LOGISTIC REGRESSION

Kumar Sundram

- Overview
- Maths – intuition

WHAT IS LOGISTIC REGRESSION

- Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).
- Like all regression analyses, the logistic regression is a predictive analysis.
- Logistic regression is used to describe the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables
- In other words, the logistic regression model predicts $P(Y=1)$ as a function of X.

LOGISTIC REGRESSION IS A TYPE OF CLASSIFICATION ALGORITHM

- Unlike actual regression, logistic regression does not try to predict the value of a numeric variable given a set of inputs.

- Instead, the output is a probability that the given input point belongs to a certain class.

- For simplicity, lets assume that we have only two classes, and the probability in question is
  - P_+ -> the probability that a certain data point belongs to the '+' class.
  - $P_- = 1 - P_+$.

- Thus, the output of Logistic Regression always lies in [0, 1].

LINEARLY SEPARABLE CLASSES

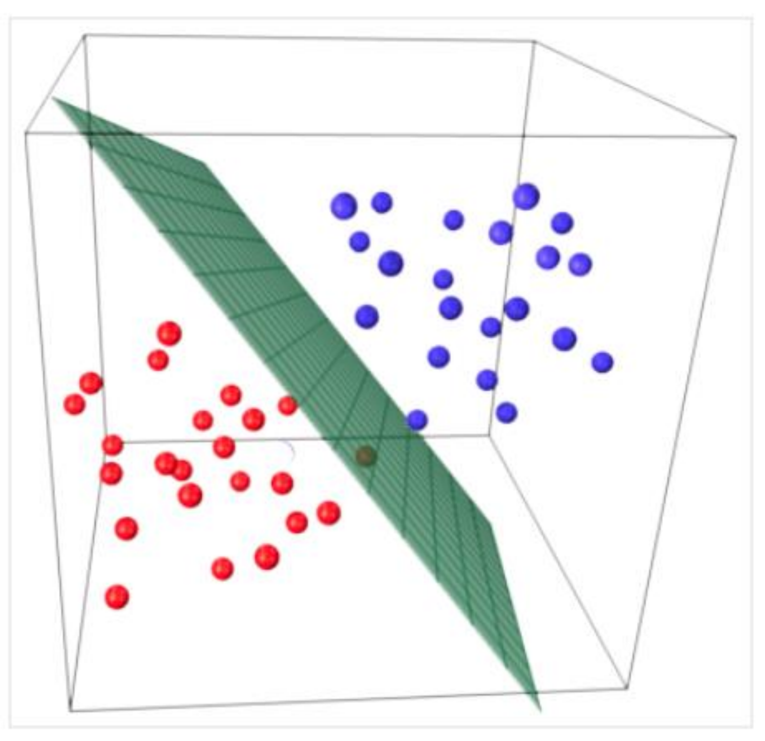
- The central premise of Logistic Regression is the assumption that input space can be separated into two nice 'regions', one for each class, by a linear boundary.

- So what does a 'linear' boundary mean?

- For 2 dimensions, its a straight line- no curving.

- For 3 dimensions, its a plane.

- This dividing plane is called a linear discriminant,
  - its linear in terms of its function,
  - it helps the model 'discriminate' between classes.



WHAT IS LOGISTIC REGRESSION

- Type of questions (DS objective) that a binary logistic regression can examine.
  - How does the probability of getting lung cancer (yes vs. no) change for every additional pound a person is overweight and for every pack of cigarettes smoked per day?
  - Do body weight, calorie intake, fat intake, and age have an influence on the probability of having a heart attack (yes vs. no)?
  - Should a bank give a person a loan? Yes/ No
  - Is an individual transaction fraudulent or not?
  - If people are likely to vote for new legislation or not?

