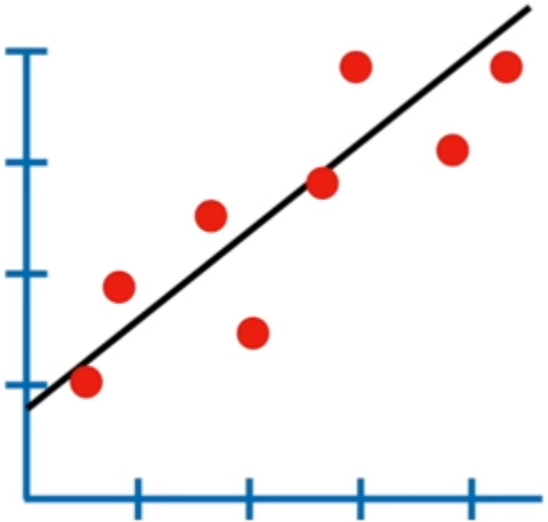
Logistic Regression

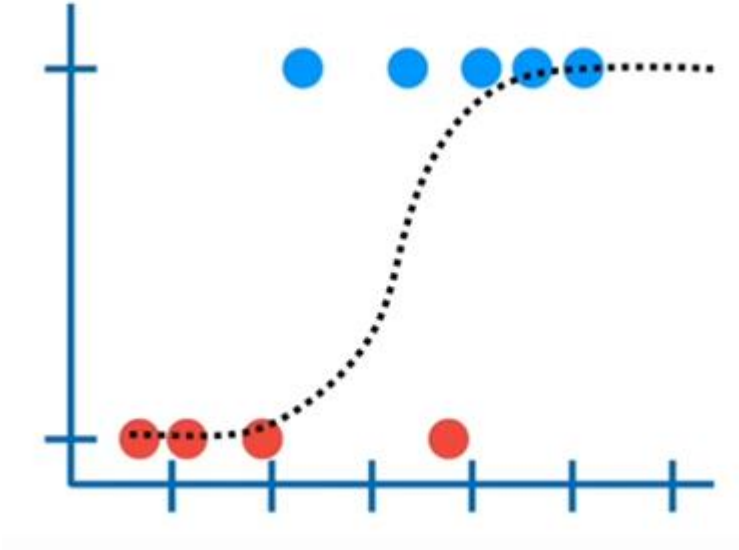
What type of target data was in Linear Regression?

- Continues
- What if discrete?
- We want to classifying something.

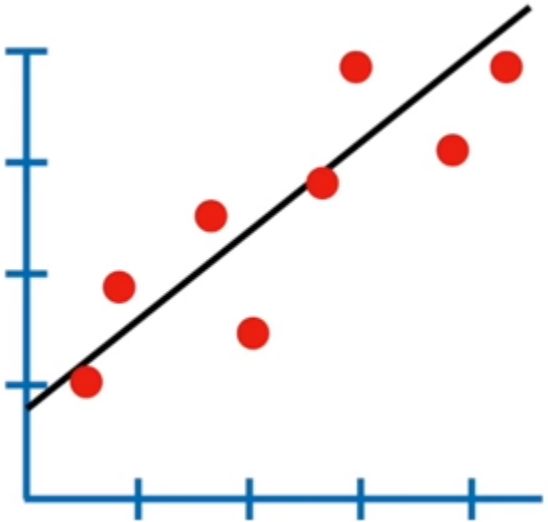
What type of target data was in Linear Regression?

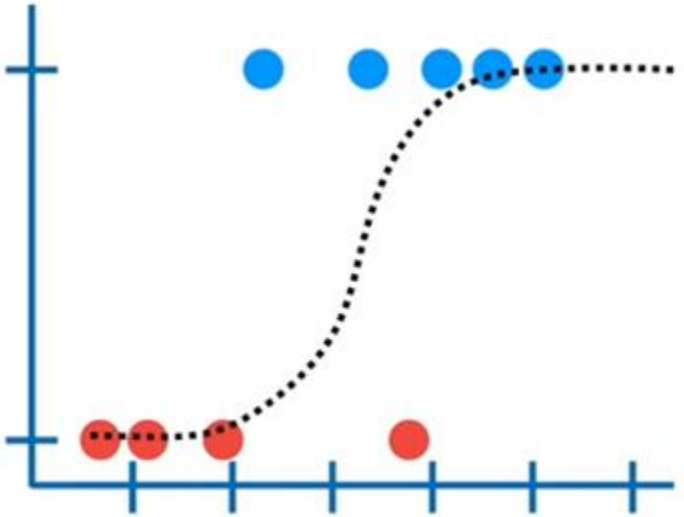


What if ?

