# Logistic Regression

## Supervised Learning

# Agenda

- Regression vs classification
- Types of classification
- Terms used in logistic regression
    - Odds vs probability
    - Log of odds
    - Odds ratio
- Logistic regression - why it is called regression?
- What is sigmoid function?
- Logistic Regression
- Log loss
- Performance measures - Confusion matrix, AUC/ROC
- Threshold optimization
- Advantages and disadvantages
- Summary

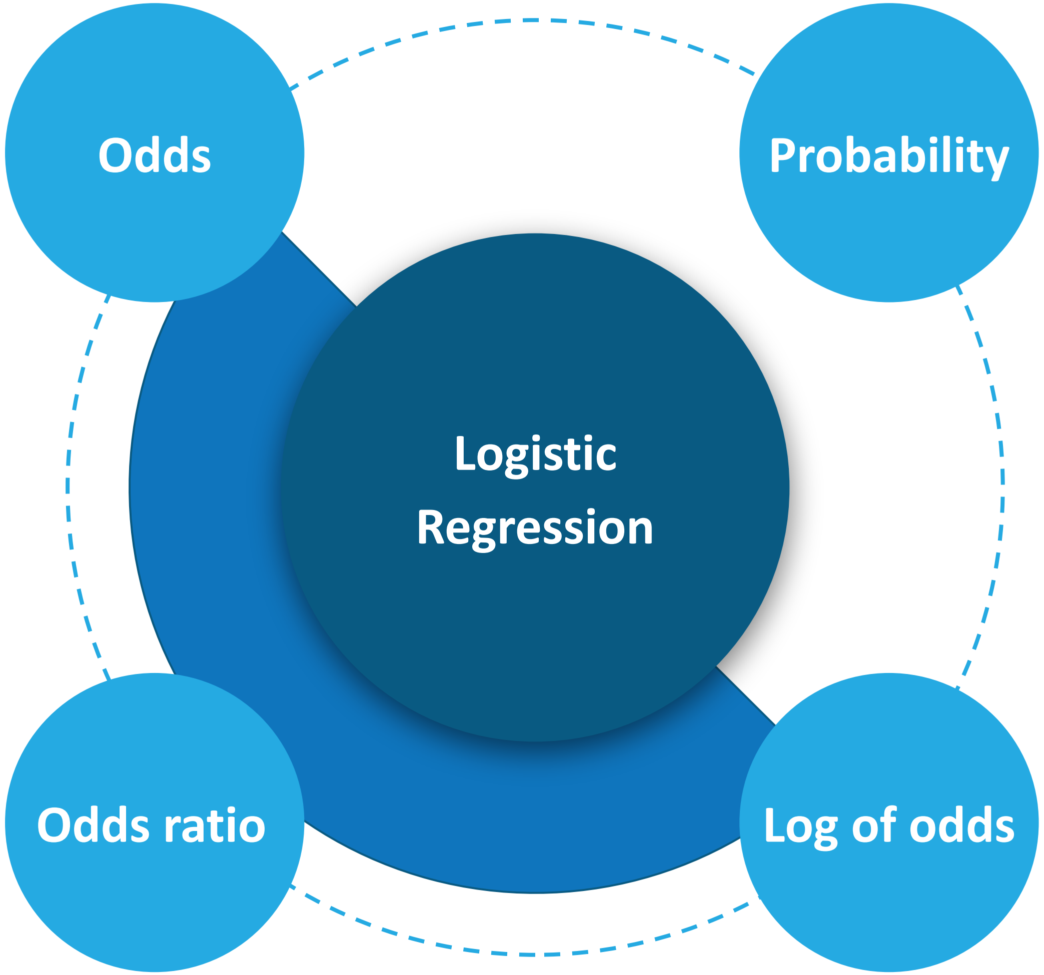
# Regression vs Classification

| Regression | Classification |
|---|---|
| ● Target variable is continuous.<br>● Involves estimating or predicting a response<br>● Example - How many marks will I get in an exam of 100 marks.<br>● Algorithms - Linear regression, regression tree | ● Target variable is discrete.<br>● Identifies a label<br>● Example - Will I pass or fail in the exam.<br>● Algorithms - Logistic regression, SVM Classifier, kNN classifier, Decision Tree classifier |

# Types of classification

- **Binary classification**
  - Example - Will I pass or fail in the exam?
  - Here we have two labels, pass and fail. We need to predict if a student is going to pass/fail based on the features. This a binary classification.
- **Multi-class classification**
  - How is the weather today? Labels - sunny, rainy, cold
  - Here we have three labels.

