LOGISTIC REGRESSION

→ Binary (2 classes)
→ Multinomial (>2)
→ Ordinal (ordered class)

→ It is a machine Learning method used in any classification problem when we want to distinguish one class from another. The most simplistic case is binary classification. We can answer either question with 'yes' or 'No'.

→ Neural Networks are built on Top of it.

→ It is statistical method used for predicting the PROBABILITY of a binary outcome based on 1 or more predictor variables. It's type of regression analysis where the dependent variable is categorical (usually binary), and the independent var can be <u>continuous</u> or <u>categorical</u>.

→ It uses the logistic function (Sigmoid function) to model the relationship b/w the dependent var & independent var. It is defined as.

$$\boxed{P(Y=1|X) = \frac{1}{1+e^{-Z}}}$$

where, → $P(Y=1|X)$ is the prob of dependent var (Y) being 1 given the independent var (X).

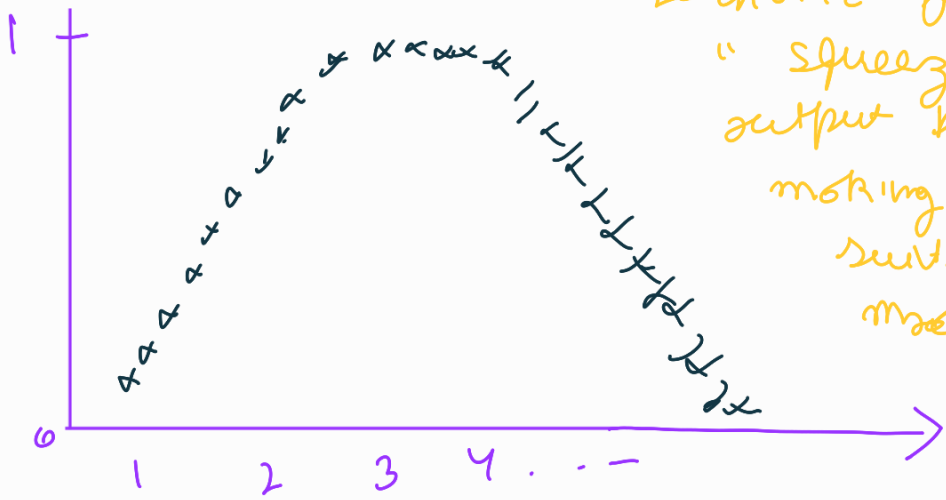
→ e is the base of natural logarithm.

→ Z is the linear combinations of the

independent vars & their coefficients

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

The coefficient β are estimated using MAXIMUM LIKELIHOOD ESTIMATOR or other optimization techniques.



LOGISTIC function "squeezes" the output b/w 0 & 1 making it suitable for modelling probabilities.

Logistic regression has a LINEAR APPROACH, but predictions are transformed using the Logistic Function. Bcz of this we can no longer understand the prediction as a Linear comb. of input variables as we can with linear regression.

Model can be stated as:

$$P(Y=1|x) = \frac{e^{(\beta_0 + \beta_1 x)}}{e^{(\beta_0 + \beta_1 x)} + 1}$$

we can change the above equation as follows:

$$\ln(P(Y=1|x)) = \underbrace{b_0 + b_1 * X}_{\text{like LINEAR regression}}$$

Q: Why do we use Logistic or Sigmoid function in Logistic Regression.

Answer: To map any real-valued input to

bounded output 0 and 1, which is suitable for representing probability. Reasons for using Logistic

1) PROBABILITY INTERPRETATION

→ helps in interpreting prob. b/w 0 & 1.

2) S-shaped curve

Logistic function has S-shaped curve, which is desirable for modeling probabilities. As the input variable changes, the output smoothly transitions from near 0 to near 1.

3) Montonicity

→ Logistic func is monotonic, means it always increases or always decreases. This property ensures that as the input variable changes, the predicted prob of the event occuring moves monotonically from 0 to 1.

