# Mohamed Noordeen Alaudeen

## Senior Data Scientist
## Logistic Regression

# What type of target data was in Linear Regression?

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

# 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?

# 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:

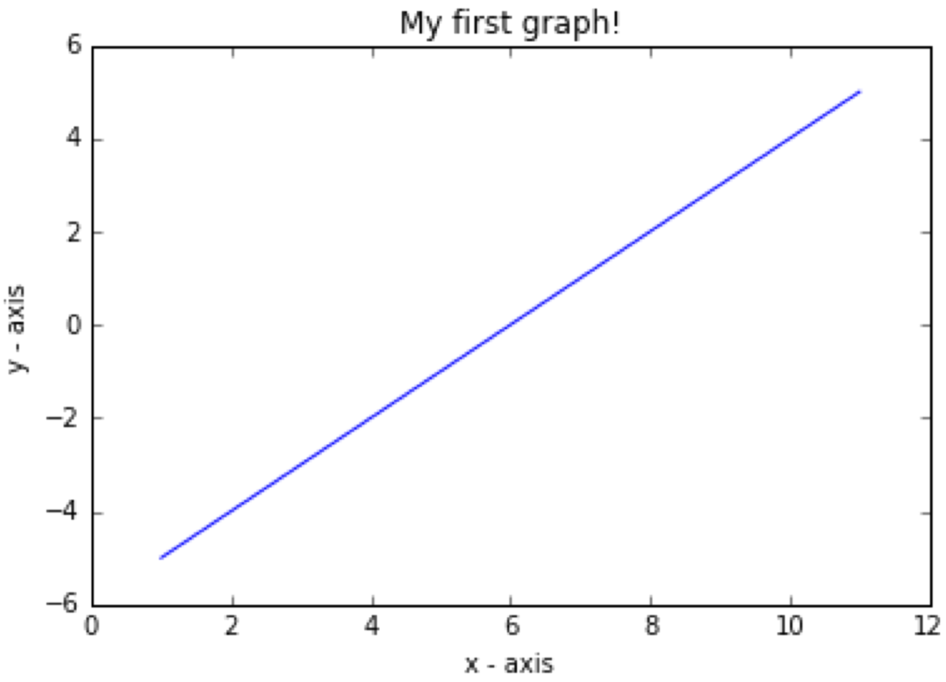
$$\text{Logodds} = \ln\left(\frac{p}{1-p}\right) \Rightarrow (-\infty, +\infty)$$