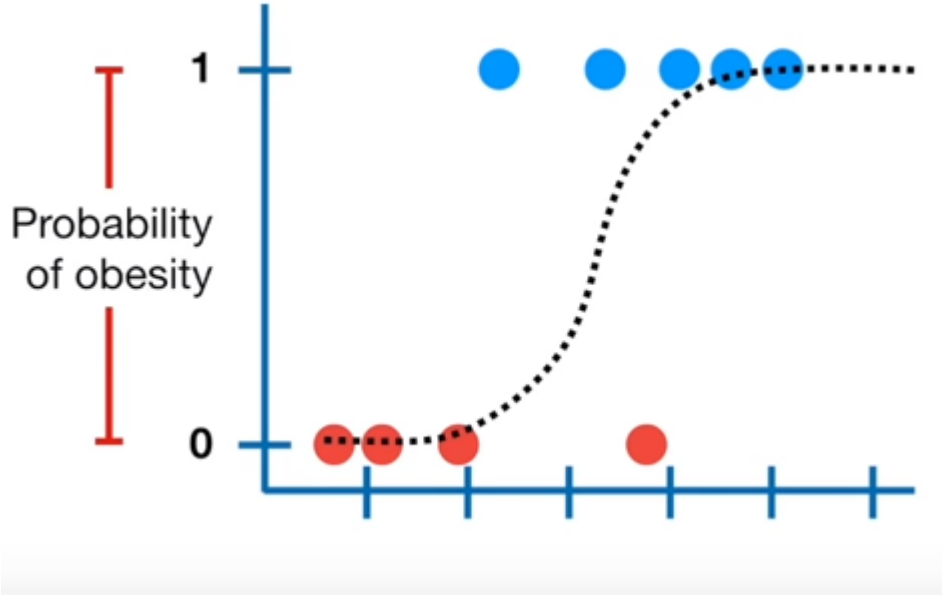


The Goal of Logistic regression





The Goal of Logistic regression



Classification problems

- Email - spam/not spam?
- Online transactions - fraudulent?
- Tumour - Malignant/benign
- Gaming - Win vs Loss
- Sales - Buying vs Not buying
- Marketing – Response vs No Response
- Credit card & Loans – Default vs Non Default
- Operations – Attrition vs Retention
- Websites – Click vs No click
- Fraud identification –Fraud vs Non Frau
- Healthcare –Cure vs No Cure

Logistic Regression

- Name is somewhat misleading.
- It is technique for classification, not regression.
- “Regression” comes from fact that we fit a linear model to the feature space.
- Involves a more **probabilistic** view of classification.

Learn from what we know.

- We would like to use something like what we know from linear regression:
- Continuous outcome = $\beta 0 + \beta 1X1 + \beta 2X2+\ldots$
- How can we turn a proportion into a continuous outcome?

Transformation of linear to logistic

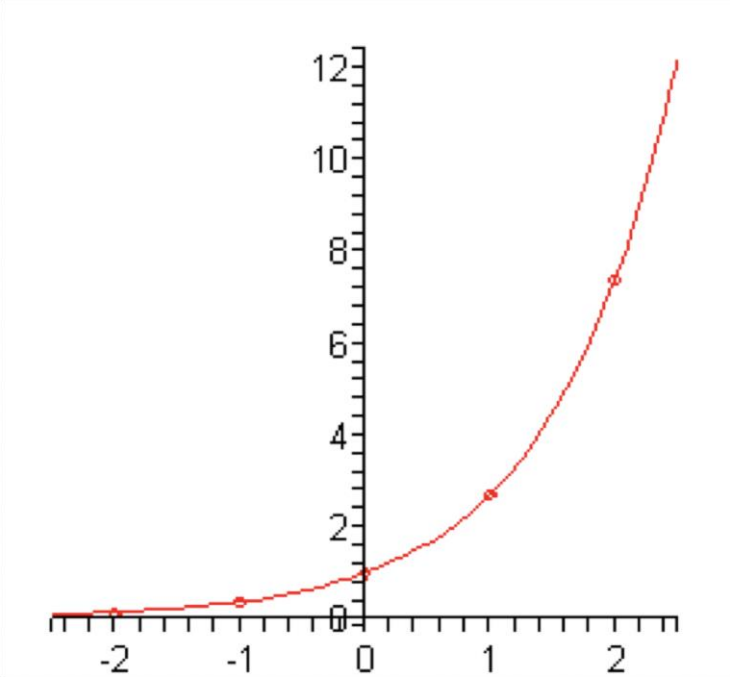
- Input : $(-\infty,+\infty)$

- Exponential
- x -2 -1 0 1 2
- f(x) 0.1353... 0.3679... 1 2.718... 7.389
- The Exponential will convert data in the range of $(0,'+\infty)$
- Any number divided by the number +1

p = 65432

p/p+1 = 0.99

Output : (0, 1)