# Terms used in Logistic Regression



Logistic Regression

Odds

Probability

Odds ratio

Log of odds

# Odds vs probability

**Odds** of an event are the ratio of number of observations in favour of an event to number of observations not in favour of the event

$$\text{Odds} = \frac{\text{No. of observations in favour of the event}}{\text{No. of observations not in favour of the event}}$$

**Probability** of an event is the ratio of number of observations in favour of an event to all possible observations

$$\text{probability}= \frac{\text{No. of observations in favour of the event}}{\text{No. of observations}}$$

# Odds vs probability

| **Hours of study** | 0.5 | 1 | 4 | 3.5 | 1.5 | 5.5 | 3 | 5 | 4.5 | 4 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Is the student passed in exam** | Fail | Fail | Pass | Pass | Fail | Pass | Fail | Pass | Pass | Pass |

For the above data, the odds of a student passing the exam is given by,

$$\text{probability=}\ \frac{\text{No. of students passing the exam}}{\text{Total no. of students}} = 6/10$$

$$\text{Odds} = \frac{\text{No. of students passing the exam}}{\text{No. of students not passing the exam}} = 6/4$$

# Log of odds

Odds of passing the exam = 6/4

log(odds of passing exam) = ln(1.5) = 0.405

Odds of not passing the exam = 4/6

log(odds of not passing exam) = ln(0.667) = - 0.405

- As we can see if we only consider odds value, the magnitude for each class value taken by variable is very different.
- Hence, the log(odds) value is considered; so that no matter whichever the class is, the magnitude remains same.
- Log of odds is the logit function used in logistic regression.

# Relation between odds and probability

If P(A) is probability of event A

$$\text{Odds} = \frac{P(A)}{1- P(A)}$$

$$\text{probability}=\frac{\text{odds}}{1+\text{odds}}$$

$$\text{Log(odds)}= \ln(P(A)/1-P(A))$$

# Odds ratio

- Odds ratio refers to the ratio of odds.
- Odds ratio can be used to determine the impact of a feature on target variable.
- For our considered example the odds ratio can be calculated as,

$$\text{Odds Ratio} = \frac{\text{Odds of a student passing the exam}}{\text{Odds of a student not passing the exam}} = \frac{6/4}{4/6} = 9/4$$

# Logistic Regression - why it is called regression?

- Logistic Regression is a classification method but it's built on the same concept as linear regression because it produces a linear decision boundary.
- The reason for producing a linear decision boundary is because our outcome depends on the additivity of the features.

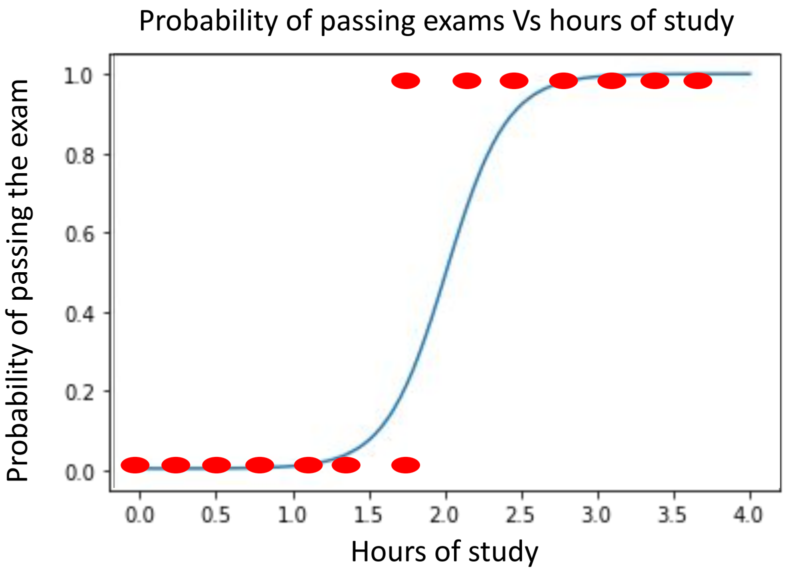
  f(x) = ax1 + bx2 + c
- Model has a linear component, but it introduces nonlinearity due to the sigmoid function.

