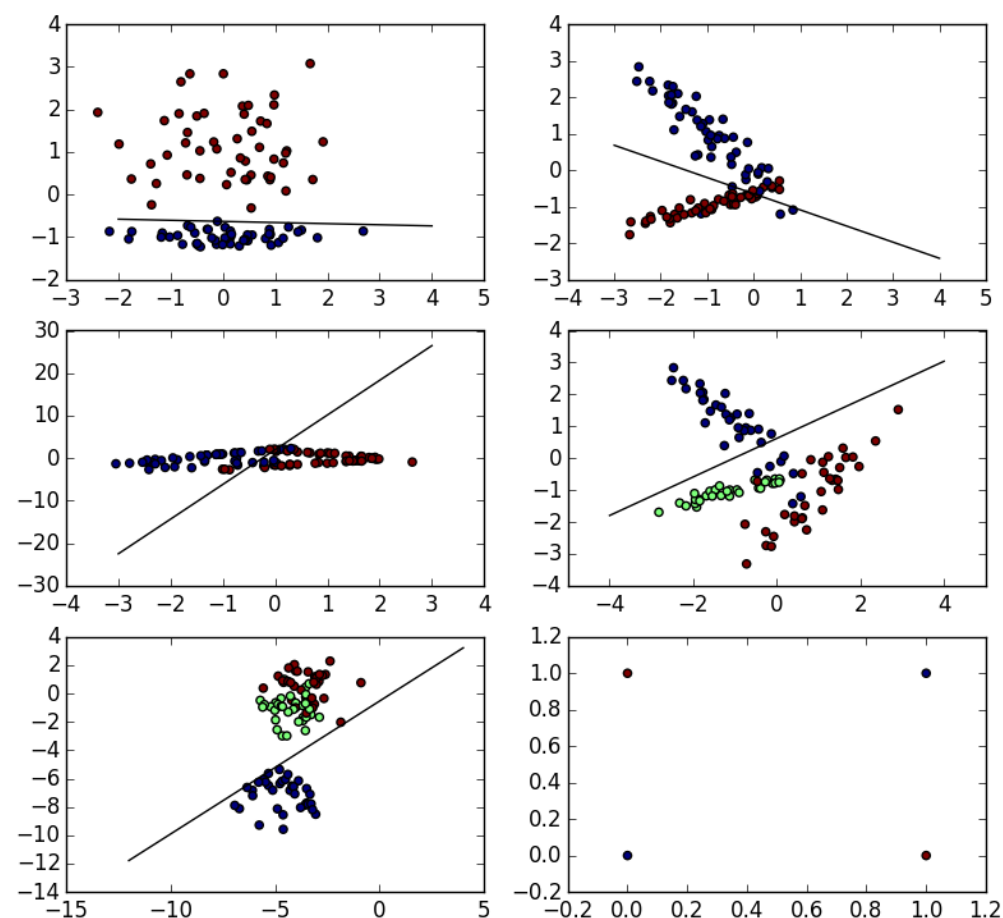