# “Logit” transformation of the probability

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

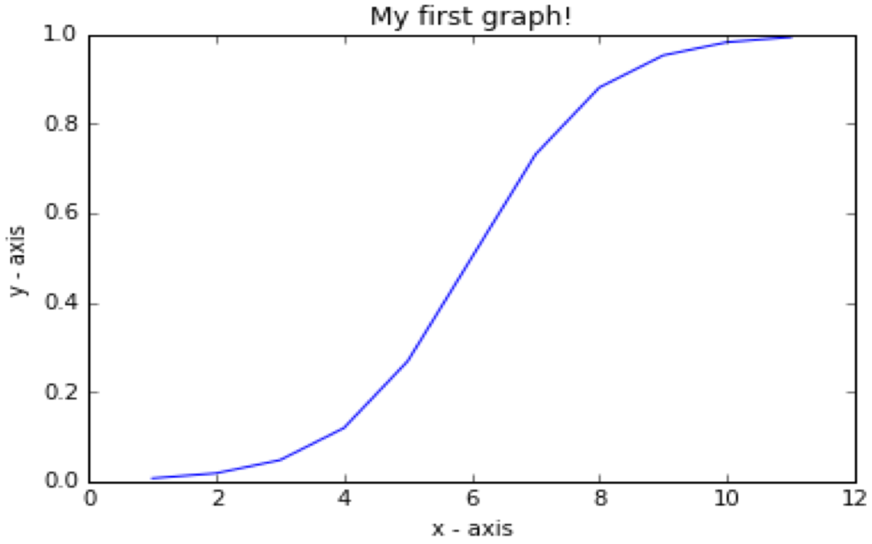
# Regression line

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

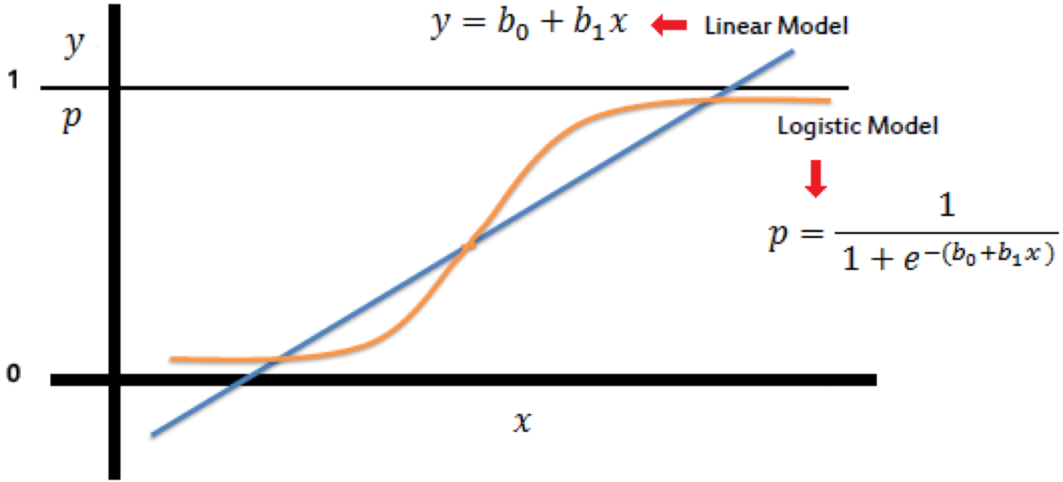


# Transformation to Classification

| 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

$g(y) = \beta o + \beta(Age)$

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)**

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

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

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

# 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

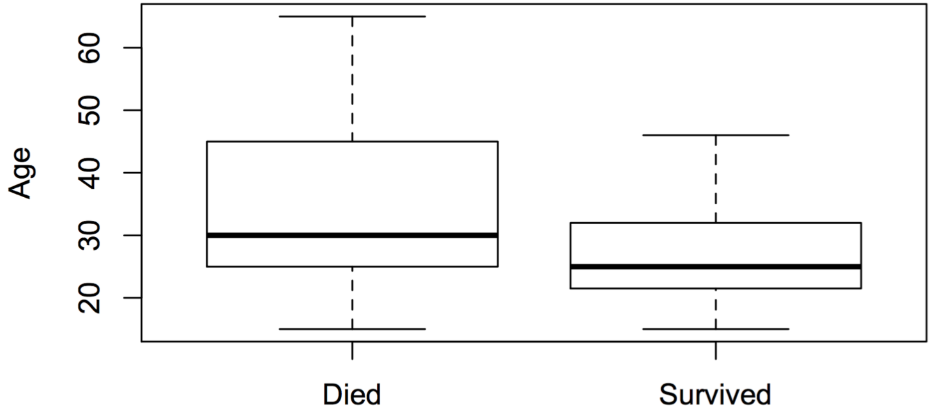
| | Age | Sex | Status |
|---|---|---|---|
| 1 | 23.00 | Male | Died |
| 2 | 40.00 | Female | Survived |
| 3 | 40.00 | Male | Survived |
| 4 | 30.00 | Male | Died |
| 5 | 28.00 | Male | Died |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 43 | 23.00 | Male | Survived |
| 44 | 24.00 | Male | Died |
| 45 | 25.00 | Female | Survived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