Transforming a proportion

- A proportion is a value between 0 and 1
- The odds are always positive:

$$\text{odds} = \left(\frac{p}{1-p}\right) \Rightarrow [0, +\infty)$$

- The log odds is continuous:

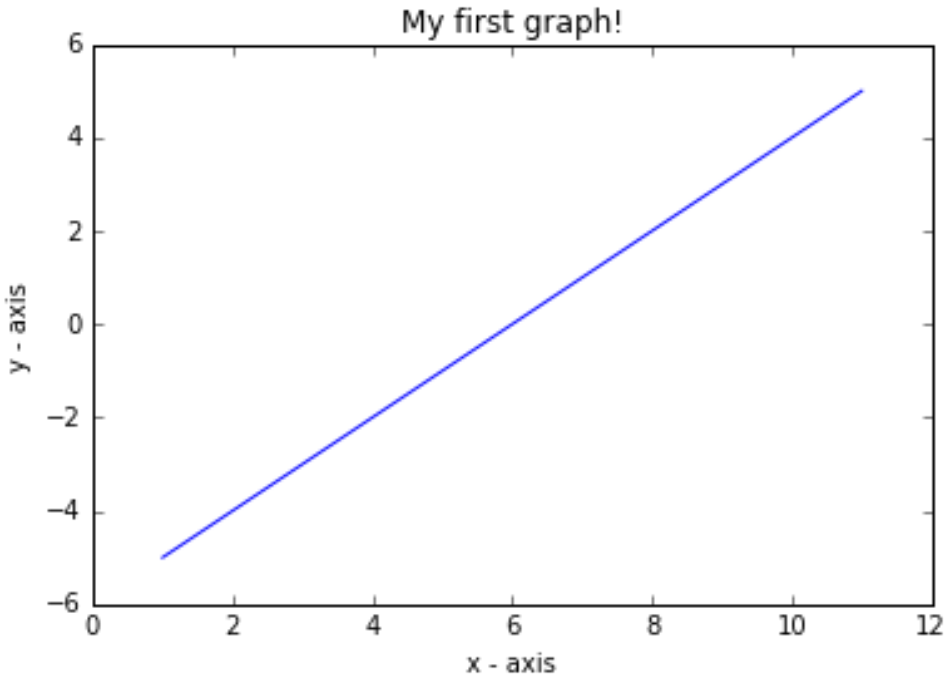
$$\text{Logodds} = \ln\left(\frac{p}{1-p}\right) \Rightarrow (-\infty, +\infty)$$

“Logit” transformation of the probability

Measure	Min	Max	Name
$\Pr(Y = 1)$	0	1	“probability”
$\frac{\Pr(Y=1)}{1-\Pr(Y=1)}$	0	∞	“odds”
$\log\left(\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right)$	-∞	∞	“log-odds” or “logit”