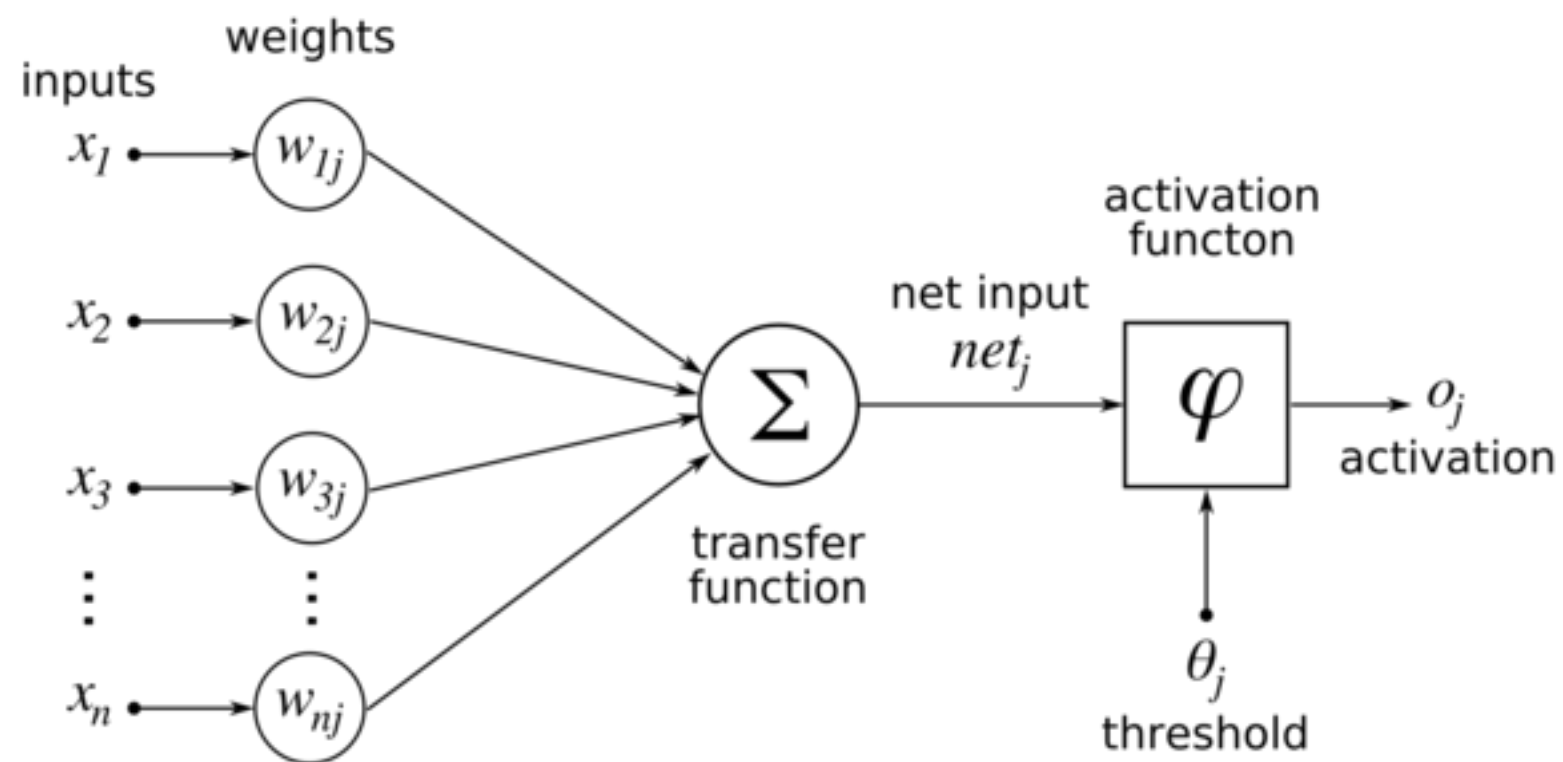
LINEAR VS. LOGISTIC REGRESSION