4) Continuity & Differentiability:

continuous & differentiable for all real numbers. This property is essential for optimization algo to converge efficiently during parameter estimation.

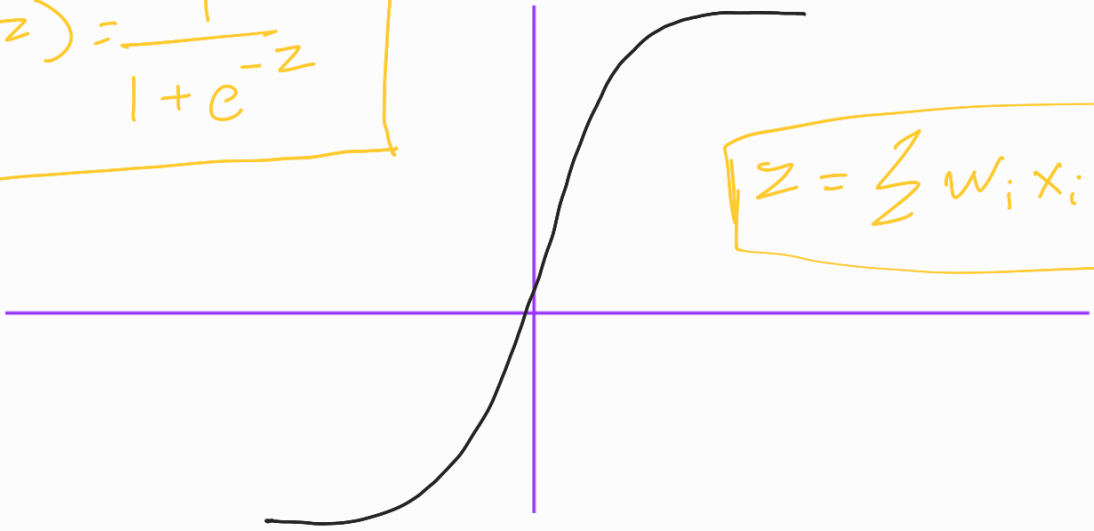
5) Easy to compute Inverse

inverse → logit function → convert prob back to odds, useful for interpreting coefficients