WHAT IS LOGISTIC REGRESSION

- Continuous Vs categorical variables
- General linear regression model : $y = b0 + b1.x1 + b2.x2 + e$
- Independent variables (Xs)
  - Continuous : age, income, height, -> use numerical values
  - Categorical : gender, ethnicity, sex, status -> use dummy variables
- Binary outcomes
  - Representing a binary outcome
    - YES | NO
    - Use dummy variables → YES: 1, NO: 0

BINARY LOGISTIC REGRESSION MAJOR ASSUMPTIONS:

- The dependent variable should be dichotomous in nature (e.g., presence vs. absent).
- There should be no outliers in the data,
  - E.g. which can be assessed by converting the continuous predictors to standardized scores, and removing values below -3.29 or greater than 3.29.
- There should be no high correlations (multicollinearity) among the predictors.
  - This can be assessed by a correlation matrix among the predictors.
- At the center of the logistic regression analysis is the task estimating the log odds of an event.
- Logistic regression requires quite large sample sizes.

EXAMPLE

- We have data on 1000 random customers from a given city. We want to know what determines their decision to subscribe to a magazine

- Subscribe : Indicates if a customer has subscribed to the magazine

- Age: Examine how age influences the likelihood of the subscription

- Other attributes : …

A LINEAR MODEL?

- Besides the outcome being binary, there is nothing special about the DV (y, subscribe)
- If a customer subscribes, the value of y is higher (from 0 to 1)
- We can apply the linear regression:-
  - y (subscribe) = $\beta_0 + \beta_1 Age + \varepsilon$
  - y (subscribe) = -1.700 + 0.064 * Age

	coefficient
const	-1.70073
age	0.0645433

INTERPRETING THE RESULTS

- If the DV is binary then the focus should be to see what makes it change from y=0 to y=1

- This is also explained as the likelihood of subscription or p (subscribe = 1)

- y (subscribe) = -1.700 + 0.064 * Age
- P (subscribe = 1) = p = -1.700 + 0.064 * Age

- Every additional year of Age, increases the probability of subscription by 6.4%

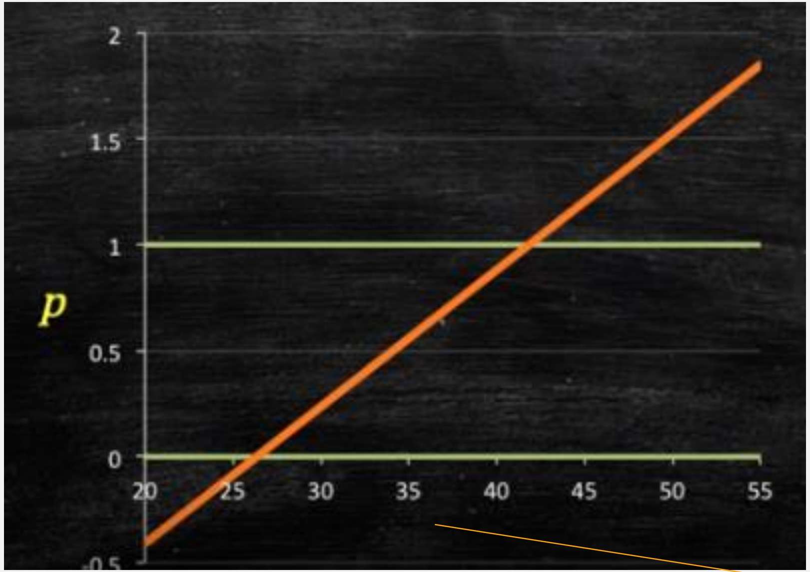
PROBLEMS WITH THE LINEAR APPROACH

- Probabilities are bounded,) 0 =< p =< 1
- The range of age in the data is 20 =< age =< 55

