Introduction to ML-Classification

DSLA COURSE

ROHIT PADEBETTU

Course Assignments

Programming Assignments

Reading Assignments

Presentation Assignments

Technical Skills Assignments

Writing Assignments

Technical Assignment

Install & Familiarize yourself with R and RStudio IDE

Programming Assignment

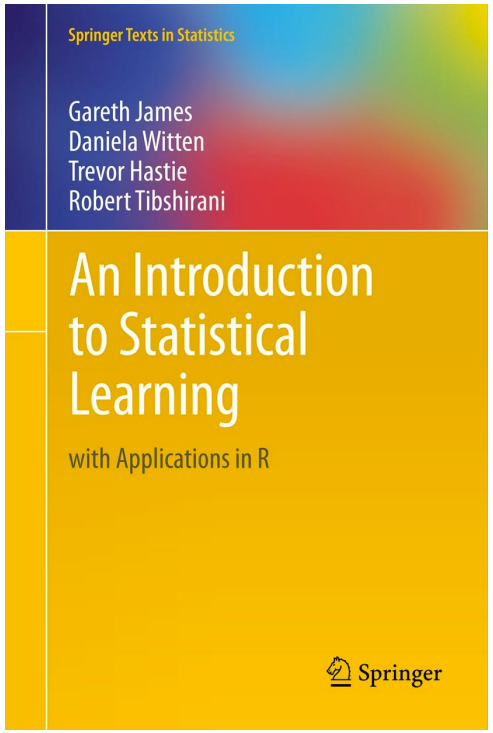
Install & Complete: Git / GitHub Tutorial

Install & Complete: Swirl – R Programming

Reading Assignment

Read Chapter 1: Introduction
Read Chapter 2: Statistical Learning

Read Chapter 3: Linear Regression
Read Chapter 4: Classification



Writing Assignment

Create & Submit by Wednesday
Aspirational Resume

What do you want your Resume to contain in 3 months ?

- Two pages maximum
- List 3 projects or jobs
- Explain the problem
- Describe mathematical or statistical techniques used
- Specific Metric you Improved

Presentation Assignment

By Tuesday Submit

1. 2 Industries hiring data scientists where you would like to work
2. 2 Types of data science problems they solve
3. Identify the Algorithms typically used to solve those problems

By Saturday Submit

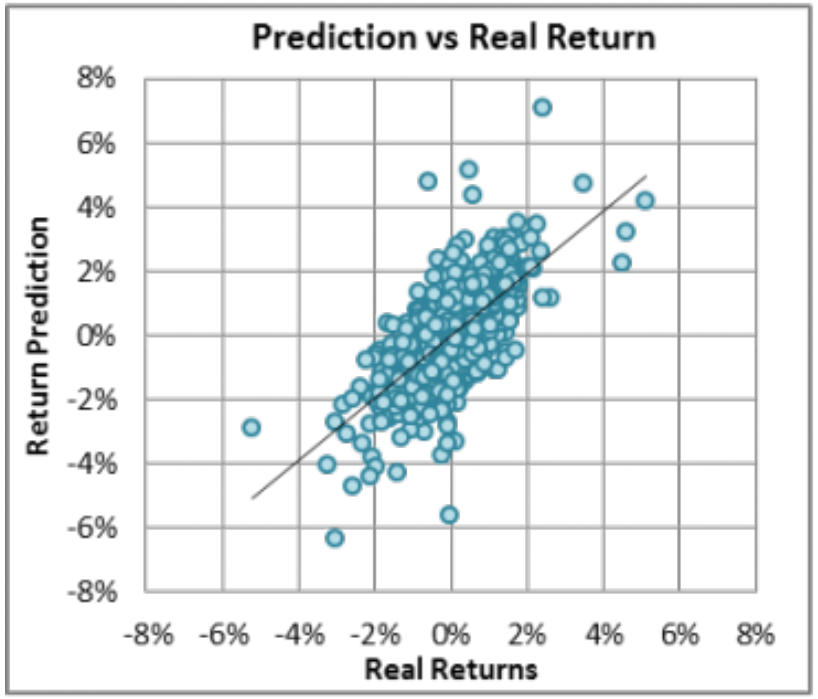
3 examples of data science problems in your personal/professional life.

For each give

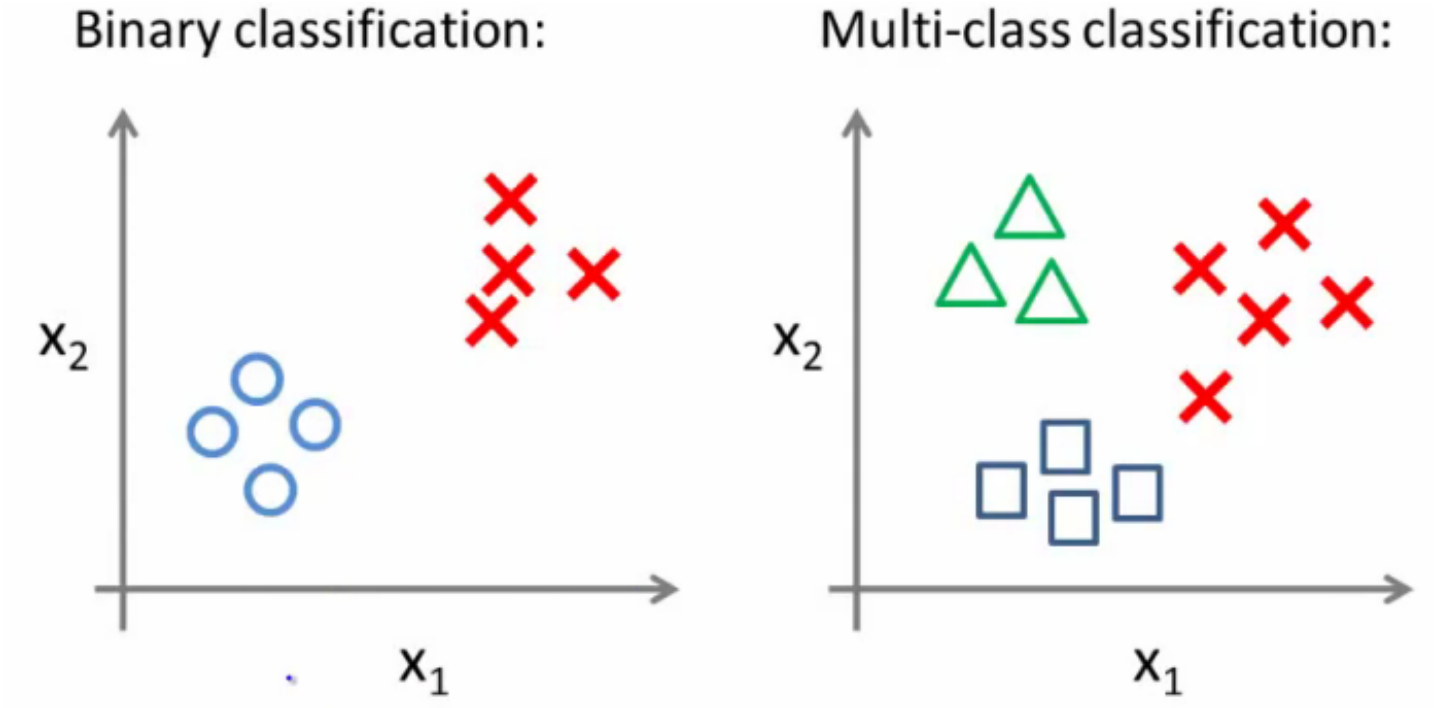
1. Relevant Question
2. Hypothesis
3. Data Source
4. Algorithm

