https://www.youtube.com/embed/io87SbDOKqM?start=60&end=97&version=3

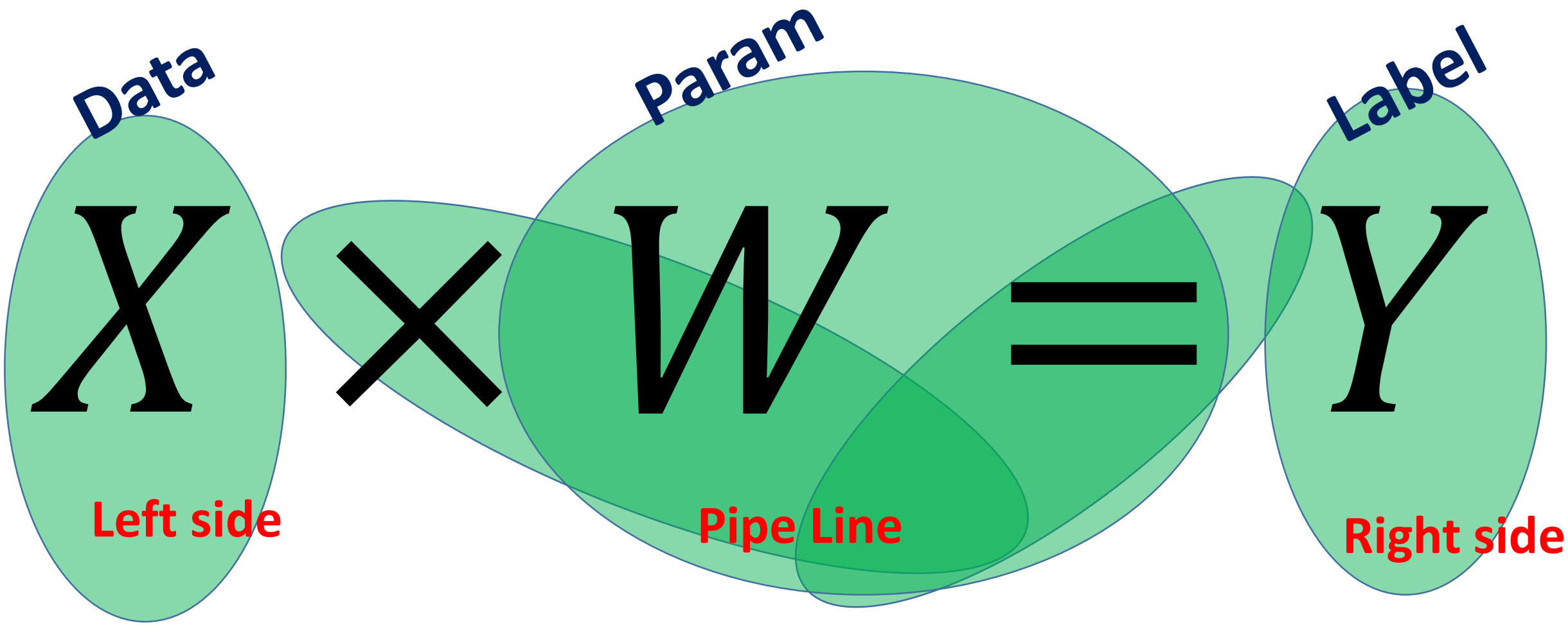
https://www.youtube.com/embed/RBnIbqW6ZhM?start=15&end=75&version=3

ML Key Terminology and living with ambiguity…

Dr. Kalidas Y.

In this lecture you will learn about 21 key phrases and known ambiguities to live with in Machine Learning World

1) key phrases… “Data, Label, Parameter”



$N \times k$ *matrix*

$k \times 1$ *vector*

$N \times 1$ *vector*

$$X \times w = y$$

2) key phrase… “Data”

X is called **Data**

$$X_{N \times k}$$

Dimensionality of a data point

Number of ***data points***

Attributes,
Features,
Dimensions,
Properties,
Columns

$$X = \begin{bmatrix} x_0^T \\ \cdots \\ \cdots \\ x_i^T \\ \cdots \\ x_N^T \end{bmatrix}$$

Single Data Point

$$x_i$$

It's a column vector

k dimensional point

Each row corresponds to…
transpose of a data point

By default assumed to be (k x 1) matrix

3) key phrase… “Targets,

…Labels, Actuals, **Ground** truth”

target vector

$$y = \begin{bmatrix} y_1 \\ \dots \\ y_i \\ \dots \\ y_N \end{bmatrix}$$

pronounced as…
“target i” or “label i”

N× 1 *vector*

4) key phrase… "Supervised"

- Given x_i and y_i pairs of points
- NOTE: (in case of problems we have seen so far…)
- x_i is a k-dimensional vector
- y_i is a scalar

We will see trivial extension to the case when y_i can be a vector as well

5) key phrase… “Data Set”