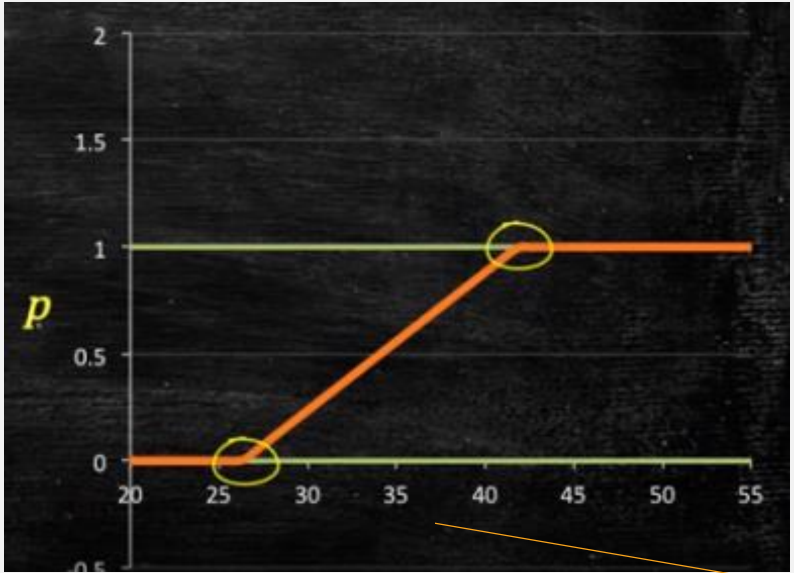
- The probability that a 35 year old person subscribes is
  - P = -1.700 + 0.064 * 35 = 0.54

- The probability that a person 25 years or 45 years subscribes?
  - P = -1.700 + 0.064 * 25 = -0.09 … possible?
  - P = -1.700 + 0.064 * 45 = +1.20

PROBABILITY PLOT

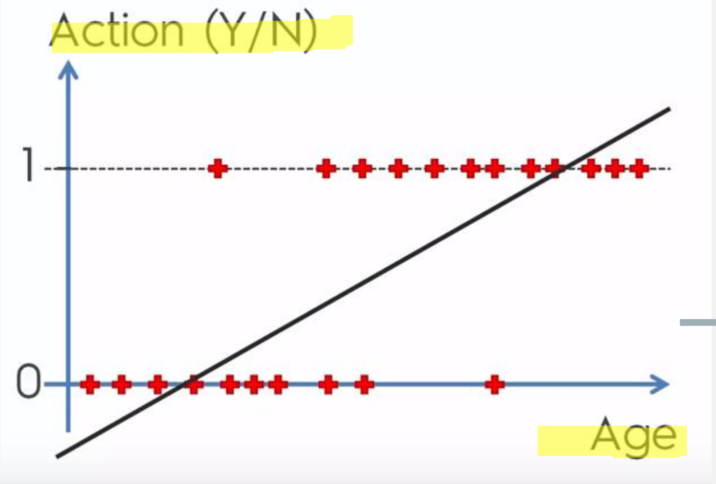


1. Customers of more than 45 years of age have probability > 1
2. Customer who are less than 25 years age, the probability is less than 0



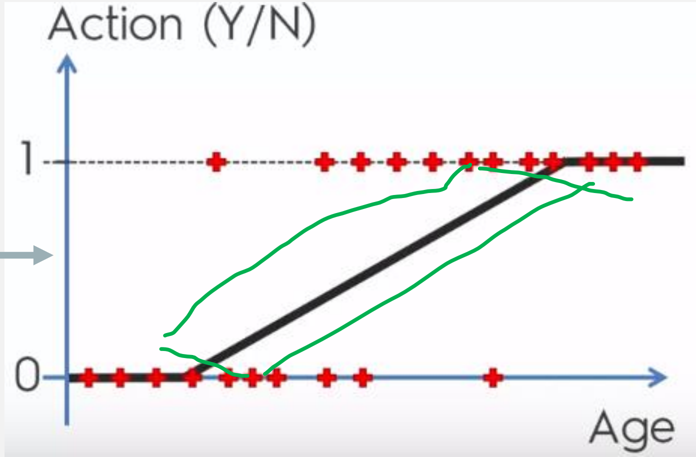
1. Any probability > 1.0, can be made 1.0
2. Any probability < 0.0, can be made 0.0

LOGISTIC FUNCTION - EXAMPLE



Predictor : X : age
Outcome : Action : y :

Depending on age, predict if the person will take the offer or not (ACTION = 1 or 0)



FIXING THE ISSUE

- We need to somehow constrain p such that $0 \leq p \leq 1$

- We need to ensure
  - Probability , p, must always be POSITVE
  - It must be ≤ 1