Regression vs Classification

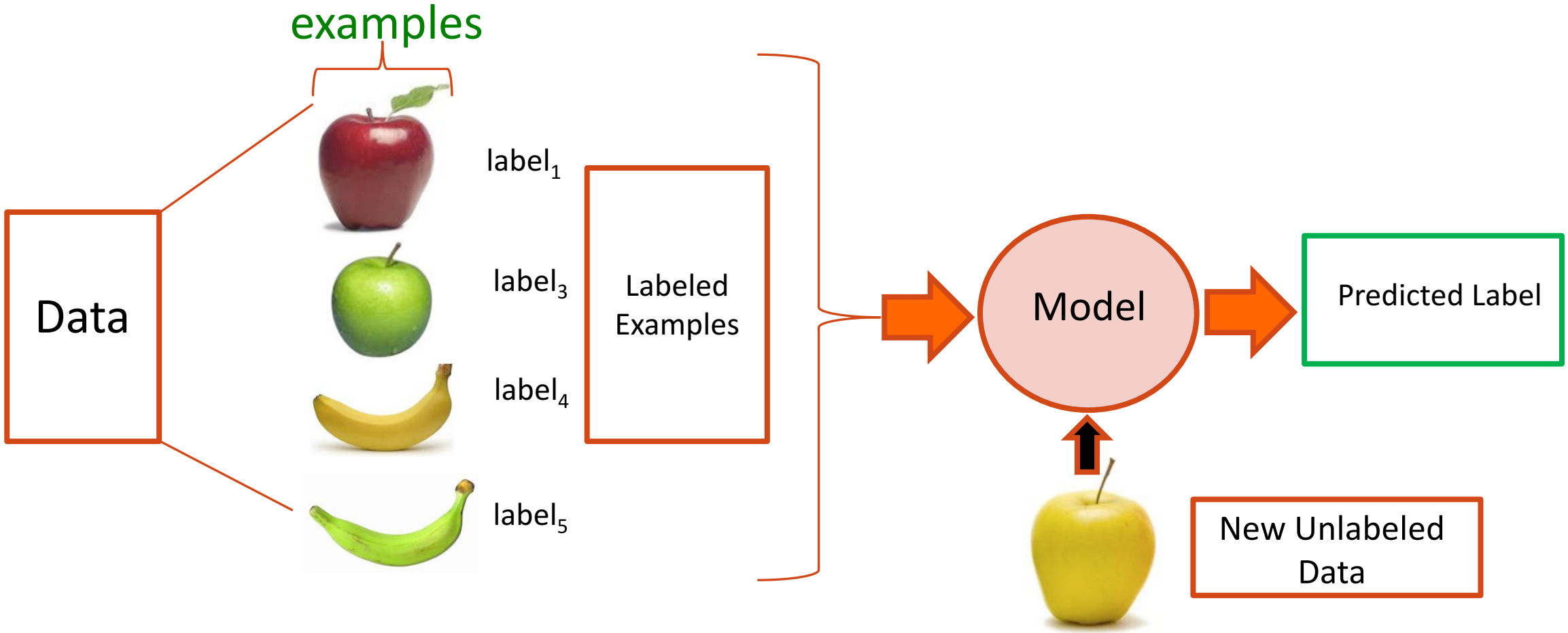
Regression **vs** **Classification**



Classification - Types



Classification



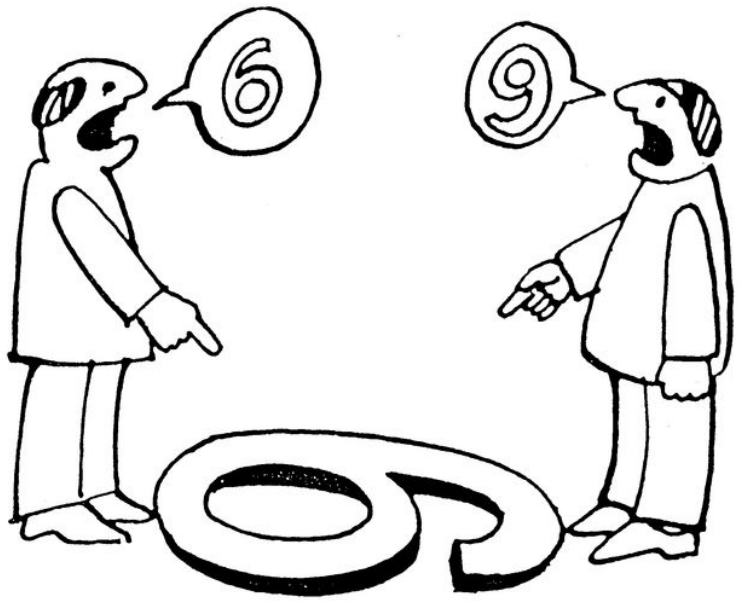
Classification – Real Life Examples

Hospital Emergency Room measures blood pressure, age, history of illness etc. of newly admitted patients. A decision needs to made regarding admitting the patients into ICU. They want to admit high risk patients. Problem is how to discriminate between high risk and low risk patients?