# 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 $P = \dfrac{1}{1+e^{-Y}}$

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

# 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:

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

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

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

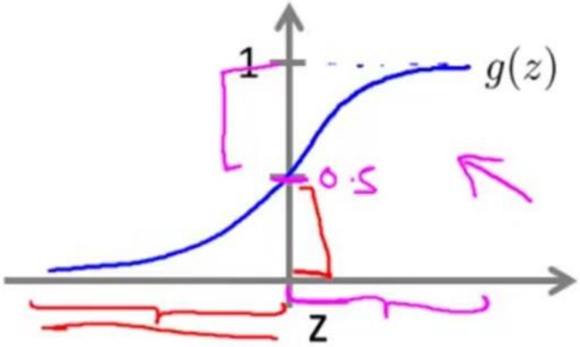
$$p = 6.16/7.16 = 0.86$$

# One Dimension

**Logistic regression**

$\rightarrow h_\theta(x) = g(\theta^T x) = P(y=1|x;\theta)$

$\rightarrow g(z) = \frac{1}{1+e^{-z}}$




Suppose predict $\boxed{\text{“}y = 1\text{“ if } h_\theta(x) \geq 0.5}$

$\rightarrow \theta^T x \geq 0$

$g(z) \geq 0.5$

when $z \geq 0$

$h_\theta(x) = g(\theta^T x) \geq 0.5$

whenever $\theta^T x \geq 0$

$z$

predict $\boxed{\text{“}y = 0\text{“ if } h_\theta(x) < 0.5}$

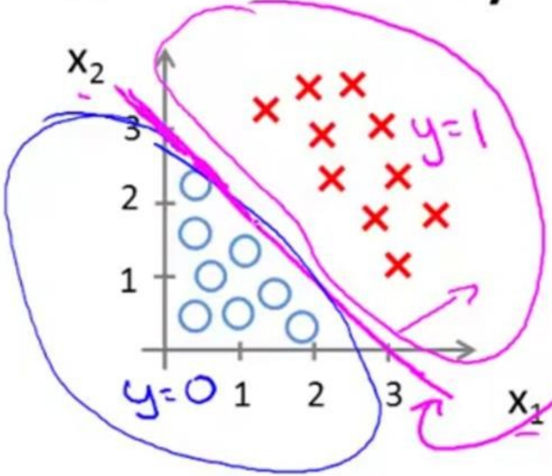
$h_\theta(x) = g(\theta^T x)$

$g(z) < 0.5$

$\rightarrow \theta^T x < 0$

# Two Dimension

## Decision Boundary




$$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

$$\rightarrow h_\theta(x) = g(\underbrace{\theta_0}_{=-3} + \underbrace{\theta_1}_{=1} x_1 + \underbrace{\theta_2}_{=1} x_2)$$

Decision boundary

Predict "$y = 1$" if $\underbrace{-3 + x_1 + x_2}_{\theta^T x} \geq 0$

$x_1, x_2$

$\rightarrow h_\theta(x) = 0.5$

$\boxed{x_1 + x_2 = 3}$

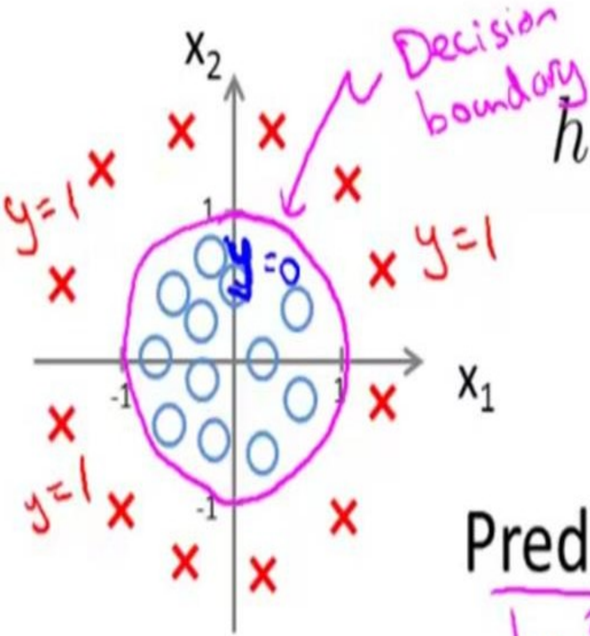
$\rightarrow x_1 + x_2 \geq 3$

$x_1 + x_2 < 3$

$\rightarrow y = 0$

# Two Dimension

## Non-linear decision boundaries




$$h_\theta(x) = g(\overset{=-1}{\theta_0} + \overset{=0}{\theta_1} x_1 + \overset{=0}{\theta_2} x_2 + \underset{=1}{\theta_3} x_1^2 + \underset{=1}{\theta_4} x_2^2)$$