WAYS TO FIX

Absolute of a number $p = \|X\|$	Square of a number $p = X^2$	Taking exponentiation $p = e^{(\beta 0 + \beta 1 Age)}$
Solves the ≤ 0 issue Does not solve the >1 issue	Solves the ≤ 0 issue Does not solve the > 1 issue	Solves the ≤ 0 issue Does not solve the > 1 issue
		$p = e^{(\beta 0 + \beta 1 Age)} / [e^{(\beta 0 + \beta 1 Age)} + 1]$

1. Even though the probability of a customer subscribing (p) is not a linear function of age, the simple transformation is now a linear function of age

2. Refer to the code demo ML-LOGIT-13-glass-coeff-logodds-prob-maths-interpretation

$$p = e^{(\beta 0 + \beta 1 Age)} / [e^{(\beta 0 + \beta 1 Age)} + 1]$$
$$1/p = [e^{(\beta 0 + \beta 1 Age)} + 1] / e^{(\beta 0 + \beta 1 Age)}$$
$$1/p = 1 + 1/ [e^{(\beta 0 + \beta 1 Age)}]$$
$$1/p - 1 = 1/ [e^{(\beta 0 + \beta 1 Age)}]$$
$$(1-p)/p = 1/ [e^{(\beta 0 + \beta 1 Age)}]$$
$$p/(1-p) = e^{(\beta 0 + \beta 1 Age)}$$
$$\boxed{\ln(p/1-p) = \beta_0 + \beta_1 Age}$$

Equation for logistic regressions

BASICALLY A SIGMOID FUNCTION

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * Age. \longrightarrow (1)$$

Now; our function gives values $(-\infty \text{ to } \infty)$

odds ratio $= \left(\frac{p}{1-p}\right)$

Probability	~~log~~(OR)	log (OR)
0	$\frac{0}{1-0} = 0$	$\log(0) \rightarrow -\infty$
1	$\frac{1}{1-1} = \infty$	$\log(\infty) \rightarrow \infty$

Bhupen.

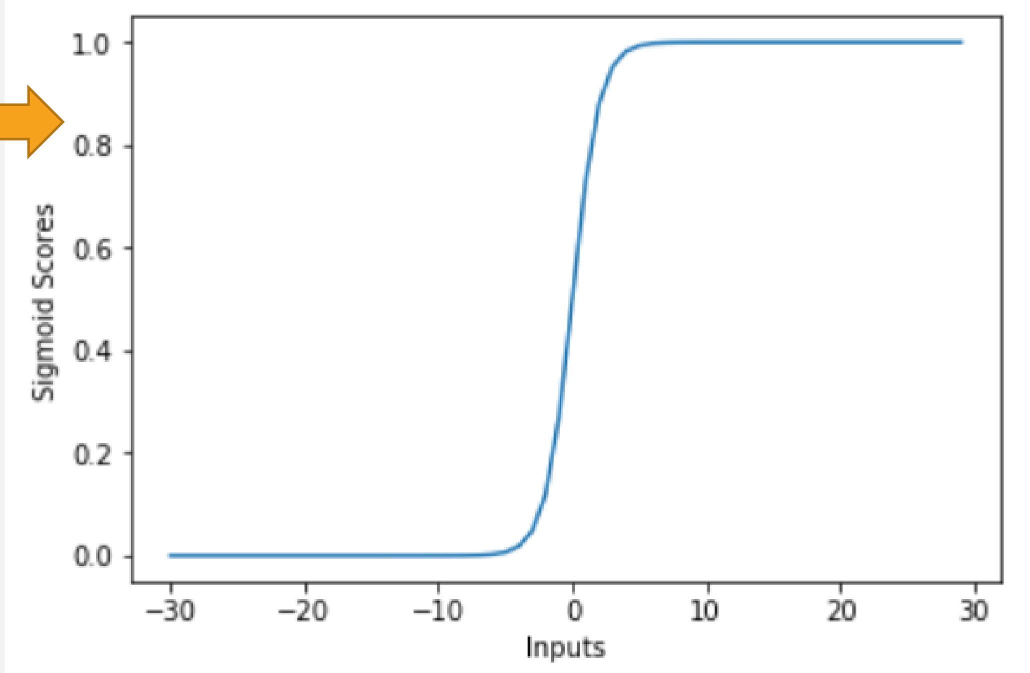
Since, $\log\left(\frac{p}{1-p}\right) = z$ (function) from (1)