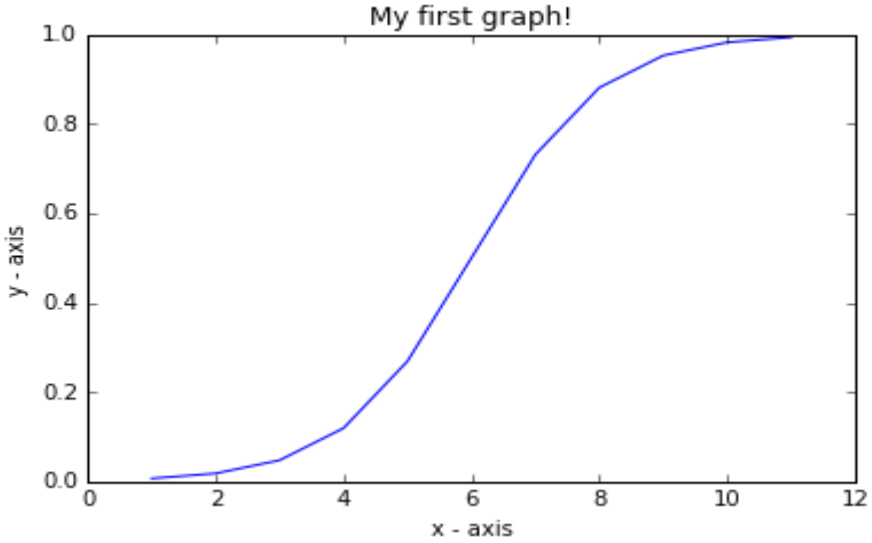
Regression line

x	Y
1	-5
2	-4
3	-3
4	-2
5	-1
6	0
7	1
8	2
9	3
10	4
11	5

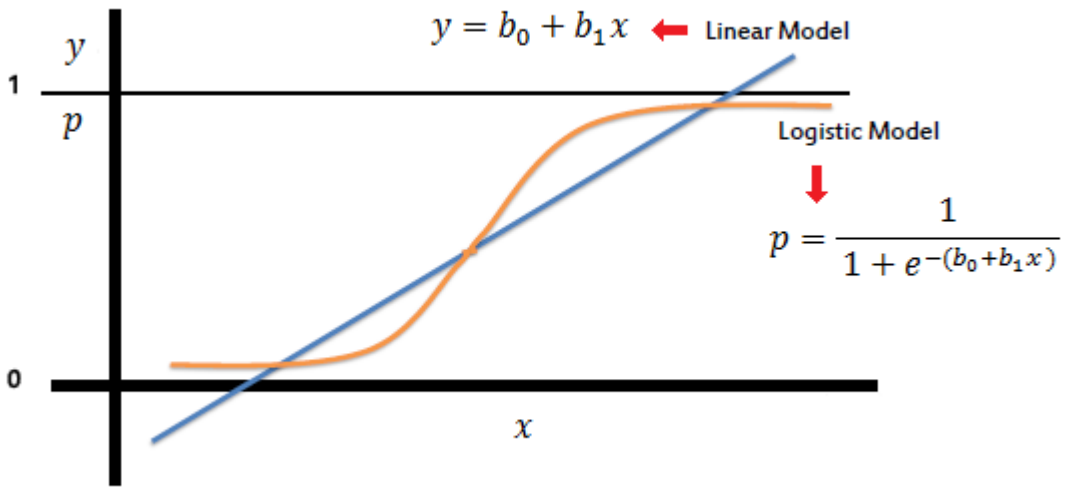


Transformation to Classification

x	Sigmoid(Y)
1	0.006692850924
2	0.01798620996
3	0.04742587318
4	0.119202922
5	0.2689414214
6	0.5
7	0.7310585786
8	0.880797078
9	0.9525741268
10	0.98201379
11	0.9933071491



Logistic Regression Equation



Simple Derivation

g(y) = βo + β(Age)

Since probability must always be positive, we'll put the linear equation in exponential form

To make the probability less than 1, we must divide p by a number greater than p. This can simply be done by:

p = exp(βo + β(Age)) / exp(βo + β(Age)) + 1 = e^(βo + β(Age)) / e^(βo + β(Age)) + 1 ----- (c)

p = e^y/ 1 + e^y --- (d)

q = 1 - p = 1 - (e^y/ 1 + e^y) --- (e)

$$\frac{p}{1-p} = e^{y}$$

$$\log\left(\frac{p}{1-p}\right) = y$$

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta(\text{Age})$$

Learning from Example

- In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There its leaders decided to attempt a new and untested rote to the Sacramento Valley. Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake. The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October. By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed)

Dataset

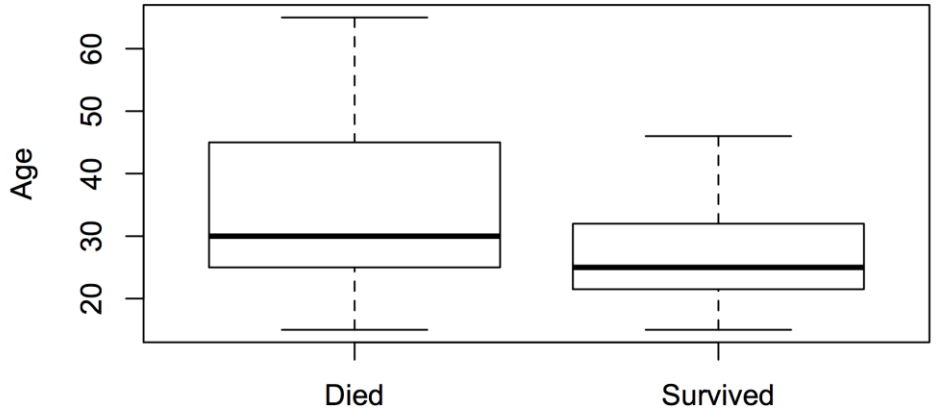
	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Exploratory Analysis

Status vs. Gender:

	Male	Female
Died	20	5
Survived	10	10

Status vs. Age:



Exploratory Analysis

- It seems clear that both age and gender have an effect on someone's survival,
- how do we come up with a model that will let us explore this relationship?

- Even if we set Died to 0 and Survived to 1, this isn't something we can transform our way out of - we need something more.