$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

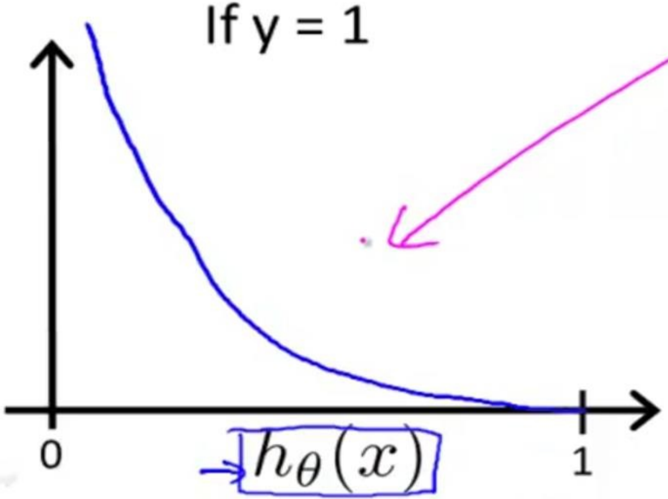
Predict "$y = 1$" if $\underbrace{-1 + x_1^2 + x_2^2 \geq 0}_{x_1^2 + x_2^2 \geq 1}$

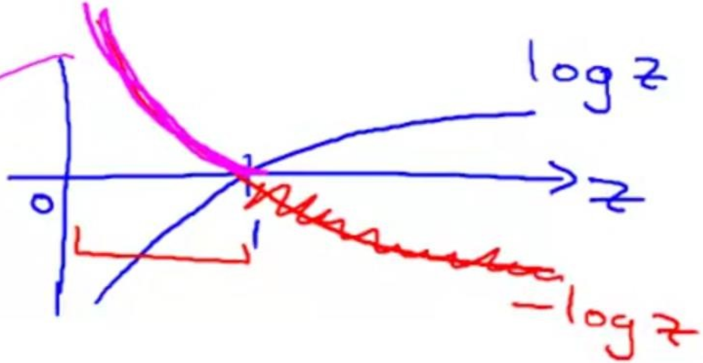
$$\boxed{x_1^2 + x_2^2 = 1}$$

# Cost Function

## Logistic regression cost function

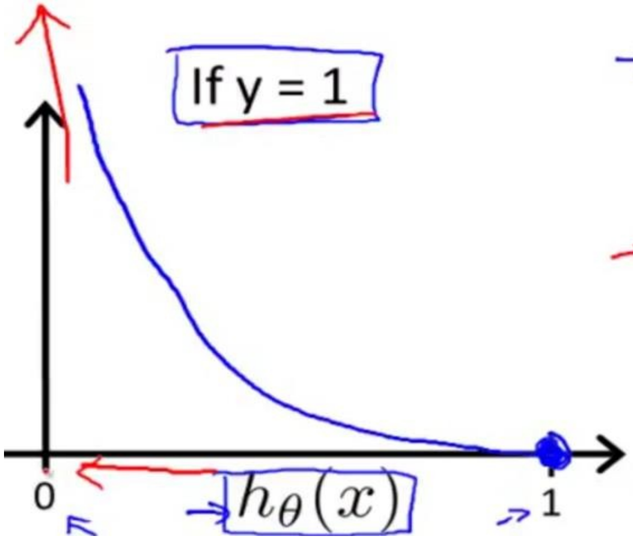
$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} \boxed{-\log(h_\theta(x))} & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$




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

# Cost Function for 1



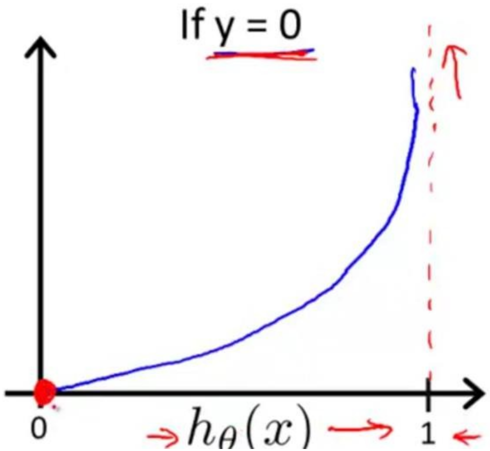

$\text{Cost} = 0$ if $y = 1, h_\theta(x) = 1$

But as $h_\theta(x) \to 0$

$Cost \to \infty$

Captures intuition that if $h_\theta(x) = 0$, (predict $P(y = 1|x; \theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.