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

$$\Rightarrow e^z(1-p) = p \Rightarrow e^z = p(1+e^z)$$

$$\Rightarrow p = \frac{e^z}{1+e^z} \Rightarrow \frac{1}{\frac{1}{e^z}+1} = \left(\frac{1}{1+e^{-z}}\right)$$

(Sigmoid function)



PROBABILITY THRESHOLD

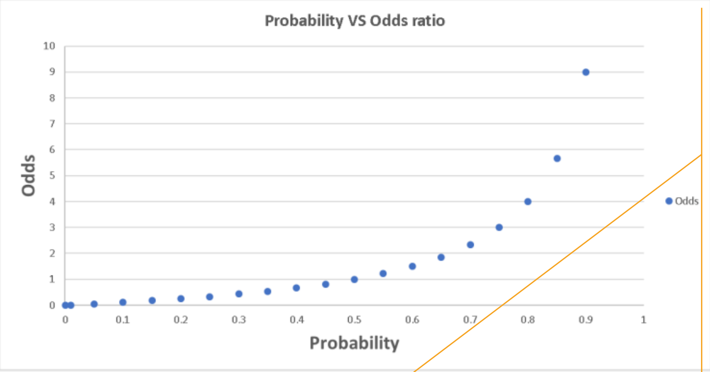
- default threshold when creating predictions is 0.5.

- How can we change the default setting ?

- Say we want the model to predict a '1' for probability greater than 0.25, not 0.5

ODDS ETC

probability	odds	logodds
0.001	0.001001001	-6.906754779
0.01	0.01010101	-4.59511985
0.05	0.052631579	-2.944438979
0.1	0.111111111	-2.197224577
0.15	0.176470588	-1.734601055
0.2	0.25	-1.386294361
0.25	0.333333333	-1.098612289
0.3	0.428571429	-0.84729786
0.35	0.538461538	-0.619039208
0.4	0.666666667	-0.405465108
0.45	0.818181818	-0.200670695
0.5	1	-1.11022E-16
0.55	1.222222222	0.200670695
0.6	1.5	0.405465108
0.65	1.857142857	0.619039208
0.7	2.333333333	0.84729786
0.75	3	1.098612289
0.8	4	1.386294361
0.85	5.666666667	1.734601055
0.9	9	2.197224577
0.99	99	4.59511985
0.999	999	6.906754779
0.9999	9999	9.210240367
0.999999	999999	13.81550956
0.999999999	1000000027	20.72326586



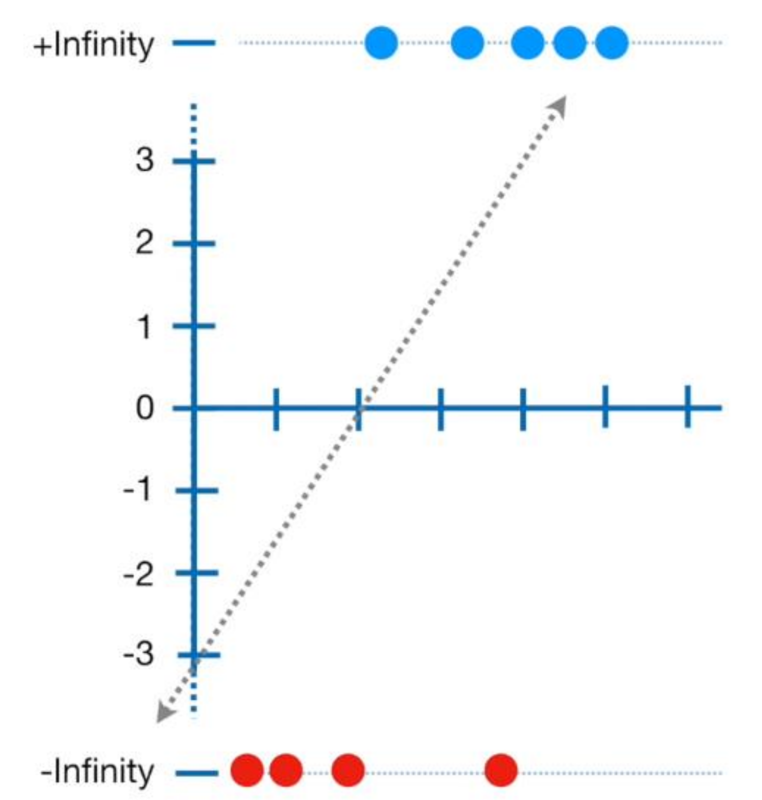
- Why do we take all the trouble doing the transformation from probability to log odds?
- One reason is that it is usually difficult to model a variable which has restricted range, such as probability.
- This transformation is an attempt to get around the restricted range problem.
- It maps probability ranging between 0 and 1 to log odds ranging from negative infinity to positive infinity.