A Bank receives hundreds and thousands of loan applications with information about salary, age, marital status, other loans, credit history, employment status etc. The bank wants to make loans to those people who are most likely going to repay the loan. Problem is how to discriminate between those who are a good credit risk and those who are bad credit risk?

Classification – Other Examples

- Face/Speech Recognition
- Character Recognition
- Spam Detection
- Customer Segmentation
- Recommendation Systems

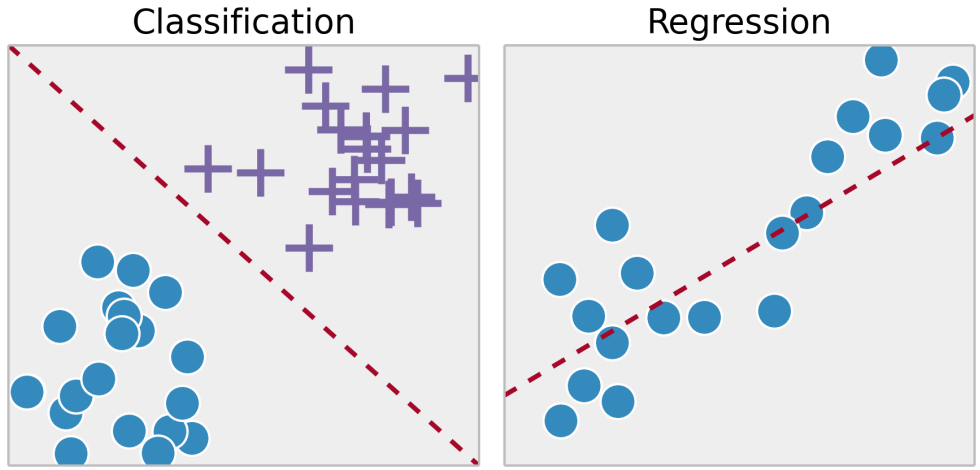


Classification – How?

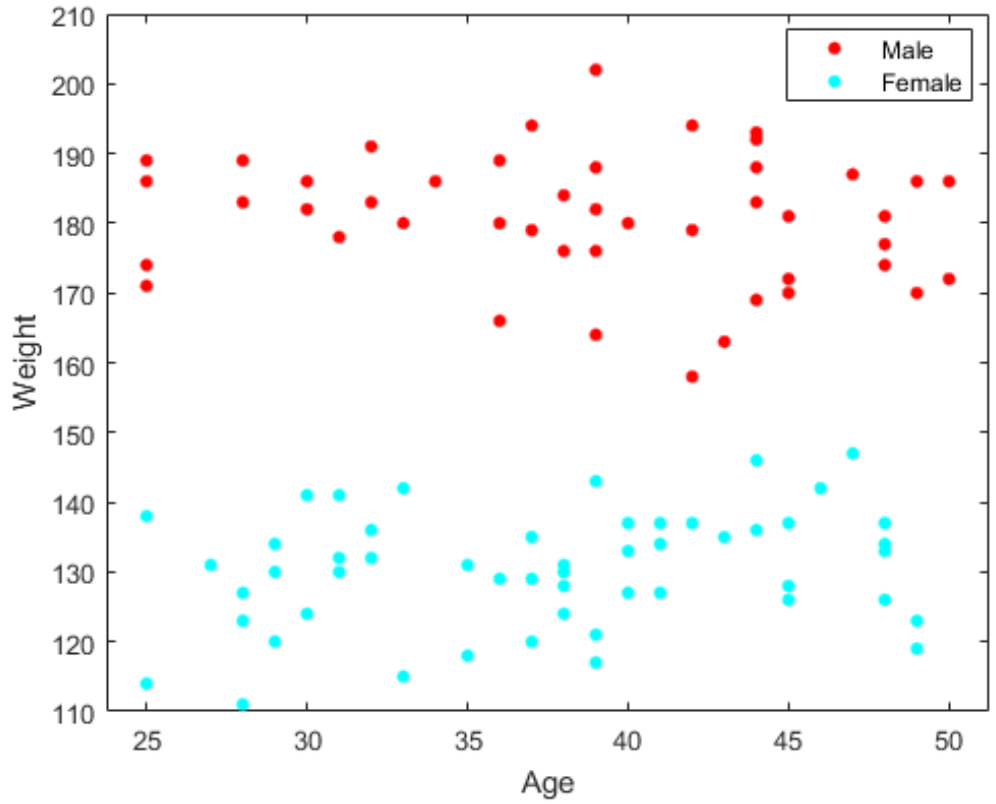
Use Machine Learning Algorithms – Duh!!

In this session we will learn about the following algorithms

- Logistic Regression (Binomial)
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- K- Nearest Neighbors (KNN)