# 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

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

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

Note: $y = 0$ or $1$ always

- For binary classification problems y is always 0 or 1
  - Because of this, we can have a simpler way to write the cost function
    - Rather than writing cost function on two lines/two cases
    - Can compress them into one equation - more efficient
- Can write cost function is
  - **$\mathrm{cost}(h_{\theta,}(x),y) = -y\log( h_\theta(x) ) - (1-y)\log( 1- h_\theta(x) )$**
    - This equation is a more compact of the two cases above

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

# Cost Function

## Logistic regression cost function

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

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

# Gradient Descent

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

Want $\min_\theta J(\theta)$:

Repeat $\{$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial\theta_j} J(\theta)$$

(simultaneously update all $\theta_j$)

$\}$

# 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:

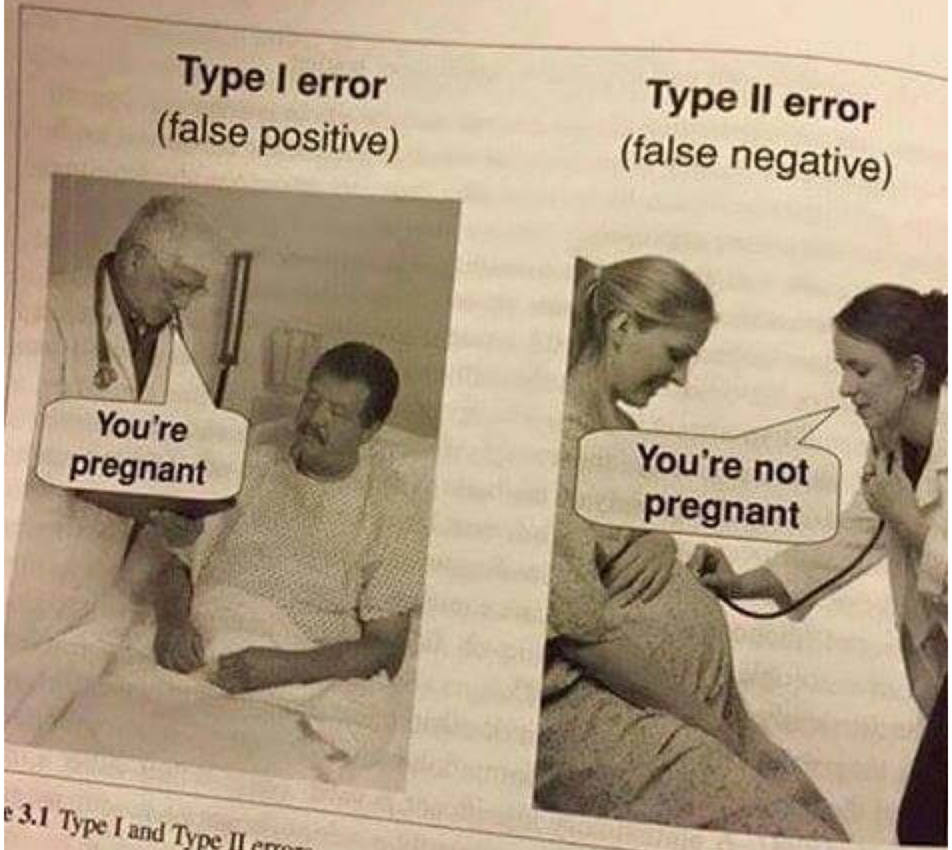
| | PREDICTED CLASS | | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| ACTUAL CLASS | Class=Yes | a: TP | b: FN |
| | Class=No | c: FP | d: TN |

