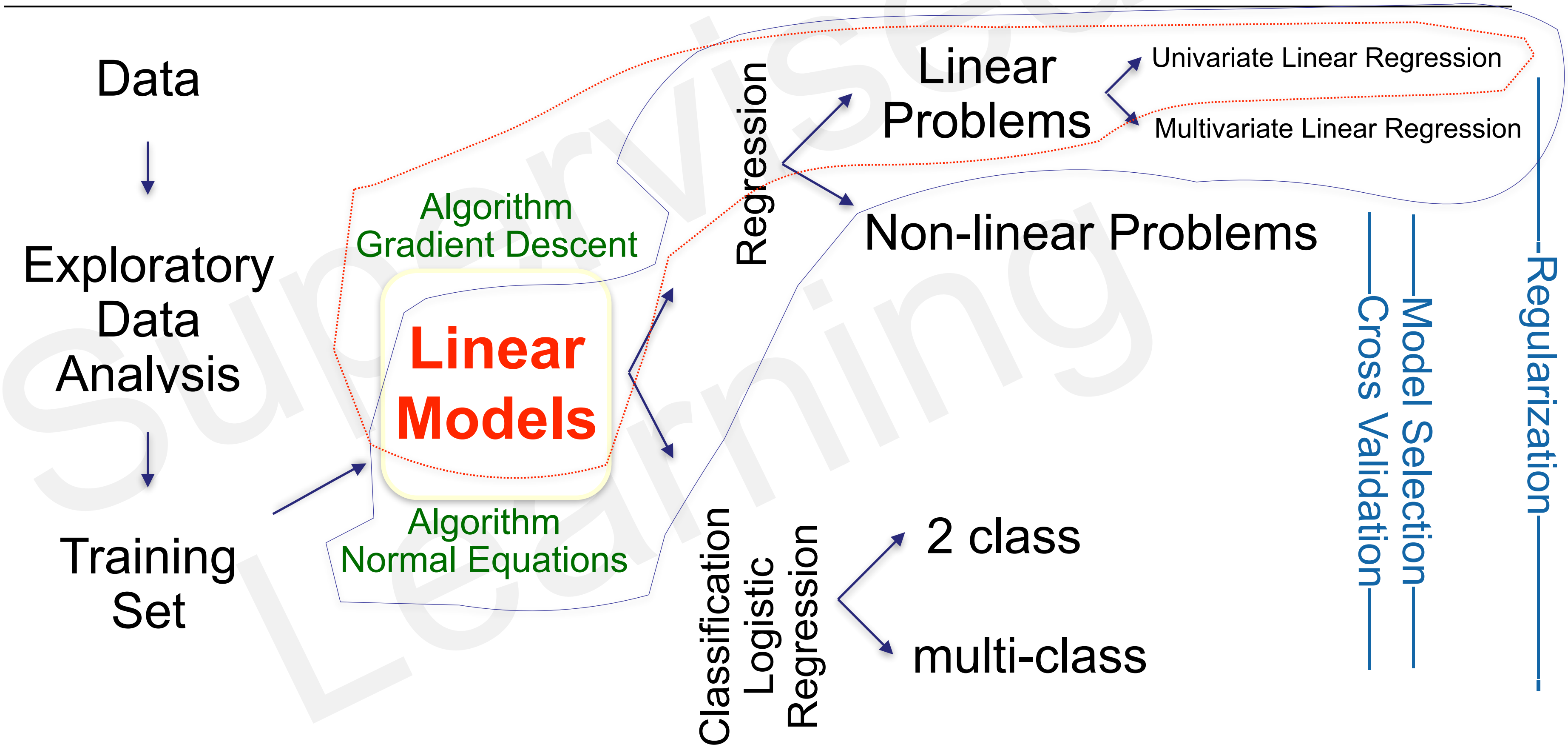
GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 5: LINEAR MODELS & LINEAR REGRESSION

WHERE ARE WE ON THE DATA SCIENCE ROAD-MAP?