- One way to think about the problem - we can treat Survived and Died as successes and failures arising from a binomial distribution where the probability of a success is given by a transformation of a linear model of the predictors.

- It turns out that this is a very general way of addressing this type of problem in regression, and the resulting models are called Logistic regression.

- All Logistic regression have the following three characteristics:
- A probability distribution describing the outcome variable
- A linear model

Linear regression $Y = b_0+b_1\times X_1+b_2\times X_2.....+b_K\times X_K$

- A link function that relates the linear model to the parameter of the outcome distribution

Linear regression $Y = b_0+b_1\times X_1+b_2\times X_2.....+b_K\times X_K$

Sigmoid Function $P = \frac{1}{1+e^{-Y}}$

$$\ln\left(\frac{P}{1-P}\right) = b_0+b_1\times X_1+b_2\times X_2.....+b_K\times X_K$$

Model Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.8185	0.9994	1.82	0.0688
Age	-0.0665	0.0322	-2.06	0.0391

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

Odds / Probability of survival for a new-born (Age=0):

Model:

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times \text{Age}$$

$$\log\left(\frac{p}{1-p}\right) = 1.8185 - 0.0665 \times 0$$

$$\frac{p}{1-p} = \exp(1.8185) = 6.16$$

$$p = 6.16/7.16 = 0.86$$

Logistic Regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

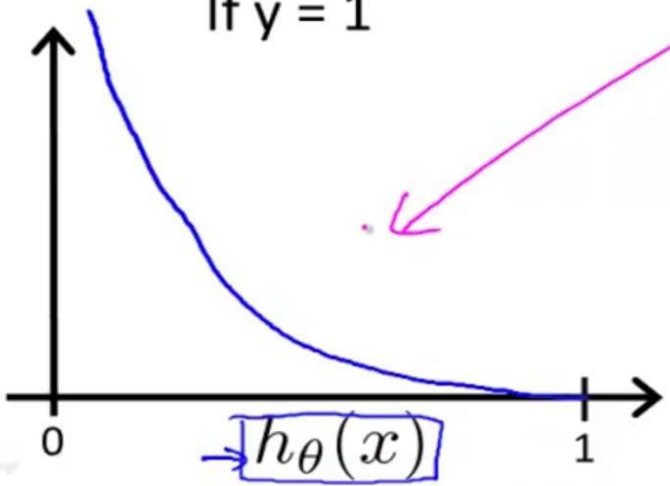
$$\text{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \qquad \text{if y} = 1$$

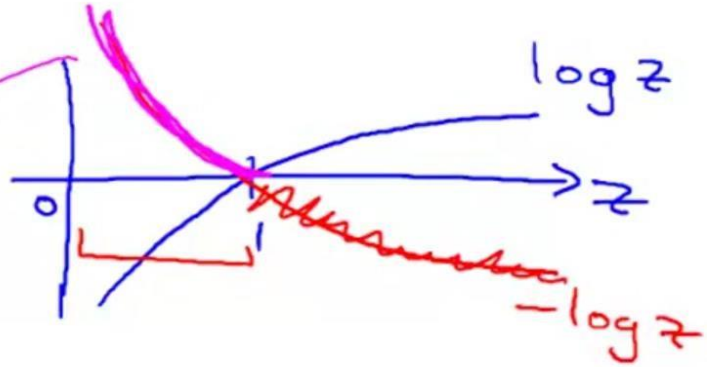
$$\text{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \qquad \text{if y} = 0$$

Cost Function

Logistic regression cost function

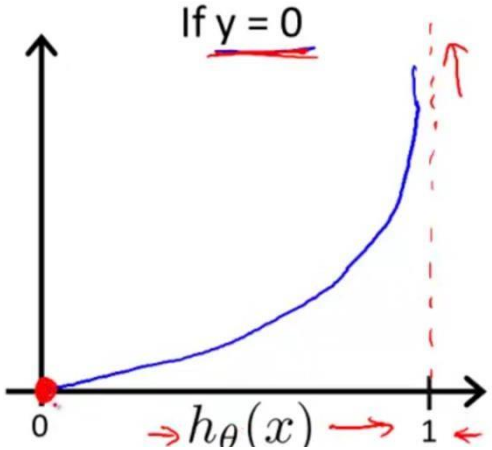
$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$





$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

Cost Function for 0



If our correct answer 'y' is 0, then the cost function will be 0 if our hypothesis function also outputs 0. If our hypothesis approaches 1, then the cost function will approach infinity.

Combining the cost function

Logistic regression cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \text{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

Multiplying by y and $(1 - y)$ in the above equation is a sneaky trick that let's us use the same equation to solve for both y=1 and y=0 cases. If y=0, the first side cancels out. If y=1, the second side cancels out. In both cases we only perform the operation we need to perform.