# What is sigmoid function?

- Logistic function is also called as sigmoid.
- Function which is given by,

$$S(x) = \frac{1}{1+e^{-f(x)}} = \frac{e^{f(x)}}{1+ e^{f(x)}}$$

- It maps predicted values to probabilities
- Value ranges between 0 and 1



# How it works - Theory

- Logistic Regression is a statistical technique that predicts probability of a target variable based on the independent features.
- It predicts the probability of occurrence of a class label. Based on these probabilities the data points are labelled.
- Probability of an outcome (y) is calculated using sigmoid function $f(x) = (1/(1+e^{-f(x)})$ which is then used to decide the class based on the threshold value.
- A threshold (or cut-off; commonly a threshold of 0.5 is used) is fixed, then

| | **Classify as** |
|---|---|
| Probability > threshold | 1 |
| Probability < threshold | 0 |

# Logistic Regression

- Logistic regression is very much similar to linear regression where the explanatory variables(X) are combined with weights values to predict a target variable of binary class(y).
- **f(x) = a+bx** here, f(x) can have values from -∞ to ∞
- **log(p/(1-p)) = f(x)**
  - Here, p is the probability that the event y occurs(Y=1) [range 0 to 1]
  - p/(1-p) is the **odds ratio** [range 0 to infinity]
  - log(p/(1-p)) is **log of odds ratio** (logit) [-∞ to ∞]
- log(p/(1-p)) = a+bx : log of p/(1-p) is linearly related to the features and can have value between -∞ to ∞

# Logistic Regression

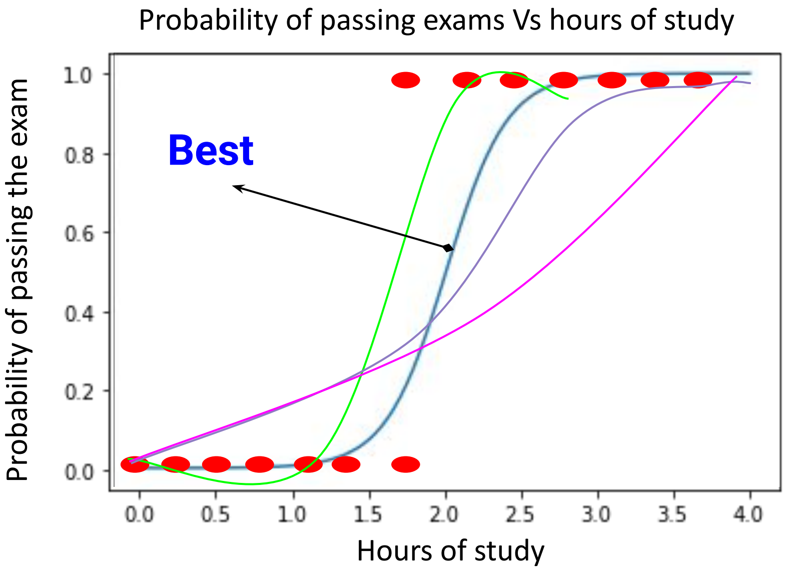
- Exponential of the logit and you have the odds for the two groups in question.
- $p/(1-p) = e^{f(x)}$ : Odds (range from 0 to infinity with values greater than 1 associated with an event being more likely to occur than to not occur and values less than 1 associated with an event that is less likely to occur)
- $P(Y) = 1/(1+e^{-f(x)})$ : Sigmoid function calculates the probability
- $p(y) = 1/(1+e^{-(a+bx)})$ : If f(x) = 0 then p = 0.5 as f(x) increases, p approaches 1 and as f(x) gets really small, p approaches 0.

**Note** - Logarithm or logit transformation is used to model the non-linear relationship between Y and X by transforming Y.

# Log loss

- Linear regression minimize the sum of squared error for finding the best fit line. This unfortunately will not work in logistic regression.
- Instead we choose to minimize the “Log Loss” or “Cross-Entropy” to find the best fit curve.

**-ylog(ŷ) - (1-y)log(1-ŷ)**



- Correct classification contributes very minimal to the sum while a incorrect classification contributes large magnitudes.

# Performance measure

- Accuracy
- Confusion Matrix
- Recall
- Precision
- F-Score
- Sensitivity ( TPR )
- Specificity
- AUC/ROC curve
- Classification report

# Confusion Matrix

| 50 { 30 P and 20 N} | Actual ( y=1) {1 is positive} | Actual ( y=0) |
|---|---|---|
| Predicted (y = 1) | TP (26) | FP (2) |
| Predicted (y = 0) | FN (4) | TN (18) |

- **True Positive :** predicted and actual values are true.
- **True Negative** : predicted and actual values are false.
- **False Positive (Type 1 Error) :** predicted is true and actual value is false.
- **False Negative (Type 2 Error) :** predicted is false and actual value is true.