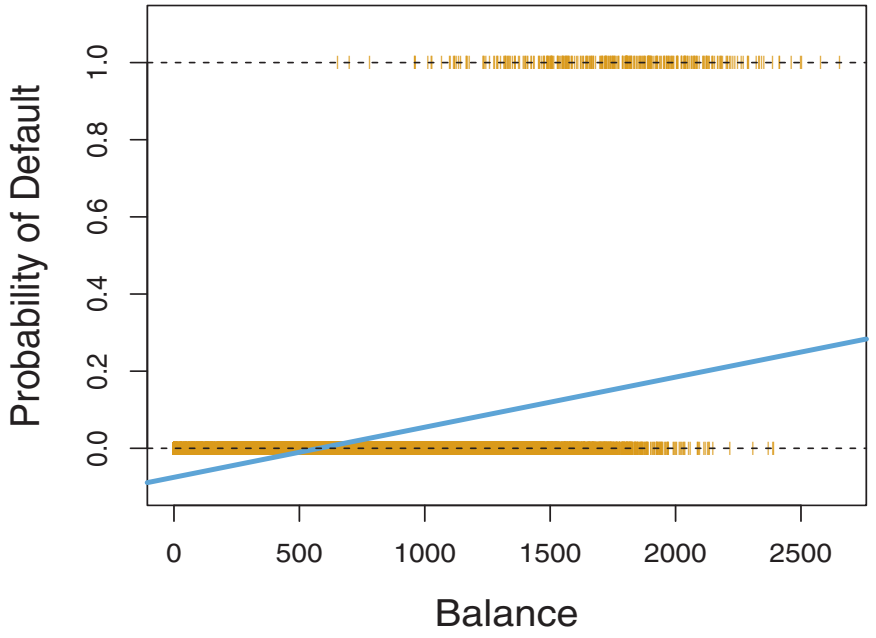
Logistic Regression- Binomial

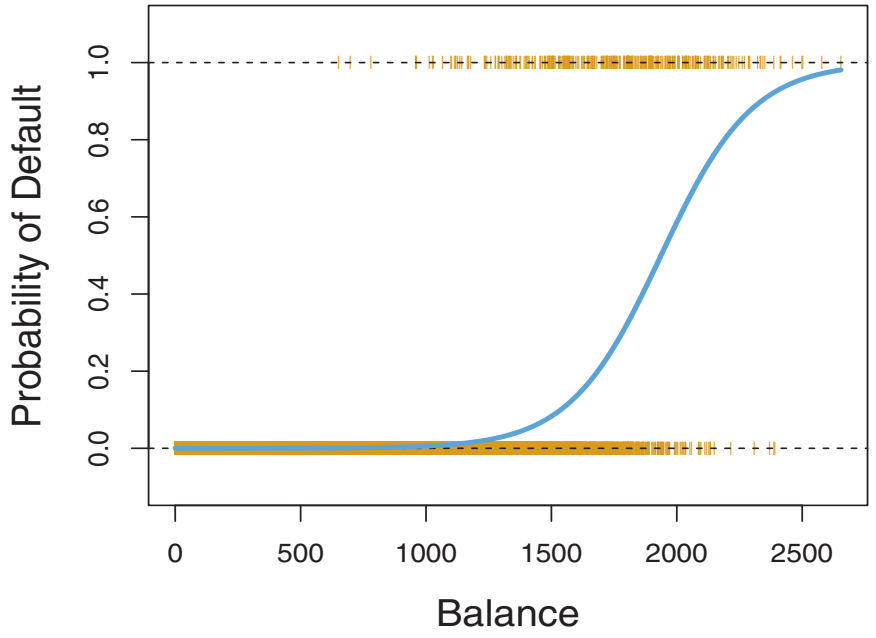


Can you predict whether a person is Male or Female based on this Age and Weight Data?

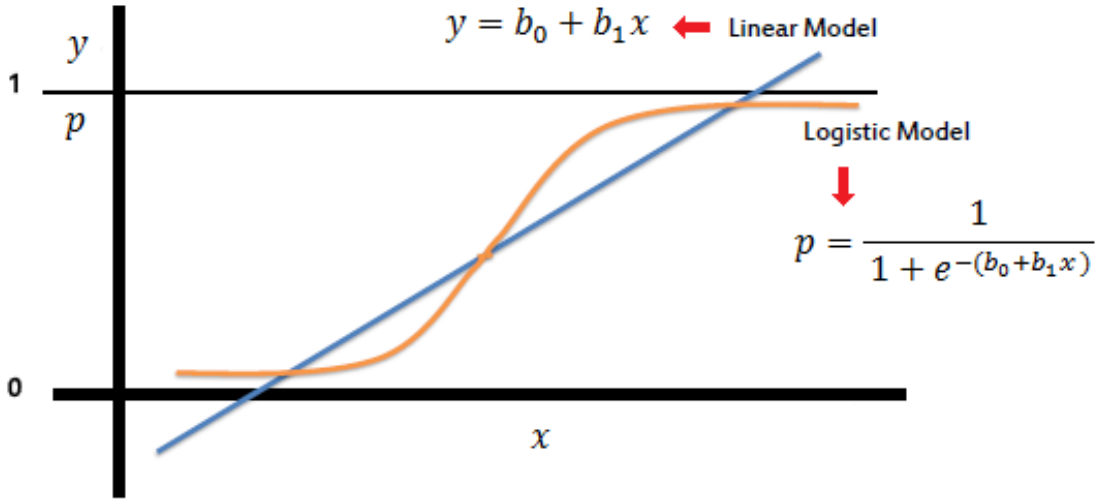
Logistic Regression vs Linear Regression

Loan Default Probability



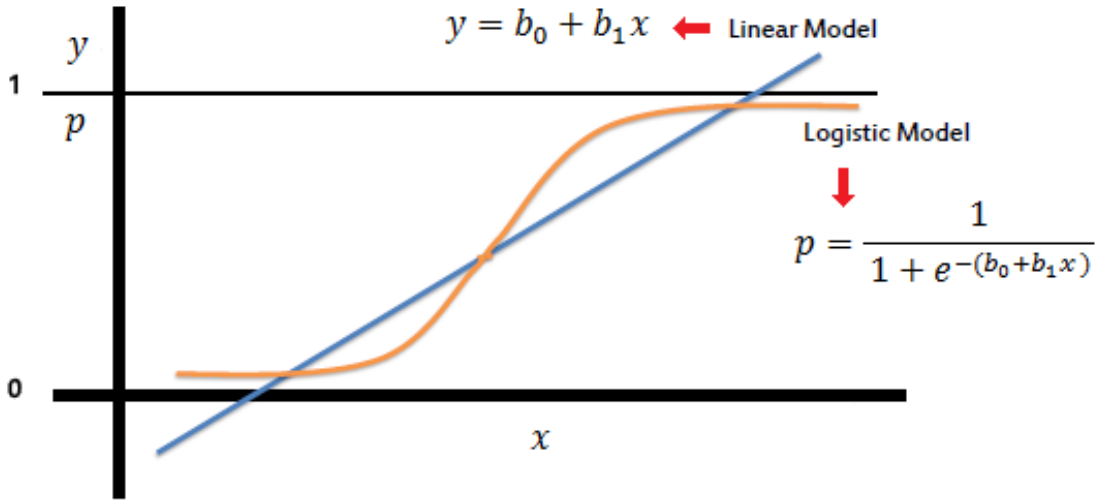


Logistic Regression- Binomial



We can estimate probability of Male or Female using a Linear Regression Model too, but don't you think Logistic Model fits better?

