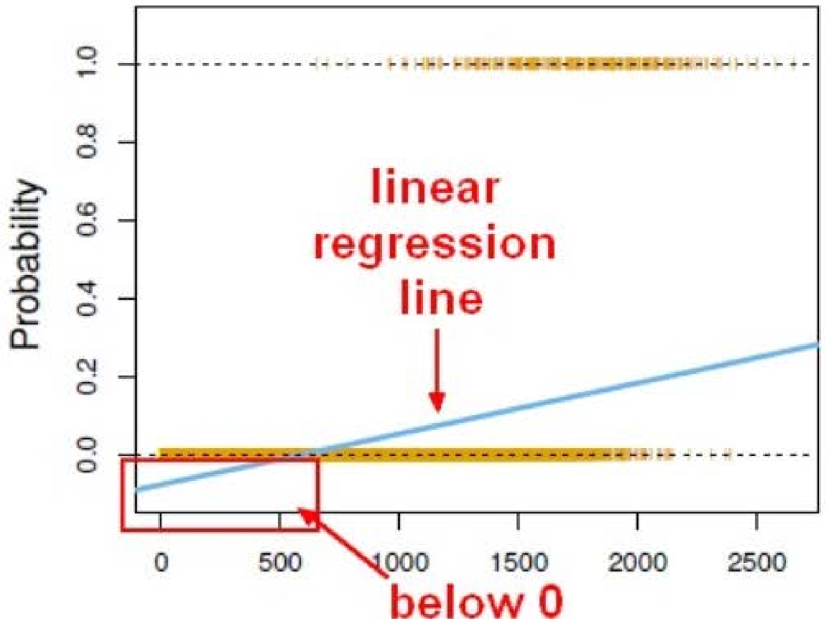
Intro to Logistic Regression

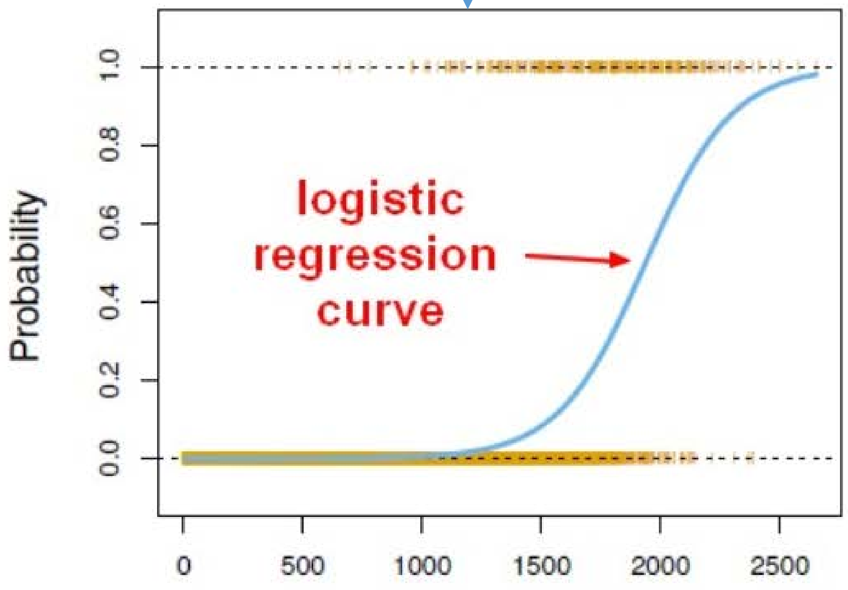
Classification

- **Logistic Regression**
  - A method of classification
  - Binary Classification – Having two classes – 0 or 1
- Regression problems – Predict continuous values
- Classification problems – Predict discrete values