Metrics for Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix:

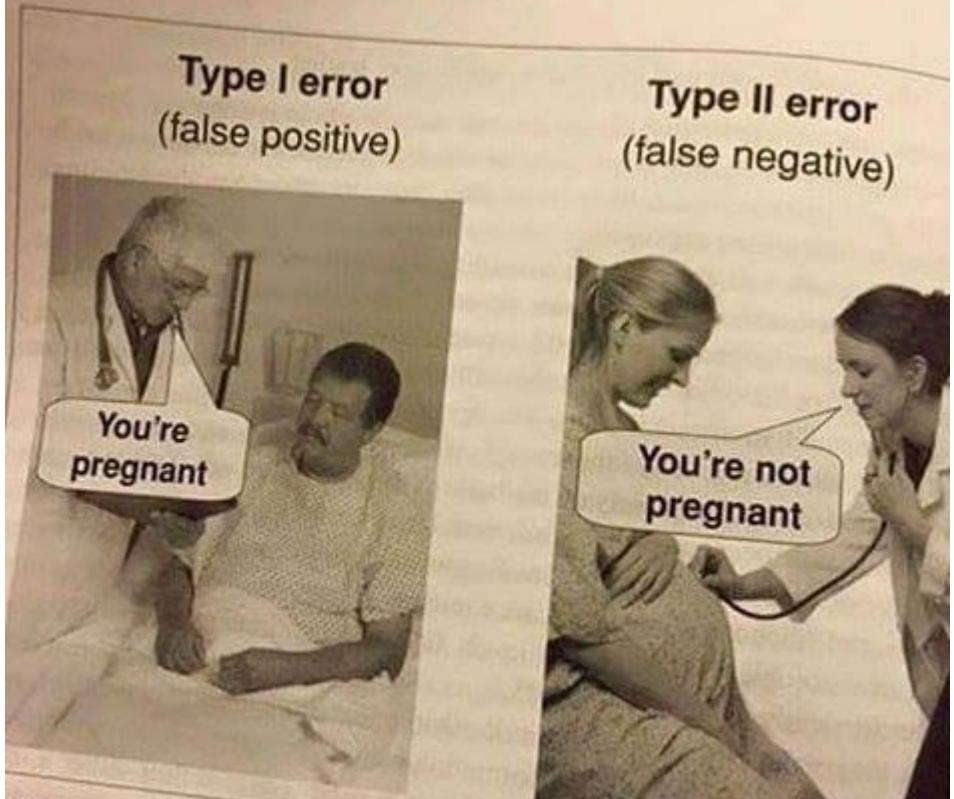
		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a: TP	b: FN
	Class=No	c: FP	d: TN

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)



e 3.1 Type I and Type II errors

Error Metrics

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a: TP	b: FN
	Class=No	c: FP	d: TN

Metric	Formula
True positive rate, recall	$\frac{TP}{TP+FN}$
False positive rate	$\frac{FP}{FP+TN}$
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F-measure	$\frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Error Metrics