Logistic Regression- Model



$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

Turns out we can use similar method we used in Linear Regression but instead of RMSE, we use Maximum Likelihood to get a good fit!

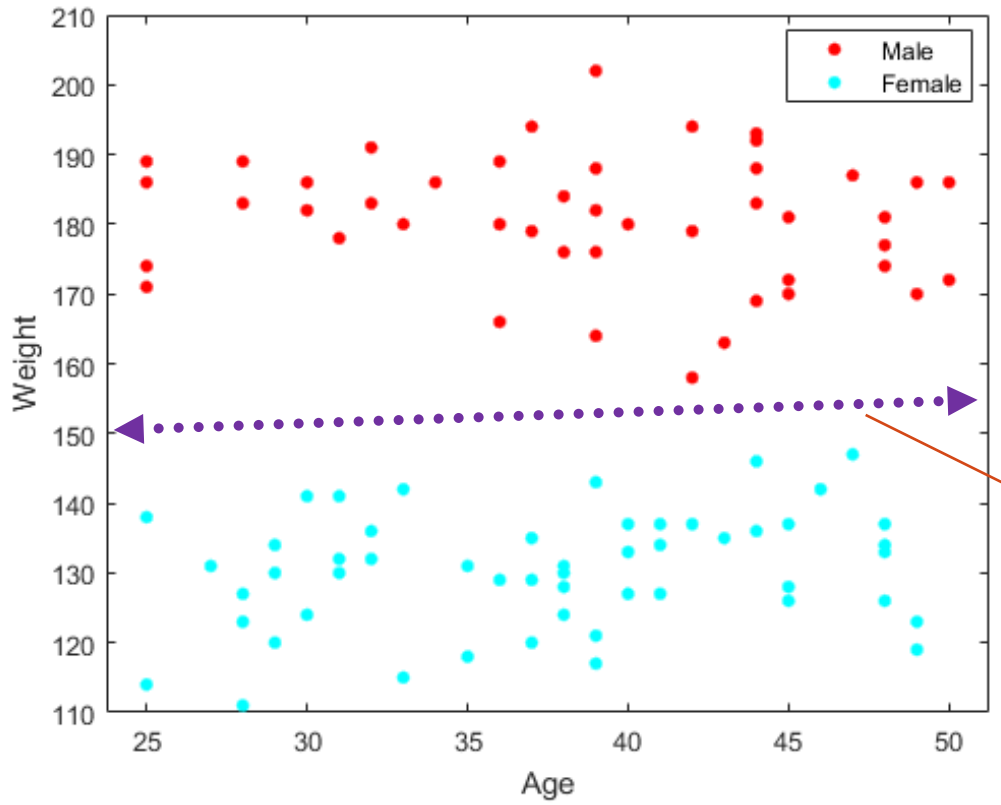
Maximum Likelihood – *Predictions as Close to the Truth as possible*

Logistic Regression – Multiple Classes

How do we use Logistic Regression when we have more than two classes ?

It is possible but there are better ways

Linear Discriminant Analysis (LDA)



Can you predict whether a person is Male or Female based on this Age and Weight Data?

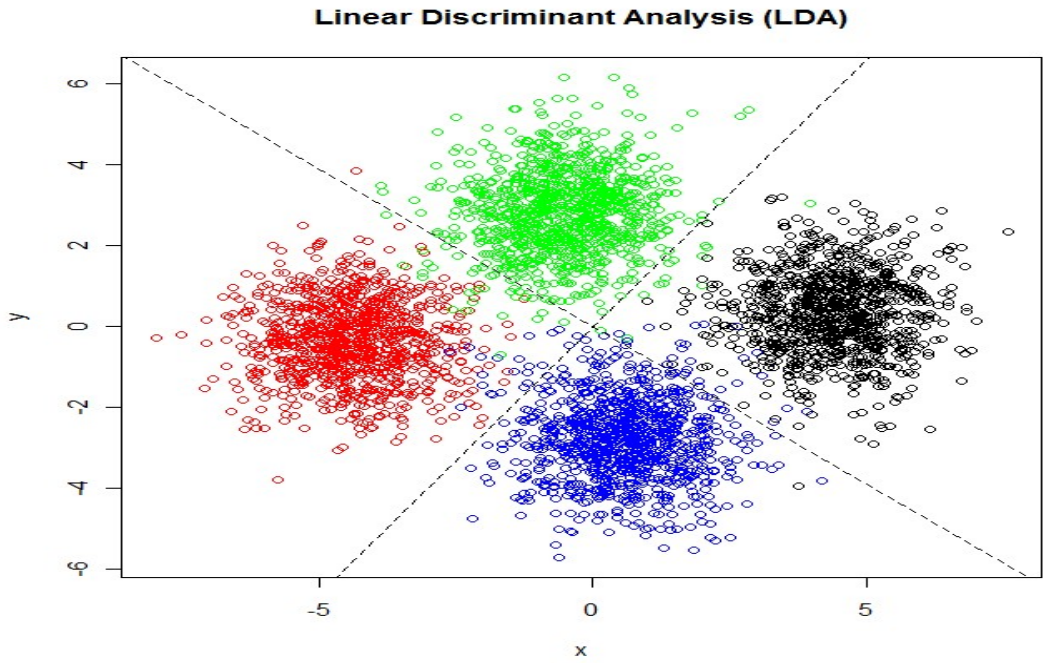
Linear Decision Boundary

Linear Discriminant Analysis (LDA)

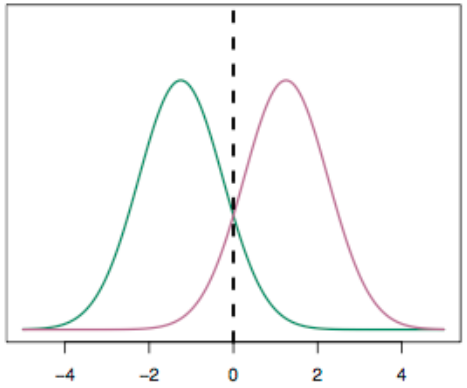
How do we use LDA when we have more than two classes ?