---

Function Plot - Sigmoid function

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



$$z = \sum w_i x_i + \text{bias}$$

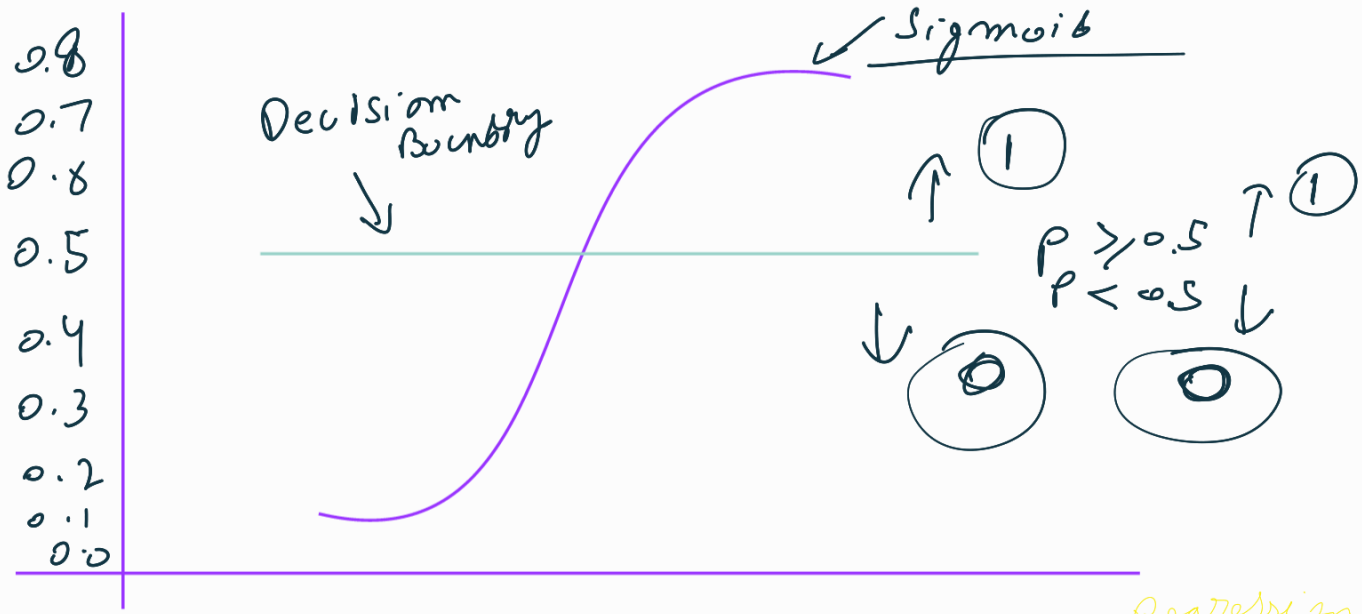
Decision Boundary

The actual application of L.R is to determine a Decision Boundary for Classification pblm.

return value → 0 or 1.

$$P \geq 0.5 \Rightarrow 1$$

$$P < 0.5 \Rightarrow 0$$



LOSS FUNCTION

(MSE, MAE, RMSE, Adj RMSE, Binary Cross entropy (for 2), Categorical cross (for multi class))

Regression loss

In Linear regression we used MSE as a

loss function.

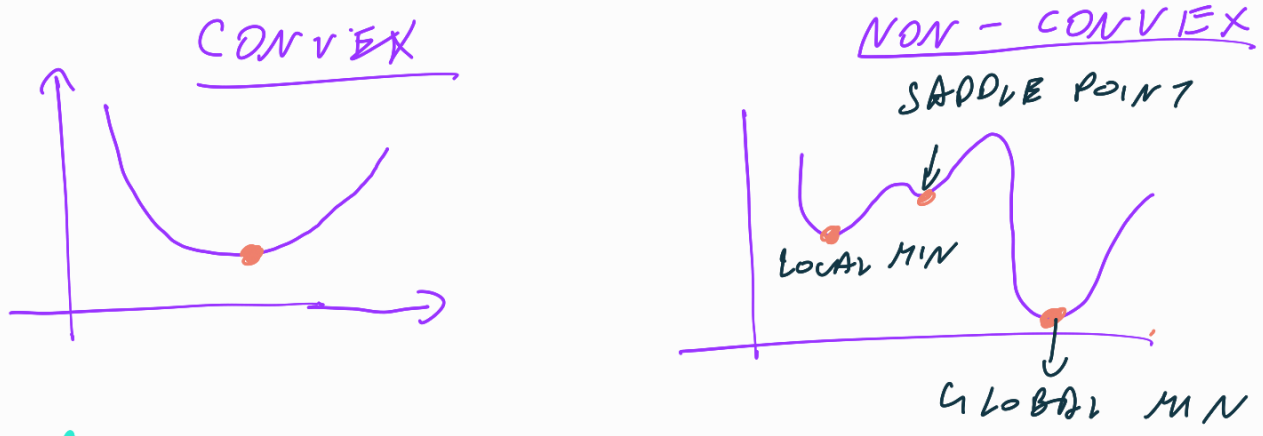
Class first a loss

Q. why MSE does not work in Logistic Reg?

Answer.

When the MSE loss function is plotted with respect to the logistic regression model weights, the curve obtained here is not a CONVEX curve, which can make very difficult to find the global minimum.

This non-convex nature of MSE in logistic regression is because the non linearity has been introduced into the model in the form of SIGMOID Function, making the relationship b/w Error & weight patterns very complex.