INTERPRET ODDS RATIOS IN LOGISTIC REGRESSION

odds	Log of odds
• probability of success of some event is .8. • probability of failure 1- .8 = .2. • The odds of success are defined as the ratio of the probability of success over the probability of failure. • So, the odds of success are .8/.2 = 4 • That is to say that the odds of success are 4 to 1.	Log(p/(1-p))
the odds increase as the probability increases or vice versa	
Probability ranges from 0 and 1.	
Odds range from 0 and positive infinity.	From –infinity to +infinity

MAIN POINTS

- In linear regression, the line is fit using the values predicated by the regression function.
- Instead in log reg, the is fit using an S shape logit function
- This S curve is basically a sigmoid function.
- The curve tells the probability of a given point, that probability is used to decide the predicated class.
- Coefficients are presented using logodds function



PROS AND CONS

Pros	Cons
Convenient probability scores for observations	Doesn't perform well when feature/dimensions/columns space is too large
Efficient implementations available across tools (businesses really like Log Reg)	Doesn't handle large number of categorical features/variables well
Multi-collinearity is not really an issue	Relies on transformations for non-linear features
Wide spread industry comfort for logistic regression solutions	

LINEAR REGRESSION VS LOGISTIC REGRESSION

linear regression	logistic regression
In linear regression, the outcome (dependent variable) is continuous.	• Binary classification; • is used when the response variable is categorical in nature. E.g. yes/no, true/false, red/green
The data is modelled using a straight line.	The probability of some obtained event is represented as a linear function of a combination of predictor variables.
Linear relationship between dependent and independent variables is required	Linear relationship between dependent and independent variables is NOT required

IMPORTANT POINTS:

- Logistic regression doesn't require linear relationship between dependent and independent variables. *It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio*

- To avoid over fitting and under fitting, we should include all significant variables.

- It requires large sample sizes because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square

- The independent variables should not be correlated with each other i.e. no multi collinearity. However, we have the options to include interaction effects of categorical variables in the analysis and in the model.

- If the values of dependent variable is ordinal, then it is called as Ordinal logistic regression

- If dependent variable is multi class then it is known as Multinomial Logistic regression.

MULTINOMIAL

Sigmoid function: used in the logistic regression model for binary classification.

Softmax function: used in the logistic regression model for multiclassification.

MULTINOMIAL LOGISTIC REGRESSION

- Binary Classification:
  - Given the subject and the email text predicting, Email Spam or not.
  - Sunny or rainy day prediction, using the weather information.
  - Based on the bank customer history, Predicting whether to give the loan or not.

- Multi-Classification:
  - Given the dimensional information of the object, Identifying the shape of the object.
  - Identifying the different kinds of vehicles.
  - Based on the color intensities, Predicting the color type.

ONE VS. REST LOGISTIC REGRESSION

- On their own, logistic regressions are only binary classifiers
  - meaning they cannot handle target vectors with more than 2 classes.