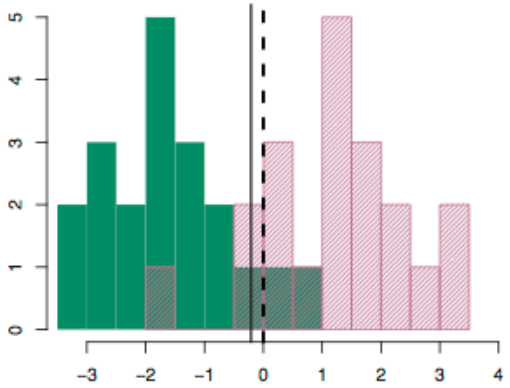
Linear Discriminant Analysis (LDA)

How do we use LDA when we have more than two classes ?



Linear Discriminant Analysis (LDA)

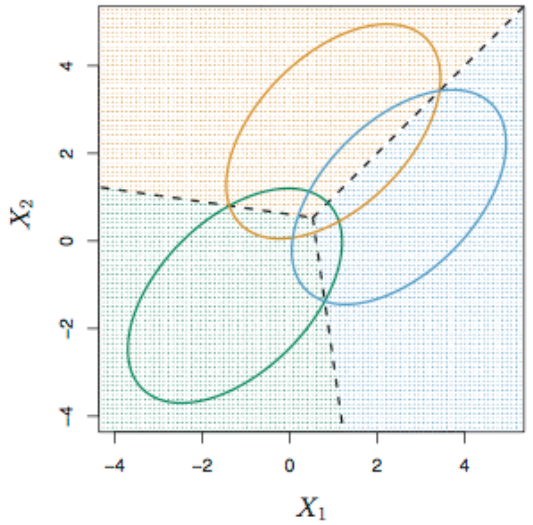


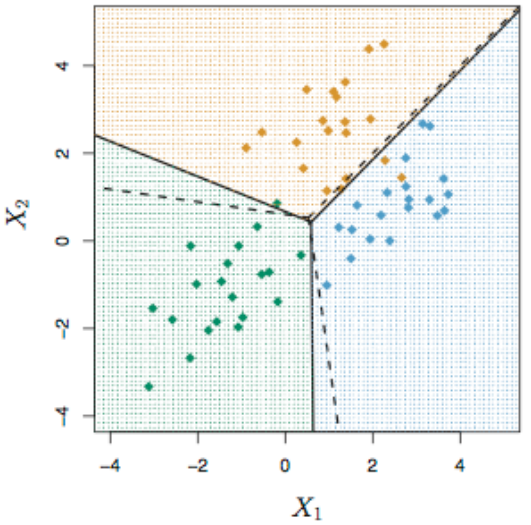


The main assumption we make here is that each of the classes are normally distributed but have different means.

Confusion Matrix

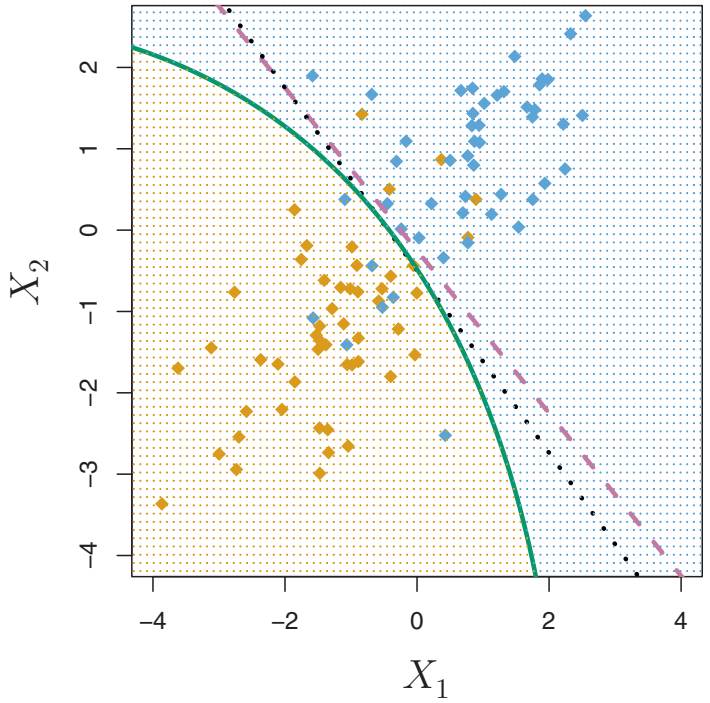
		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

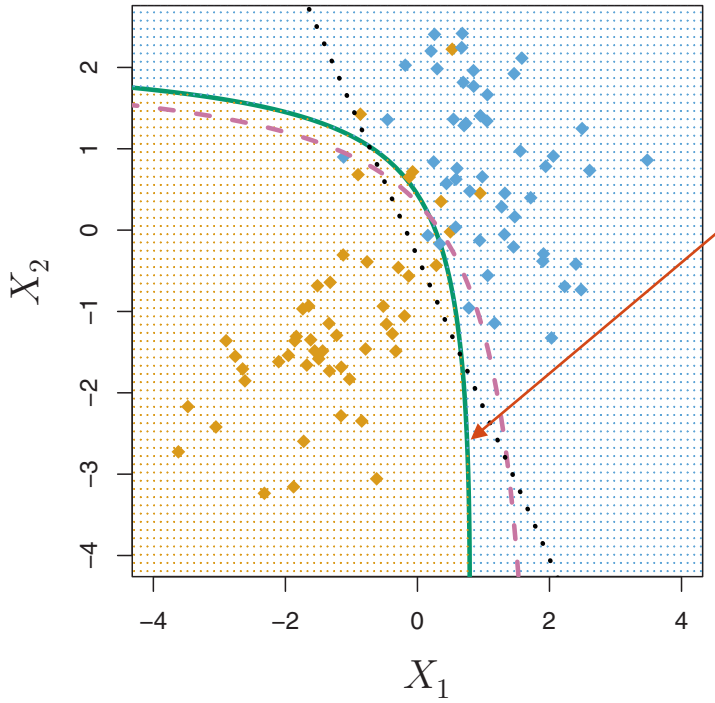




Quadratic Discriminant Analysis

What if the data is such that a straight line isn't the best way to separate the data?





More flexible boundary

QDA is generally better when there is more data.

QDA can be prone to over-fitting for small data.

Quadratic Discriminant Analysis

What is the difference between LDA/QDA and Logistic Regression ?