LOGISTIC AND LINEAR REGRESSION



LOGISTIC REGRESSION



METRICS FOR MEASUREMENT

		true class		total
		EFR	LFR	
predicted class	EFR	True Positives (TP)	False Positives (FP)	predicted EFR
	LFR	False Negatives (FN)	True Negatives (TN)	predicted LFR
		true EFR	true LFR	

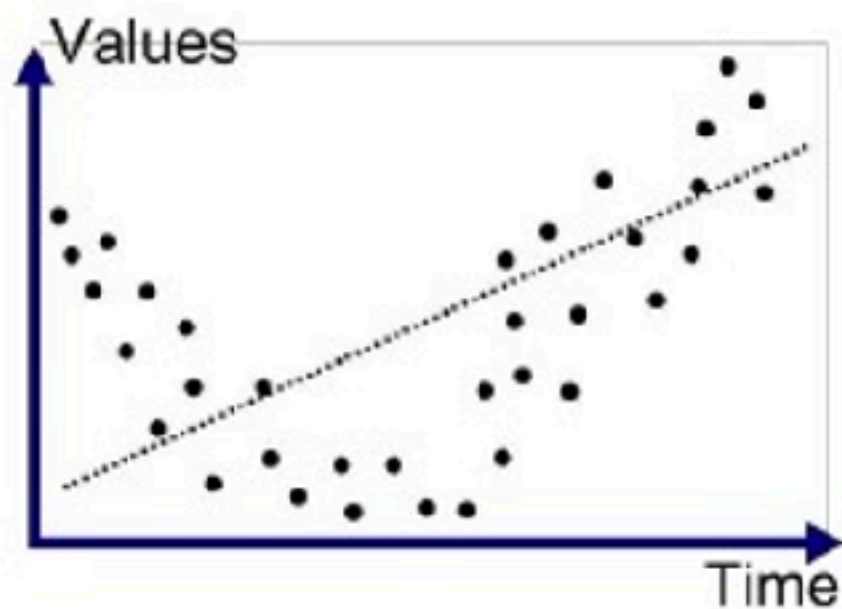
$$PR = \frac{TP}{TP+FP}$$

$$RE = \frac{TP}{TP+FN}$$

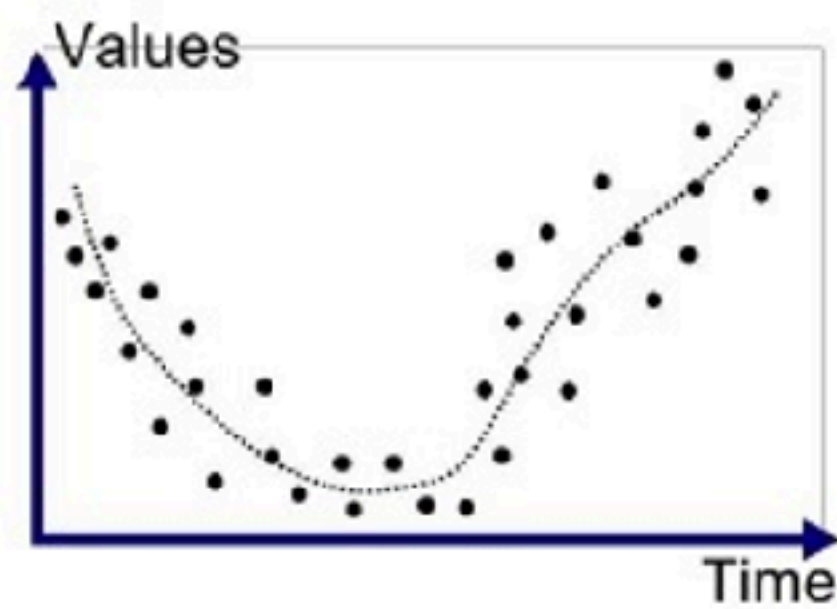
$$CA = \frac{TP+TN}{TP+TN+FP+FN}$$

$$F_1 = \frac{2TP}{2TP+FP+FN}$$

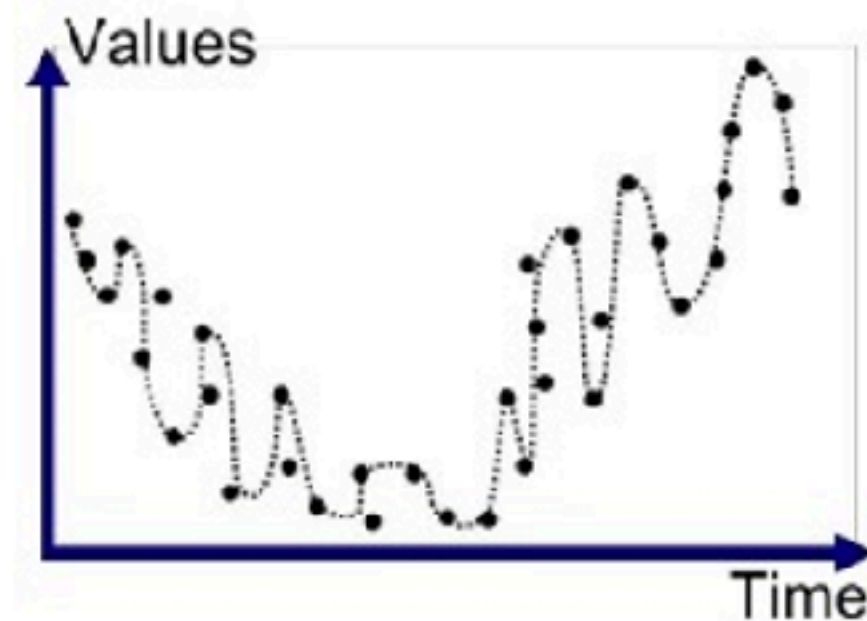
OVERFITTING



Underfitted



Good Fit/Robust



Overfitted

LETS DO CODING !!!

TASKS FOR THIS WEEK

- ▶ Write a blog post on:
 - ▶ What is the difference between logistic and linear regression
- ▶ In the code of logistic regression, plot the confusion matrix and analyse what kind of errors are happening (False positives or False negatives?)
 - ▶ Push your updated code/notebook in GitHub
 - ▶ Update your readme as "Done" after pushing your code

THANK YOU – NEXT WEEK

Week

Topics

Week 1

Intro to ML
Discovering ML Use Cases & ML in Business

Week 2

Python- Hands On
Supervised Learning & Regression

Week 3

Neural Network - 1
Neural Network -2 (Bias, Variance) & Hands ON



Week 4

Kernel Learning & SVM
Practical Advice for ML projects.

Week 5

Boosting
Decision Trees, Random Forest, & xgBoost

Week 6

Unsupervised Learning
Clustering & Dimensionality Reduction

Week 7

Time Series Data Analysis
Imputation & Prediction Systems

Week 8

ML Use Cases from Products & Research