- extensions to logistic regression allows to handle more than 2 classes

- In one-vs-rest logistic regression (OVR)
  - a separate model is trained for each class predicted whether an observation is that class or not (thus making it a binary classification problem).
  - It assumes that each classification problem (e.g. class 0 or not) is independent.

SIGMOID AND SOFTMAX FUNCTION

Softmax Function	Sigmoid Function
Used for multi-classification in logistic regression model.	Used for binary classification in logistic regression model.
The probabilities sum will be 1	The probabilities sum need not be 1.
Used in the different layers of neural networks.	Used as activation function while building neural networks.
The high value will have the higher probability than other values.	The high value will have the high probability but not the higher probability.

COMPARING LOGISTIC REGRESSION WITH OTHER MODELS

Advantages of logistic regression:	Disadvantages of logistic regression:
Highly interpretable, Outputs well-calibrated predicted probabilities	Presumes a linear relationship between the features and the log-odds of the response
Model training and prediction are fast	
No tuning is required (excluding regularization)	
Features don't need scaling	
Can perform well with a small number of observations	

SKLEARN.LINEAR_MODEL.LOGISTICREGRESSION

- class sklearn.linear_model.LogisticRegression(penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='warn', max_iter=100, multi_class='warn', verbose=0, warm_start=False, n_jobs=None)

- penalty : str, 'l1' or 'l2', default: 'l2'

  Used to specify the norm used in the penalization.

- C : float, default: 1.0

  Inverse of regularization strength; must be a positive float.

- solver : str, {'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'}, default: 'liblinear'.

  Algorithm to use in the optimization problem.

  For small datasets, 'liblinear' is a good choice, whereas 'sag' and 'saga' are faster for large ones.

  For multiclass problems, only 'newton-cg', 'sag', 'saga' and 'lbfgs' handle multinomial loss; 'liblinear' is limited to one-versus-rest schemes.

SKLEARN.LINEAR_MODEL.LOGISTICREGRESSION

- class sklearn.linear_model.LogisticRegression(penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='warn', max_iter=100, multi_class='warn', verbose=0, warm_start=False, n_jobs=None)

multi_class : str, {'ovr', 'multinomial', 'auto'},

default: 'ovr'

If the option chosen is 'ovr', then a binary problem is fit for each label.

solver : str, {'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'},

default : 'liblinear'.

- small datasets, 'liblinear' is a good choice,
- large datasets, 'sag' and 'saga'

multiclass problems - only 'newton-cg', 'sag', 'saga' and 'lbfgs' handle multinomial loss;

one-versus-rest schemes - 'liblinear'

L2 penalty - 'newton-cg', 'lbfgs' and 'sag'
L1 penalty - 'liblinear' and 'saga'

QS

Is Logistic regression mainly used for Regression?

A) TRUE
B) FALSE

Is it possible to apply a logistic regression algorithm on a 3-class Classification problem?

A) TRUE
B) FALSE

Which of the following methods do we use to best fit the data in Logistic Regression?

A) Least Square Error
B) Maximum Likelihood
C) Jaccard distance
D) Both A and B

QS

Which of the following evaluation metrics can not be applied in case of logistic regression output to compare with target?

A) AUC-ROC
B) Accuracy
C) Logloss
D) Mean-Squared-Error

One of the very good methods to analyze the performance of Logistic Regression is AIC, which is similar to R-Squared in Linear Regression. Which of the following is true about AIC?

A) We prefer a model with minimum AIC value
B) We prefer a model with maximum AIC value
C) Both but depend on the situation
D) None of these

QS

Standardisation of features is required before training a Logistic Regression.

A) TRUE
B) FALSE

Suppose you have been given a fair coin and you want to find out the odds of getting heads. Which of the following option is true for such a case?

A) odds will be 0
B) odds will be 0.5
C) odds will be 1
D) None of these

QS

The logit function(given as l(x)) is the log of odds function. What could be the range of logit function in the domain x=[0,1]?

A) $(-\infty, \infty)$
B) $(0,1)$
C) $(0, \infty)$
D) $(-\infty, 0)$