Logistic Regression

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 x.$$

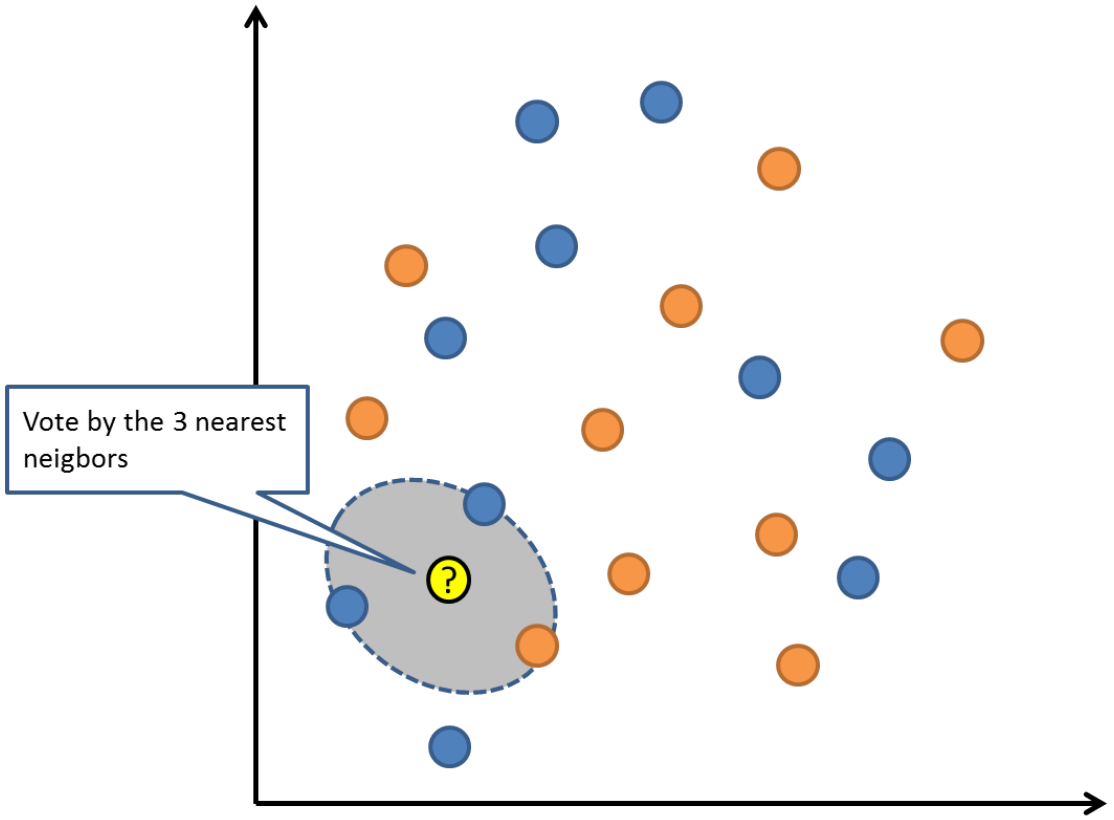
Model parameters are computed by Maximizing Likelihood

LDA/QDA

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x,$$

Model parameters are computed using mean and variance from the data sample

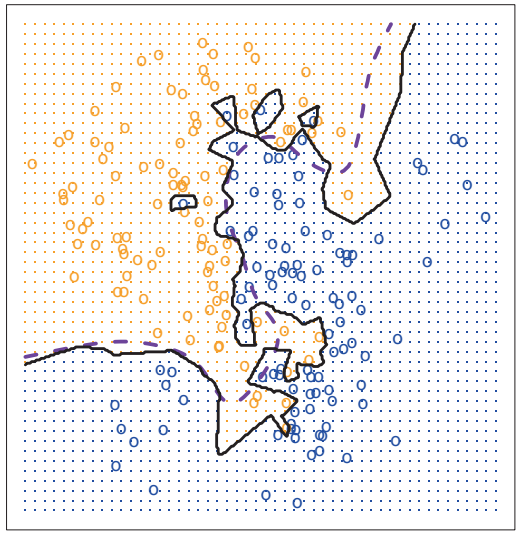
K – Nearest Neighbors (KNN)



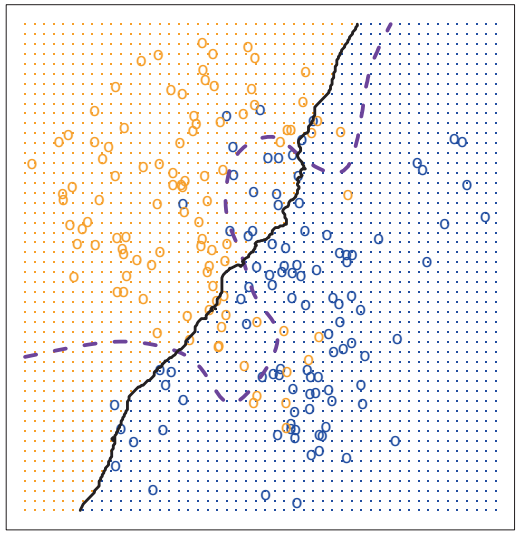
- Non-Parametric approach
- Only input is value of K – how many neighbors?
- Each point is assigned a class based on votes of the K nearest neighbors
- Non-Linear classification boundary
- K better be odd number. Why?
- Various Distance Measures. Euclidean , Manhattan etc.
- Is an age difference of 50 year same as salary difference of $50? Scale and Normalize Data.