# Confusion Matrix for multi-class

- It displays the frequency table for each class.

| Total - 90 | Actual Class 1 ( 30) | Actual Class 2 (30) | Actual Class 3(30) |
|---|---|---|---|
| Predicted Class 1 | 26 | 1 | 1 |
| Predicted Class 2 | 3 | 27 | 1 |
| Predicted Class 3 | 1 | 2 | 29 |

# Precision and Recall

**Precision** : TP / (TP + FP) it is the ratio of true positives to the total positive predictions.

**Recall** : TP / (TP + FN) its is the ratio of true positives to the total actual positives observations. It is also known as sensitivity or TPR.

**Specificity:** TN/(TN+FP) it is the ratio of true negatives to the total actual negative observations.

**False positive rate:** (1-Specificity) it is the ratio of false positives to the total actual negative observations.

# Classification report

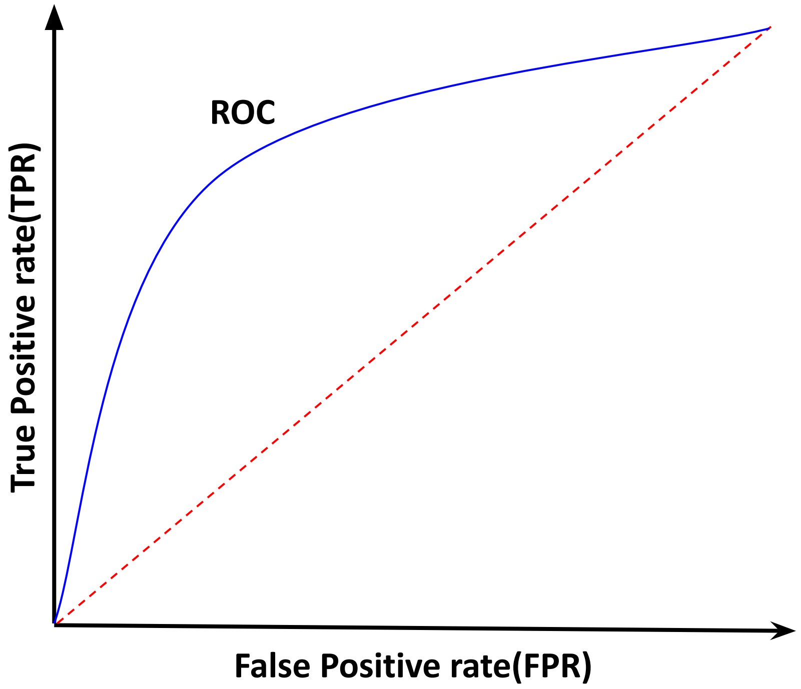
- It displays the precision, recall, F1, and support scores for the model.
- **F1 Score :** it is the harmonic mean of precision and recall

  F1 Score = 2*(Recall * Precision) / (Recall + Precision)
- **Support :** it shows the number of samples of actual occurrences of the class in the data

  If all the classes are not represented equally in the training data then our results might not be good and we may have to explore sampling technique.

# ROC curve - Receiver Operating Characteristic

- ROC curve is a plot between TPR vs FPR, as the threshold value is increased from zero to one
- As we know that a good model need to correctly identify positive as  positive and a negative as a negative.
- Hence it should have high sensitivity (TPR) and low FPR (1-specificity) which will result in the ROC curve close to left corner of the plot.
- ROC curve shows that if AUC is more than 0.7 ( ROC line above the diagonal line ) then the model is able to make the correct predictions for most of the thresholds.
- A model with no discrimination ability will have an ROC curve which is the 45 degree diagonal line.



# Threshold optimization

- Logistic regression return a probability value which ranges between 0 to 1.
- To assign a class based on the returned probability value, a cut-off value ( Threshold) is used to segregates the instances into classes.
- Let’s say, threshold = 0.5. Instances with probability value > 0.5 will belong to one class and with value <0.5 will belong to another class.
- You can use ROC/AUC curve or check FP and FN to optimize the threshold value.

**Note** - By optimizing the threshold we trade off TP for FN, and FP for TN. Changing threshold does not change the coefficients of the model it just updates the FP and FN.

# Advantages and disadvantages

- **Advantages**
  - Simple and easy to implement
  - Can easily address multiclass targets
  - Very fast at classifying unknown records and good accuracy for many simple data sets.
- **Disadvantages**
  - Constructs linear boundaries
  - Assumes that variables are independent

# Hands on

# Summary

In this module we discussed:

- Regression vs classification
- Types of classification
- Odds vs probability, log of odds and odds ratio
- Logistic regression - why it is called regression?
- What is sigmoid function
- Logistic regression and Log loss
- Performance measures - Confusion matrix, AUC/ROC
- Threshold optimization
- Advantages and disadvantages

# THANK YOU

# Happy learning ☺️