Intro

Transform to a logistic regression curve



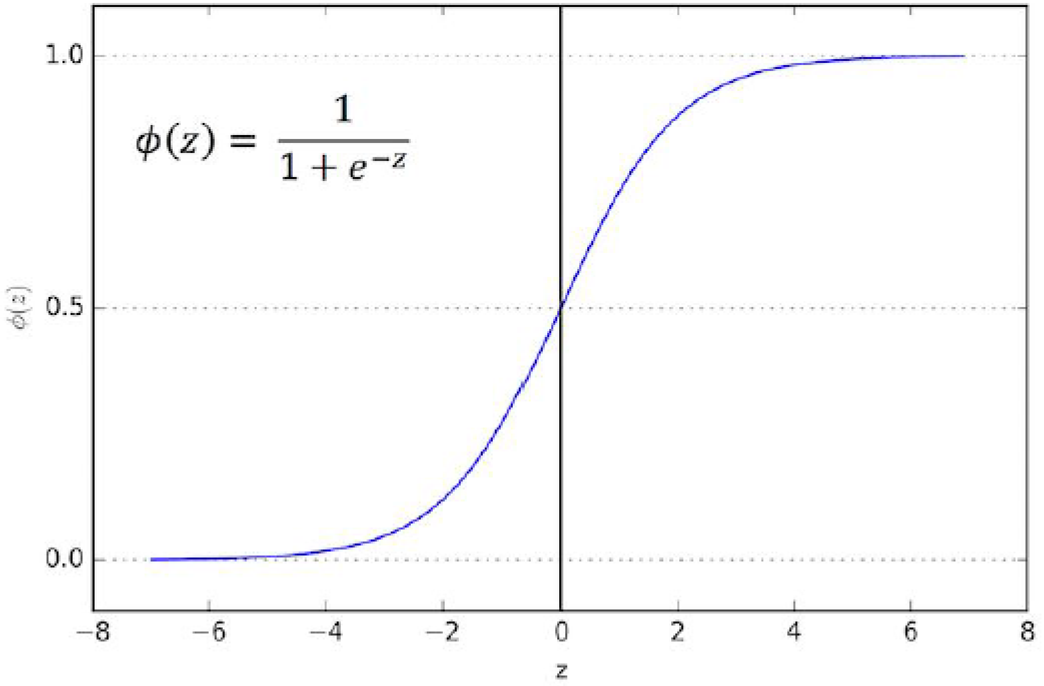


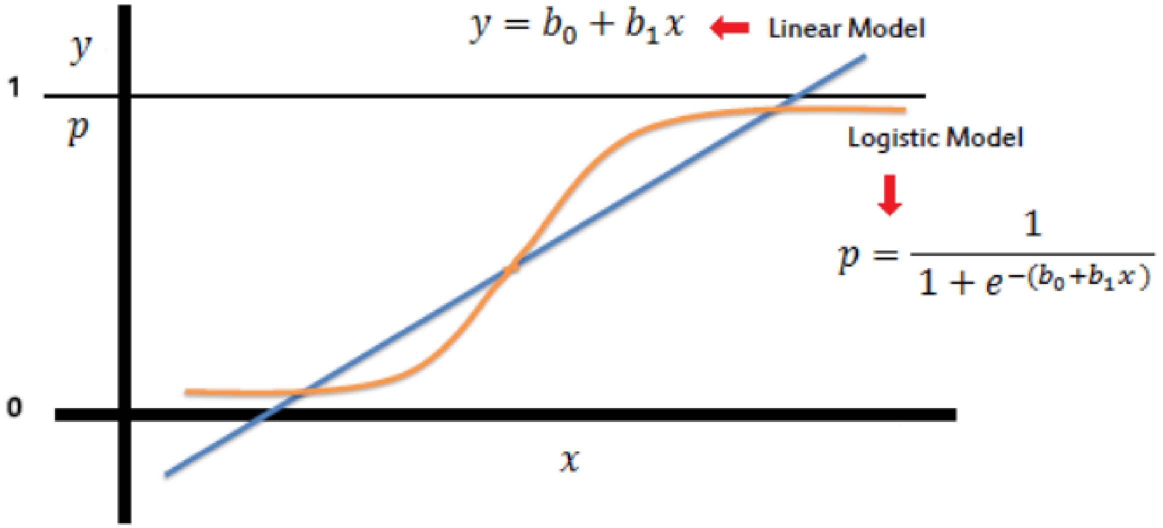
Normal Linear regression

On Binary groups

Sigmoid Function

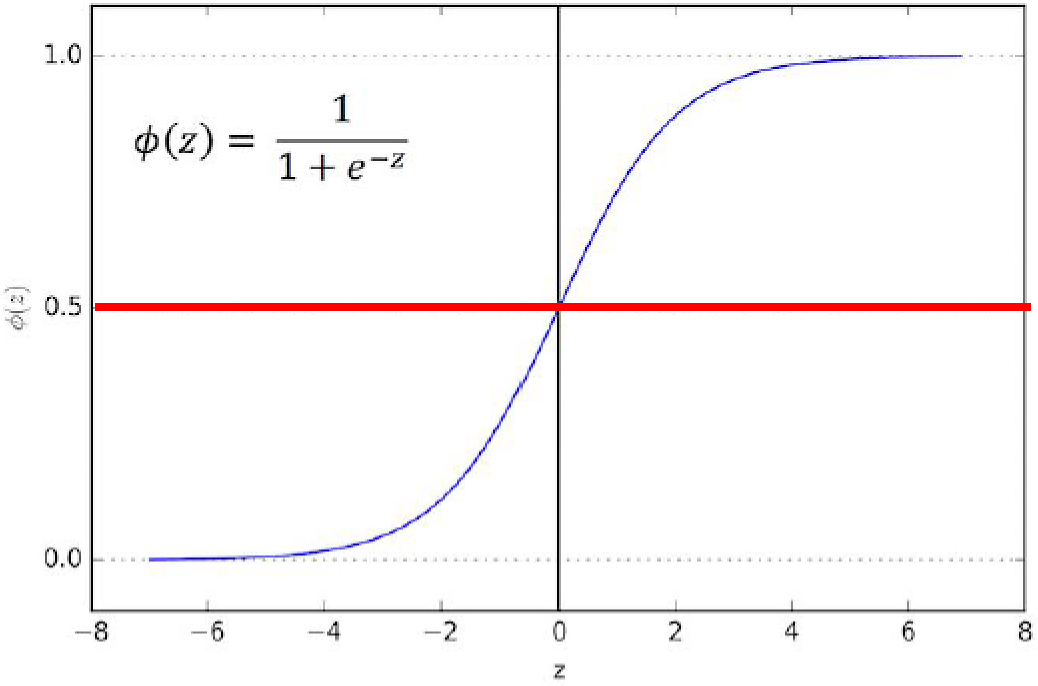
- The Sigmoid (aka Logistic) Function takes in any value and outputs it to be between 0 and 1





Binary Classifier

- This results are in form of probability
- Defined always between 0 to 1
- We can set a cutoff point at 0.5, anything below it results in class 0, anything above is class 1



Evaluate

- CONFUSION MATRIX : Used to evaluate classification problems
- *“The confusion matrix shows the ways in which your classification model is confused when it makes predictions*.”

		Actual class		
		Cat	Dog	Rabbit
Predicted class	Cat	5	2	0
	Dog	3	3	2
	Rabbit	0	1	11

Confusion Matrix

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Example: Test for presence of disease
NO = negative test = False = 0
YES = positive test = True = 1

	Prediction	
Actual	0	1
0	TN	FP
1	FN	TP

“**true positive**” for correctly predicted event values.
“**false positive**” for incorrectly predicted event values.