Pair of points

in

$$D = \{(x_i, y_i)\}_{i \in [1..N]}$$

Set of Points

Pronounced as…

“x i y i” with i over 1 to N

For each point i

1 to N

6) key phrase… "Model"

f(x) is a scalar i.e. 1x1

Linear Model

$$f(x) = x^T w$$

x^T is (1 x k) vector

w is (kx1) vector

This function invocation is called PREDICTION for a given data point

$$f(x) = x \cdot w$$

DOT PRODUCT FORM

7) key phrase… "Prediction"

"prediction for i^{th} point"

$$y_i' = f(x_i) = x_i \cdot w$$

pronounce as "y prime i"

$$y' = \begin{bmatrix} y'_1 \\ y'_2 \\ \dots \\ y'_i \\ \dots \\ y'_N \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \dots \\ f(x_i) \\ \dots \\ f(x_N) \end{bmatrix} = \begin{bmatrix} x_1 \cdot w \\ x_2 \cdot w \\ \dots \\ x_i \cdot w \\ \dots \\ x_N \cdot w \end{bmatrix}$$

8) key phrase… "Loss function"

$$L(w) = \left|\left|y - y'\right|\right|^2$$

$$L(w) = \sum_{i=1}^{i=N} (y_i - x_i \cdot w)^2$$

Actual value

Predicted value

$$L(y_i, f(x_i)) = (f(x_i) - y_i)^2$$

Squared Loss Function

9) key phrase… "Gradient of Loss function"

$$L(w) = \sum_{i=1}^{i=N} (y_i - x_i \cdot w)^2$$

$$\nabla L = \frac{\partial L}{\partial w} = 2\, X^T (Xw - y)$$

its only, a notational convenience…

differentiating with respect to w

Squared Loss Function

For example,

$$L(W) = (XW - Y)^T (XW - Y)$$

Multi Variate - Loss Function (or simply just *Loss Function*)

For example,

$$L(w) = (Xw - y)^T (Xw - y)$$

$w_{K\times 1}$ is K dimensional vector

Popular two - Regression Loss Functions

$$L(w) = (Xw - y)^T (Xw - y)$$

Squared Error

$$= ||Xw - y||_2 = \sum_{i=1}^{i=N} (y_i - x_i \cdot w)^2$$

L2- Norm

$$L(w) = ||Xw - y||_1 = \sum_{i=1}^{i=N} |y_i - x_i \cdot w|$$

Absolute Error

L1- Norm

Popular… Regression Loss Functions

$$L(w) = \frac{1}{N}\sum_{i=1}^{i=N}(y_i - x_i \cdot w)^2$$

Mean Squared Error (MSE)

$$L(w) = \frac{1}{N}\sum_{i=1}^{i=N}|y_i - x_i \cdot w|$$

Mean Absolute Error (MAE)

Popular… Regression Loss Functions

***Root** Mean Squared Error (RMSE)*

$$L(w) = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}(y_i - x_i \cdot w)^2}$$

10) key phrase… "BUILDING A MODEL"

- Determining a w such that $L(w)$ is minimized

  Symbolically denoted as,

$$w^* = \underset{w}{\operatorname{argmin}}\, L(w)$$

- NOTE: $\min_{w} L(w)$ is different from $\underset{w}{\operatorname{argmin}}\, L(w)$
- $\min_{w} L(w)$ means **minimum value** of Loss function across various values of w
- $\underset{w}{\operatorname{argmin}}\, L(w)$means **minimizing vector** w for a given Loss function

11) key phrase… “Training/Learning/Fitting a model”

- “Building a model” on a given data set

12) key phrase… “model deployment”

- “Building a model” and using it “to predict” on “new data points”

13) key phrase… "model evaluation"

$$L(w) = \sum_{i=1}^{i=N} (y_i - w \cdot x_i)^2$$

this is an implicit form.. "data set" is assumed to have been given

explicitly compute loss function value on a given "data set"

data set

$$L(w, D) = \sum_{(x,y) \in D} (y - x \cdot w)^2$$

point in a data set

each point in a data set is a tuple

14) key phrase… “model performance”

- **Loss** function value in “model evaluation” being **low**

15) key phrase… “model maintenance”

- “model performance” being high over several months or weeks
- IF “model performance” is LOW
- THEN
  - Identify suitable “new data set”
  - re-“Train model”
  - re-”Evaluate model”
  - re-”Deploy model”
- ELSE
  - continue using

16) key phrase… “production model”

- A “deployed model” already being used in a real world setting

17) key phrase… “Training and Testing”

- We need to mimic or simulate real world scenario

- A “production model” always “sees new data”

- With a “given data set”, we need to mimic this scenario

- Create “train set” to build “temporarily a production model”
- Create “test set” to do “model evaluation” on “new data”

18) key phrase… “Training Set”

- A “data set” “of points” reserved for “training a model”

19) key phrase… “Test Set”

- A “data set” “of points” reserved for “model evaluation”

20) key phrase… “Training/Emperical Error”

- “model performance” on “training set”
- Loss function value on “training set”

21) key phrase… “Test/Generalization Error”

- “model performance” on “test set”
- Loss function value on “test set”