Data

Exploratory Data Analysis

Training Set

Algorithm Gradient Descent

Linear Models

Algorithm Normal Equations

Regression

Linear Problems

Univariate Linear Regression

Multivariate Linear Regression

Non-linear Problems

Classification Logistic Regression

2 class

multi-class

Cross Validation

Model Selection

Regularization

Finding the minimum of the cost function determines values for θ that optimally (in the sum of squared errors sense) model the training set

A second algorithm for finding the minimum of the cost function are via the Normal Equations

Standard minimization problem

Take the derivative and set equal to zero

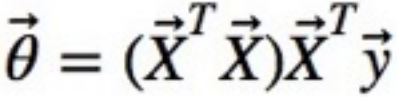
Using the training set can derive a group of simultaneous equations

This is not an iterative algorithm, rather it is purely analytical and solves for the final values of theta directly

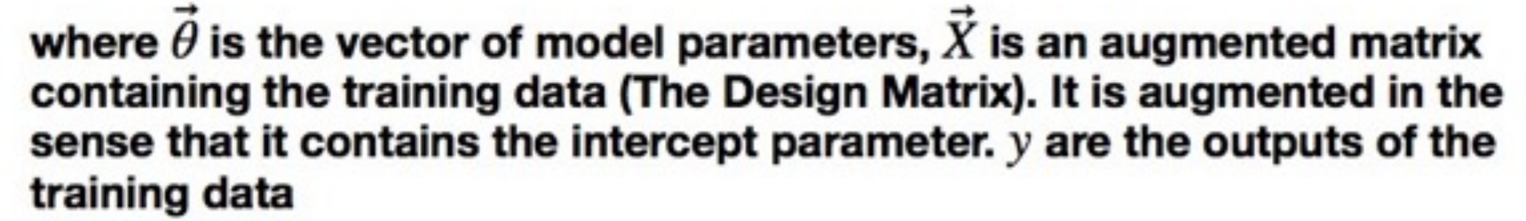
The matrix containing the training data is called ‘The Design Matrix’

Although there is no strict need to feature scale, by force of habit I generally always scale my features

KEY CONCEPTS - NORMAL EQUATIONS



$$\vec{\theta} = (\vec{X}^T \vec{X})\vec{X}^T \vec{y}$$



where $\vec{\theta}$ is the vector of model parameters, $\vec{X}$ is an augmented matrix containing the training data (The Design Matrix). It is augmented in the sense that it contains the intercept parameter. y are the outputs of the training data

Key Concepts - Linear Models and Multivariate Linear Regression

$y = \theta_0 + \theta_1 x$ Univariate Linear Regression

$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \ldots \theta_n x_n$ Multivariate Linear Regression

Extend the number of features beyond 1.

KEY CONCEPTS - ASSUMPTIONS FOR LINEAR REGRESSION

1. Linear relationship - the relationship between features and output is linear
2. Multivariate normality - the features are identically distributed
3. No or little multicollinearity - the features are independent
4. No auto-correlation - the residuals are independent of each other
5. Homoscedasticity - the variance in the residuals is ‘constant'

If the features are not independent and exhibit multi-collinearity, the normal equations can become ill-posed

Adding a regularizer solves this problem

KEY CONCEPTS - GRADIENT DESCENT VS NORMAL EQUATIONS

1. Gradient Descent

i. An iterative algorithm
ii. Need to know the cost function
iii. Start with an initial random guess for the model parameters
iv. At each iteration we alter the parameters a little to reduce the cost function
v. Stop when a minimum of the cost function is reached
vi. The training data must be scaled before we used this algorithm
vii. We need to determine a hyper-parameter, called the learning rate
viii. Why do you need to know about this algorithm??

KEY CONCEPTS - GRADIENT DESCENT VS NORMAL EQUATIONS

2. Normal Equations

i. It is possible to solve for the model parameters directly!!

ii. Enter the training data (inputs and outputs) into a set of equations and determine the values of theta

iii. No need to scale the inputs

iv. No hyper-parameter to estimate

v. Involves the calculation of the inverse of a matrix

vi. As the dataset gets large this becomes very computationally expensive

vii. Equations can become unstable if linear regression assumptions are violated