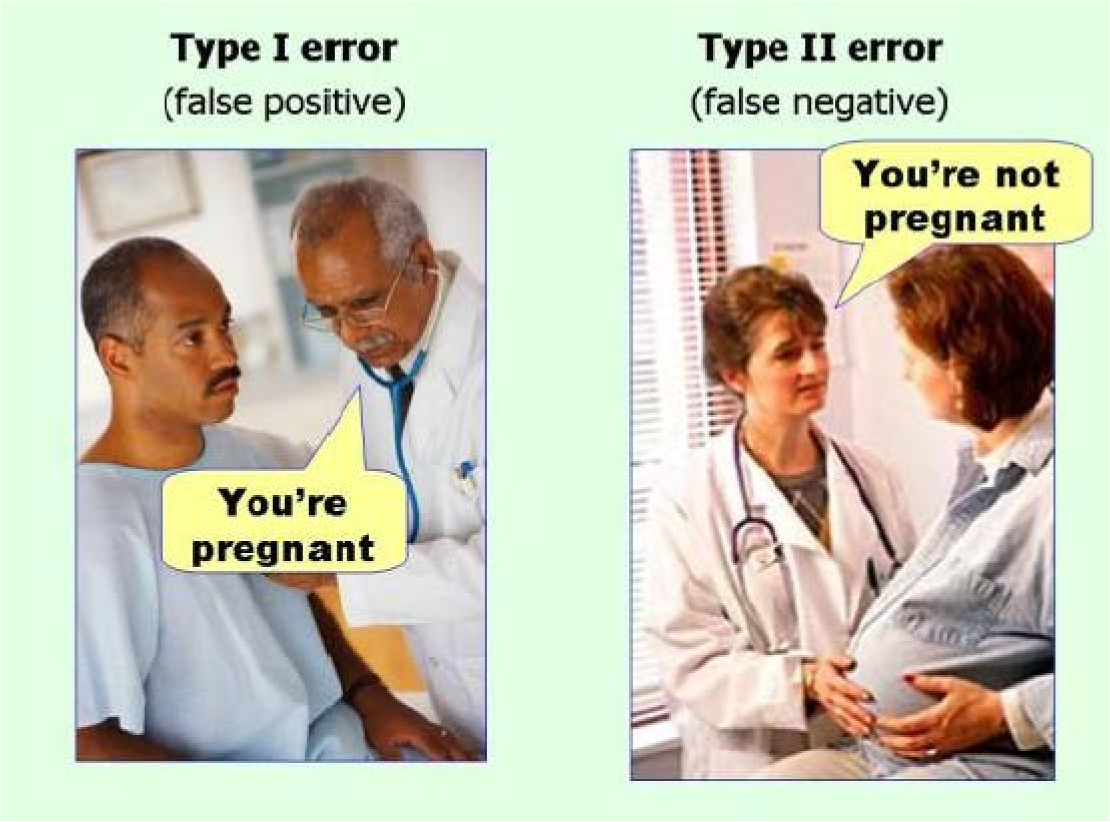
“**true negative**” for correctly predicted no-event values.
“**false negative**” for incorrectly predicted no-event values.

Confusion Matrix

Type 1 error

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Type 2 error



Precision , Recall , F1

Precision: When it predicts yes, how often is it correct?

TP/predicted yes = 100/110 = 0.91

True Positive Rate: When it's actually yes, how often does it predict yes?