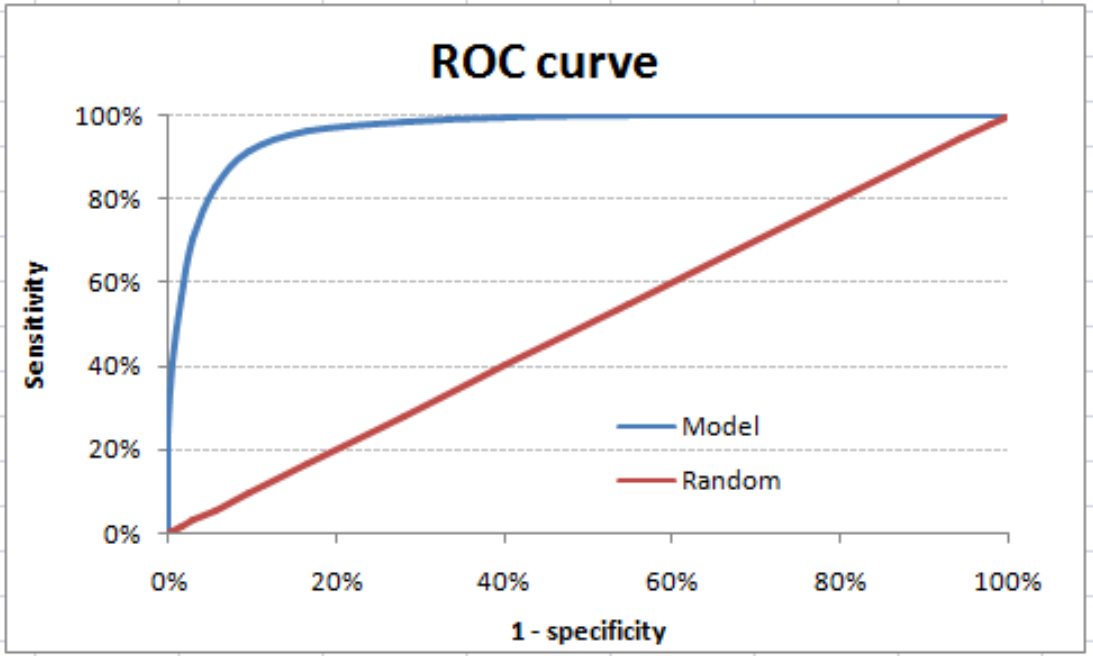
Metrics	Formula	Evaluation Focus
Accuracy (acc)	$\frac{tp + tn}{tp + fp + tn + fn}$	In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
Error Rate (err)	$\frac{fp + fn}{tp + fp + tn + fn}$	Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated.
Sensitivity (sn)	$\frac{tp}{tp + fn}$	This metric is used to measure the fraction of positive patterns that are correctly classified
Specificity (sp)	$\frac{tn}{tn + fp}$	This metric is used to measure the fraction of negative patterns that are correctly classified.
Precision (p)	$\frac{tp}{tp + fp}$	Precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.
Recall (r)	$\frac{tp}{tp + tn}$	Recall is used to measure the fraction of positive patterns that are correctly classified
F-Measure (FM)	$\frac{2 * p * r}{p + r}$	This metric represents the harmonic mean between recall and precision values
Geometric-mean (GM)	$\sqrt{tp * tn}$	This metric is used to maximize the *tp* rate and *tn* rate, and simultaneously keeping both rates relatively balanced

https://medium.com/greyatom/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b

The ROC curve

- In a Receiver Operating Characteristic (ROC) curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points.
- AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability.
- It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

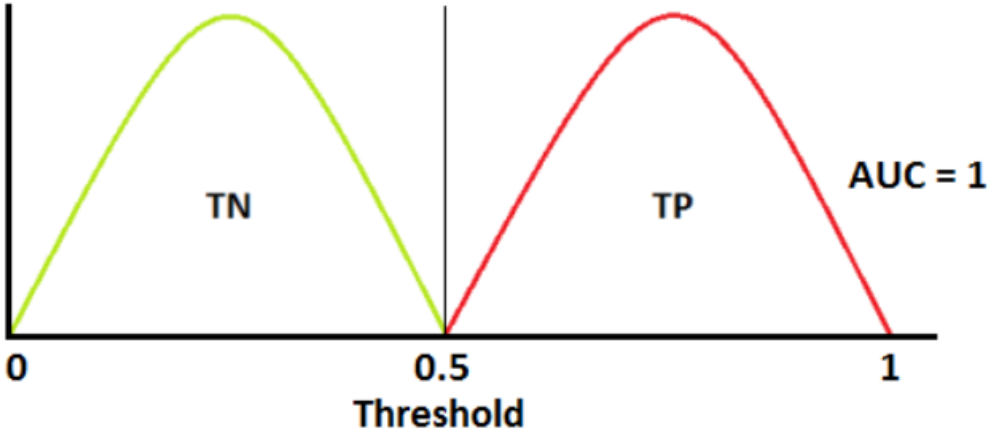
The ROC curve

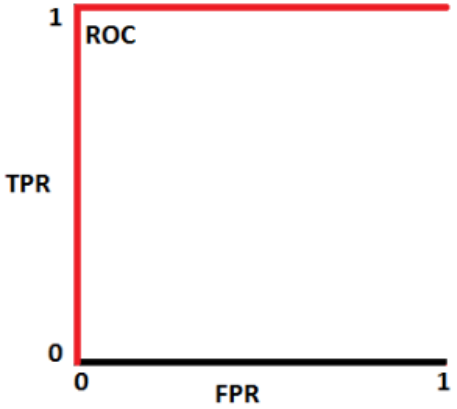


- 90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

The ROC curve

Red distribution curve is of the positive class (patients with disease) and green distribution curve is of negative class(patients with no disease)