K – Nearest Neighbors (KNN)

KNN: K=1

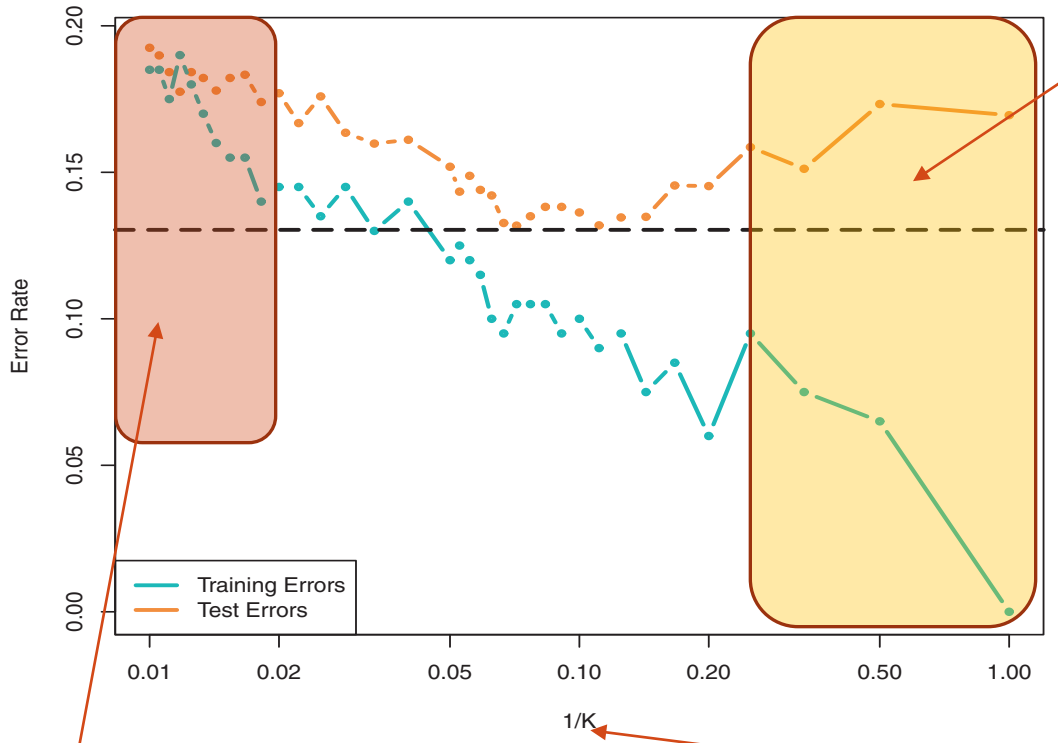


KNN: K=100



- Very Flexible
- Computationally expensive
- Low K leads to over fitting (overly flexible)
- High K leads to under fitting (linear)
- Cross Validation (?) for K

KNN – Bias vs Variance



High Variance

High Bias

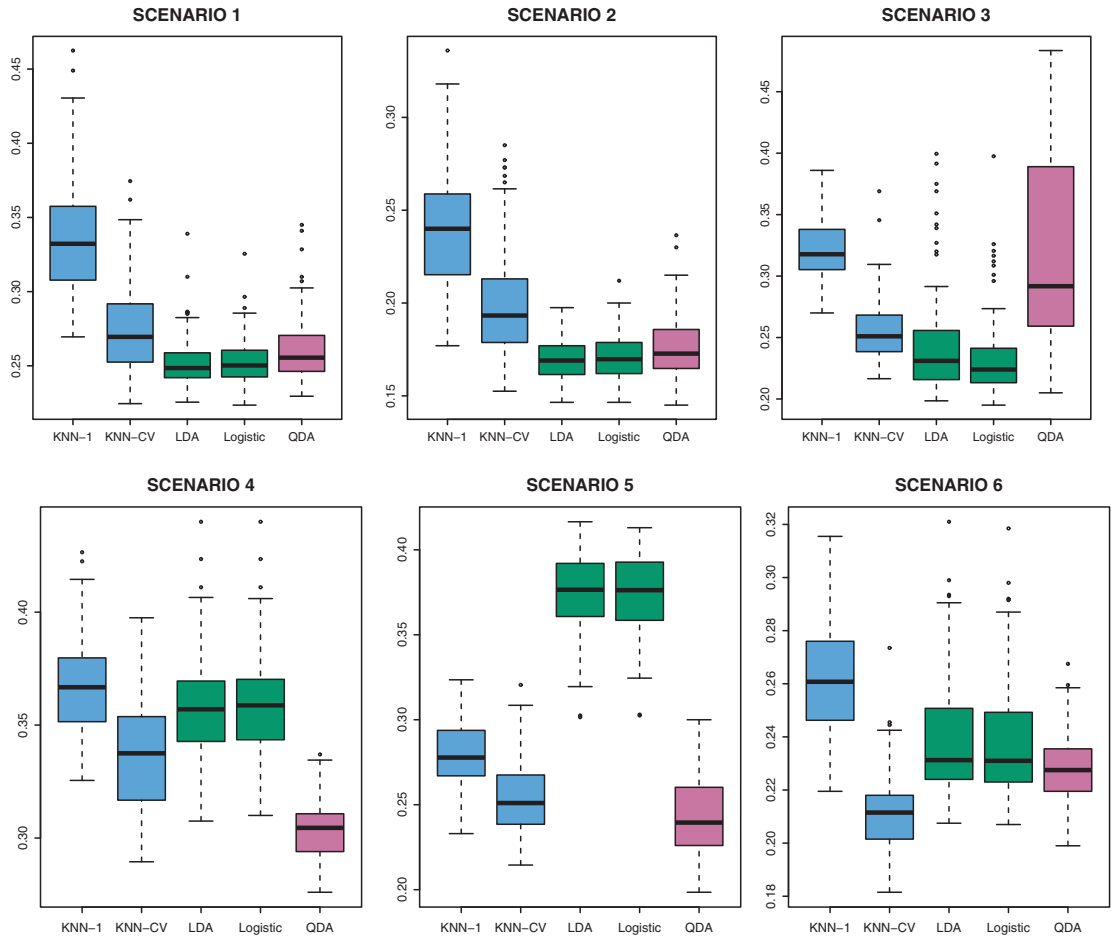
- Test error follows a ‘U’ shape
- As K decreases (more flexible model) , training error decreases, but test error increases.
- Over fit leads to High Variance
- As K increases (more linear model), training error increases, but test error decreases and then increases again.
- Under fit leads to High Bias

1/K increase corresponds to K decreasing from 100 as you move left to right

Which Classifier is better?

Answer

It depends on the data !



Confusion Matrix

What is a Confusion Matrix?

A **confusion matrix** is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.

Confusion Matrix

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

Remember this from LDA discussion?
97.2% Accuracy – not bad huh!!

Why do we need this?

- Is Classification Accuracy alone a good measure ?
- What if there are unequal observations in two classes ?
- What are the types of Errors the model is making ?

Confusion Matrix

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	**Specificity** = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

Classification