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

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

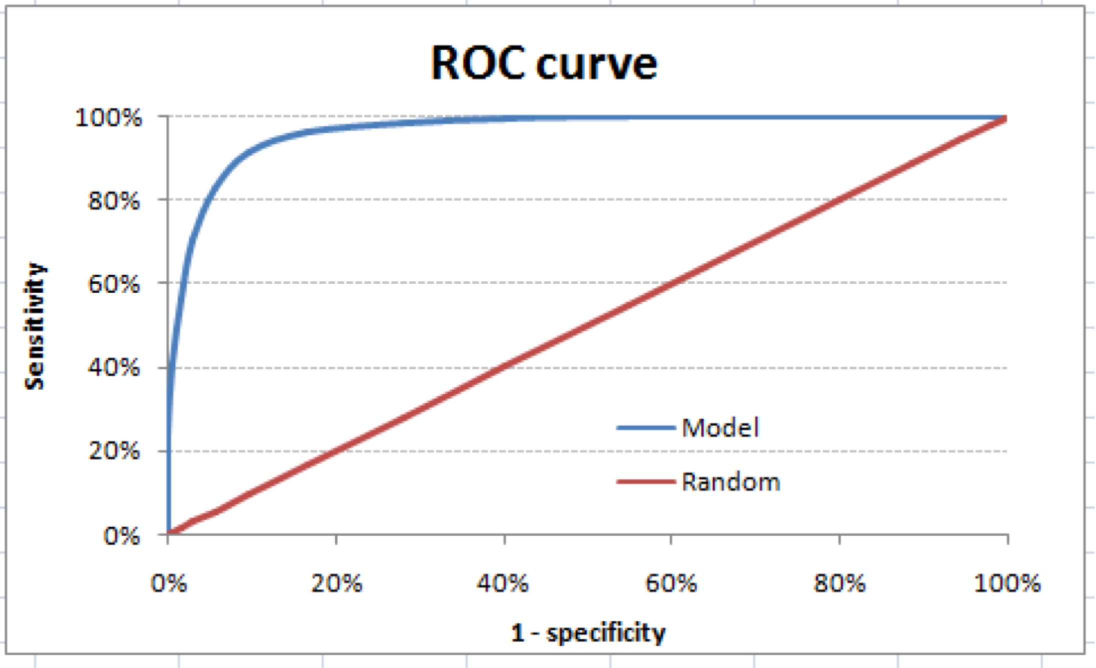
# Error Metrics

| 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.



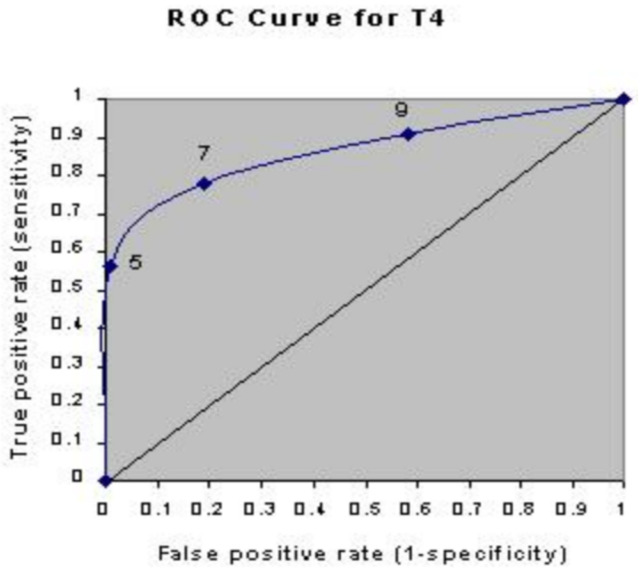

- 90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

# The ROC curve

| T4 value | Hypothyroid | Euthyroid |
|---|---|---|
| 5 or less | 18 | 1 |
| 5.1 - 7 | 7 | 17 |
| 7.1 - 9 | 4 | 36 |
| 9 or more | 3 | 39 |
| **Totals:** | 32 | 93 |

| Cutpoint | True Positives | False Positives |
|---|---|---|
| 5 | 0.56 | 0.01 |
| 7 | 0.78 | 0.19 |
| 9 | 0.91 | 0.58 |

| Cutpoint | Sensitivity | Specificity |
|---|---|---|
| 5 | 0.56 | 0.99 |
| 7 | 0.78 | 0.81 |
| 9 | 0.91 | 0.42 |



# Logistic Regression Merits

- Simple and efficient.
- Low variance.
- It provides **probability** score for observations.

# Logistic Regression Demerits

- Doesn't handle **large** number of categorical features/variables well.
- It requires transformation of non-linear features.

# Additional Data

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 1.6331 | 1.1102 | 1.47 | 0.1413 |
| Age | -0.0782 | 0.0373 | -2.10 | 0.0359 |
| SexFemale | 1.5973 | 0.7555 | 2.11 | 0.0345 |

$$H_0 : \beta_{age} = 0$$
$$H_A : \beta_{age} \neq 0$$

$$Z = \frac{\hat{\beta_{age}} - \beta_{age}}{SE_{age}} = \frac{-0.0782 - 0}{0.0373} = -2.10$$

$$\begin{aligned} \text{p-value} &= P(|Z| > 2.10) = P(Z > 2.10) + P(Z < -2.10) \\ &= 2 \times 0.0178 = 0.0359 \end{aligned}$$