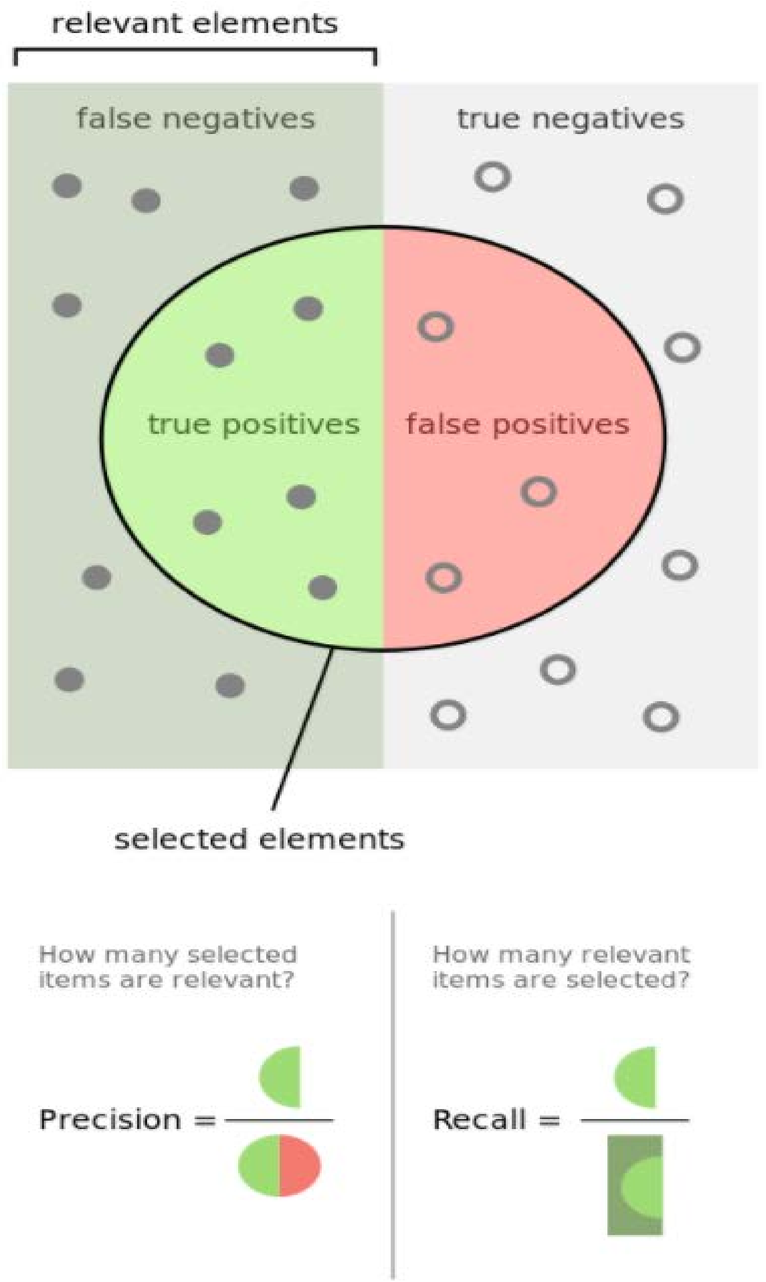
TP/actual yes = 100/105 = 0.95

also known as "Sensitivity" or "Recall"

F1 score: Harmonic average of the precision and recall.

1 : Best value at 1 (perfect precision and recall)

0 : Worst at 0.



Confusion Matrix - Terminology

		Predicted	
		YES	NO
Actual	YES	TP	FN Type 1 error
	NO	FP Type 2 error	TN

Sensitivity (recall)
TP/ TP + FN

False Negative rate
FN/ TP + FN

False Positive rate
FP/ FP + TN

Specificity
TN/ FP + TN

Precision
TP/ TP + FP

False omission rate
FN/ FN + TN

False discovery rate
FP/ TP + FP

Negative predictive value
TN/ FN + TN

Accuracy
TP + TN/ Total

Error Rate
(FP + FN)/Total

F1 Score
2TP / (2TP + FP + FN)

$$F_1 = \frac{2}{\frac{1}{\mathrm{recall}} + \frac{1}{\mathrm{precision}}}$$

Accuracy & Specificity

Accuracy: Overall, how often is the classifier correct?

(TP+TN)/total = (100+50)/165 = 0.91

Error Rate: Overall, how often is it wrong?

(FP+FN)/total = (10+5)/165 = 0.09

equivalent to 1 minus Accuracy

False Positive Rate: When it's actually no, how often does it predict yes?

FP/actual no = 10/60 = 0.17

Specificity: When it's actually no, how often does it predict no?