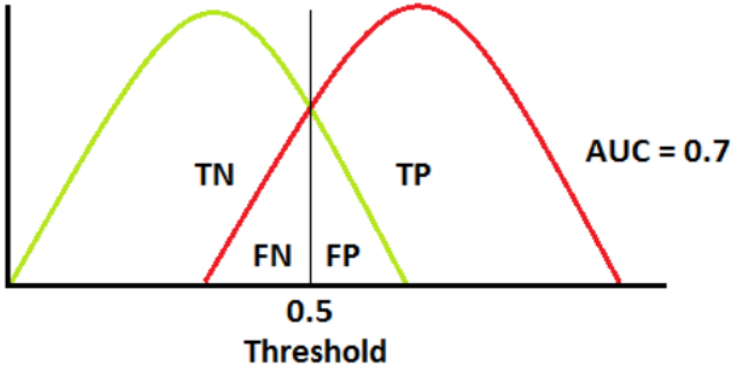


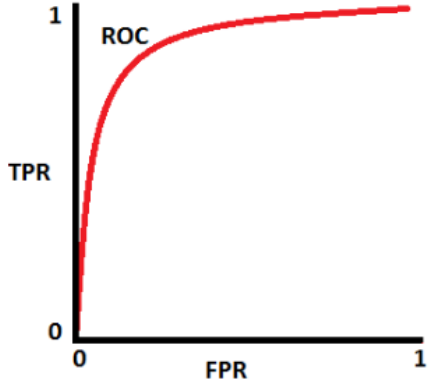


This is an ideal situation. When two curves don’t overlap at all means model has an ideal measure of separability. It is perfectly able to distinguish between positive class and negative class.

The ROC curve

Red distribution curve is of the positive class (patients with disease) and green distribution curve is of negative class(patients with no disease)

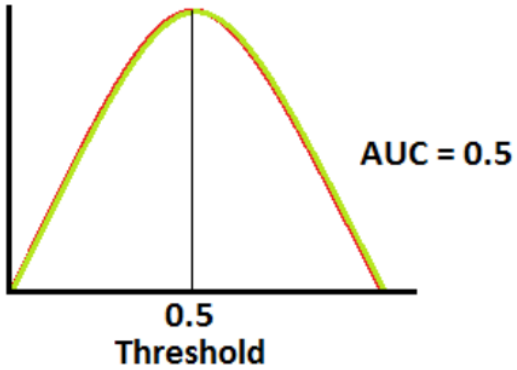


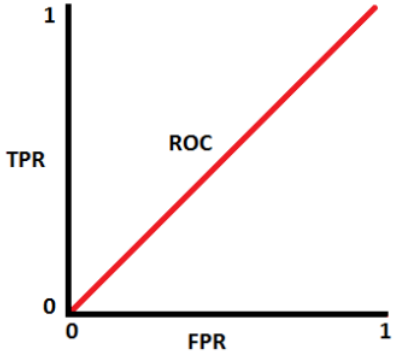


When two distributions overlap, we introduce type 1 and type 2 error. Depending upon the threshold, we can minimize or maximize them. When AUC is 0.7, it means there is 70% chance that model will be able to distinguish between positive class and negative class

The ROC curve

Red distribution curve is of the positive class (patients with disease) and green distribution curve is of negative class(patients with no disease)

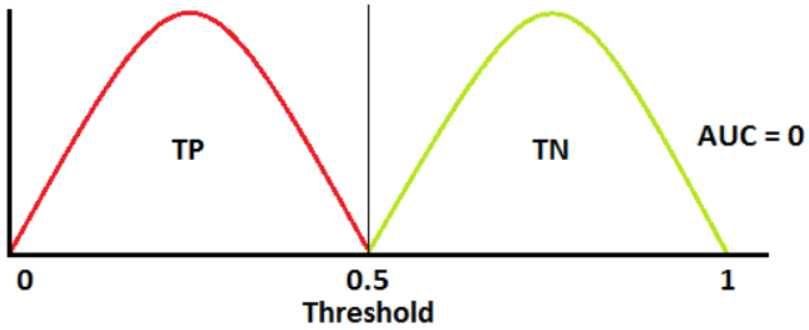


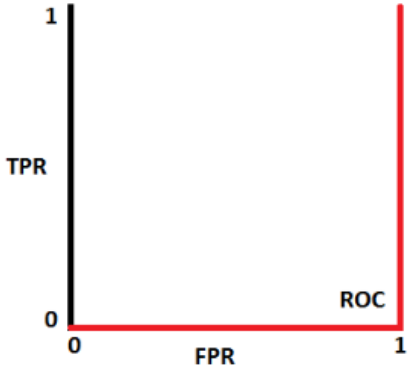


This is the worst situation. When AUC is approximately 0.5, model has no discrimination capacity to distinguish between positive class and negative class

The ROC curve

Red distribution curve is of the positive class (patients with disease) and green distribution curve is of negative class(patients with no disease)





When AUC is approximately 0, model is actually reciprocating the classes. It means, model is predicting negative class as a positive class and vice versa