- TN/actual no = 50/60 = 0.83
- equivalent to 1 minus False Positive Rate

Confusion Matrix - Terminology

		Predicted: YES	Predicted: NO
Actual	YES	TP = 100	FN = 5 Type 2 error
	NO	FP = 10 Type 1 error	TN = 50

Sensitivity (recall)
100/ 105 = 0.95

False Negative rate
5/105 = 0.05

False Positive rate
10/ 60 = 0.17

Specificity
50/60 = 0.83

Precision
100/ 110= 0.91

False omission rate
5/55 = 0.09

False discovery rate
10/110 = 0.09

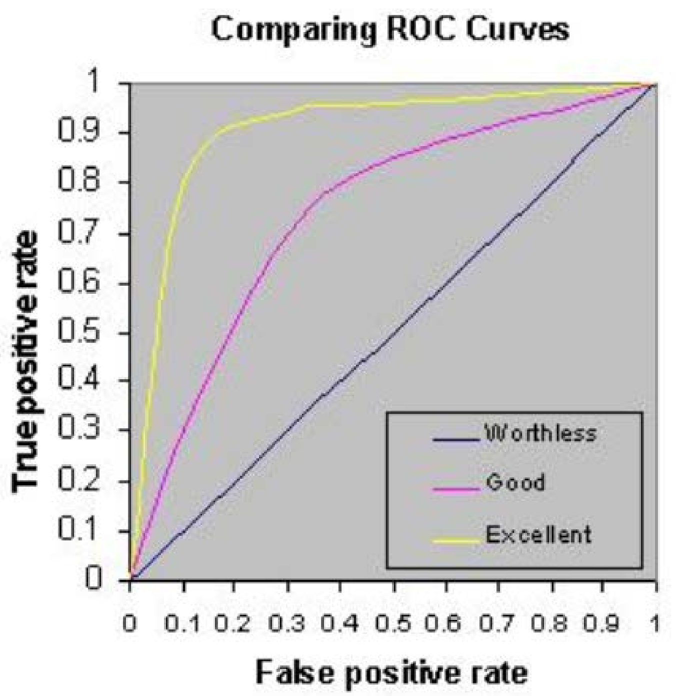
Negative predictive value
50/55 = 0.91

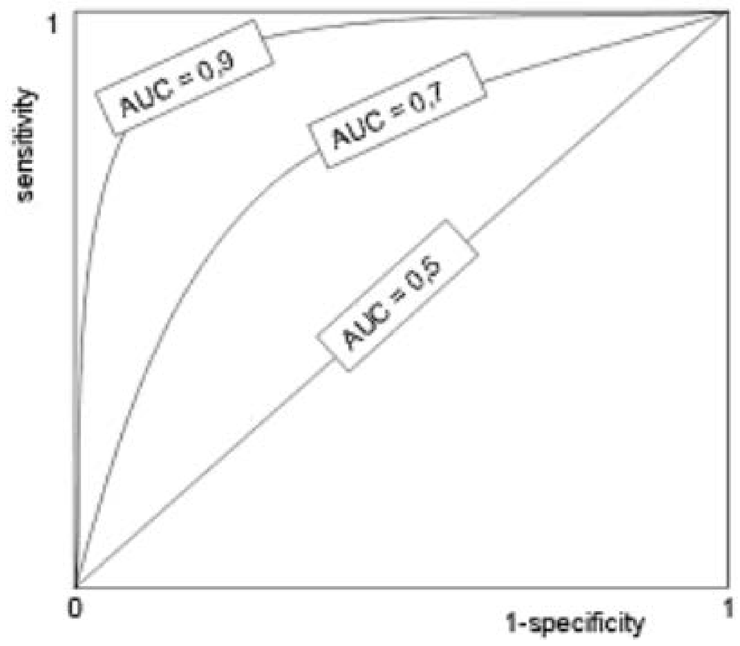
Accuracy
100+50/ (100+5+10+50) = 0.91

Error Rate
1-0.91 = 0.09

F1 Score
2*100 / (2*100 + 10 + 5) = 0.93

ROC and AUC





ROC curve - visualize the performance of a binary classifier

Receiver operating characteristic

AUC - Summarize its performance in a single number

More AUC better the model.

Refer - http://www.dataschool.io/roc-curves-and-auc-explained/

Logistic Regression Examples

1. Spam versus “Ham” emails
2. Loan Default (yes/no)
3. Disease Diagnosis