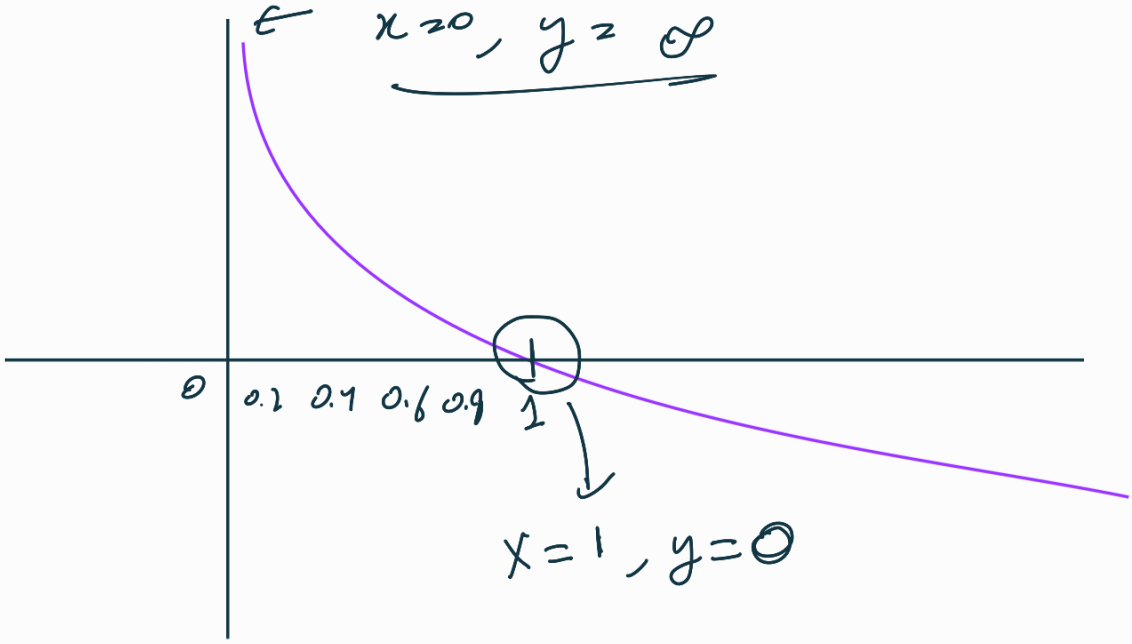


Saddle point :- where curve is upward in one direction & downward in other direction

Mathematically, it is a point where first derivative with respect to one variable is zero, but the function does not reach the local maximum or minimum at that point.

→ Loss function heavily penalize the samples which are away from the desired value.

PLOT $y = -\log(x)$



Binary Cross-Entropy Loss function / Log Loss

$$\text{log Loss} = \frac{1}{n}\sum_{i=1}^{N} \underbrace{-y\log(\hat{y})}_{} \underbrace{-(1-y)\log(1-\hat{y})}_{}$$

if actual $y = 1$ then this part is ACTIVE

if actual $y = 0$ then this part is active

So, if the label is +ve or 1 then $\rightarrow p$

" " " " -ve or 0 " $\rightarrow (1-p)$

COST FUNCTION

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

$$z = \theta^T x$$

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

To minimize, log loss we use optimization technique called GRADIENT DESCENT, update the parameter in the direction where it reduces the loss function.

STEPS for OPTIMIZATION using G.D.

1) Initialize params :-> weight & bias, randomly
2) Compute prediction -> find $\hat{y}$
3) Compute Gradient -> find gradient of log loss w.r.t (θ and b)
   weight, bias
4) Update Params :- update weight & Bias.

$$\theta_{new} = \theta_{old} - \alpha \times \text{gradient}$$

$$b_{new} = b_{old} - \alpha \times \text{gradient}$$

$\forall$ α = learning rate.

5) Repeat Step 2 - 4 :- until Convergence

---

if you compute for multiple classification then Binary cross entropy will not work, here you have to use Categorical Cross entropy.

Categorical cross entropy (for multiple classes / labels)

$$\boxed{\text{Loss} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{c} y_{ij} \log(\hat{y}_{ij})}$$

Hypothesis function

$$\boxed{h_\theta(x) = \theta_0 + \theta_1 x}$$ // $y = mx + c$

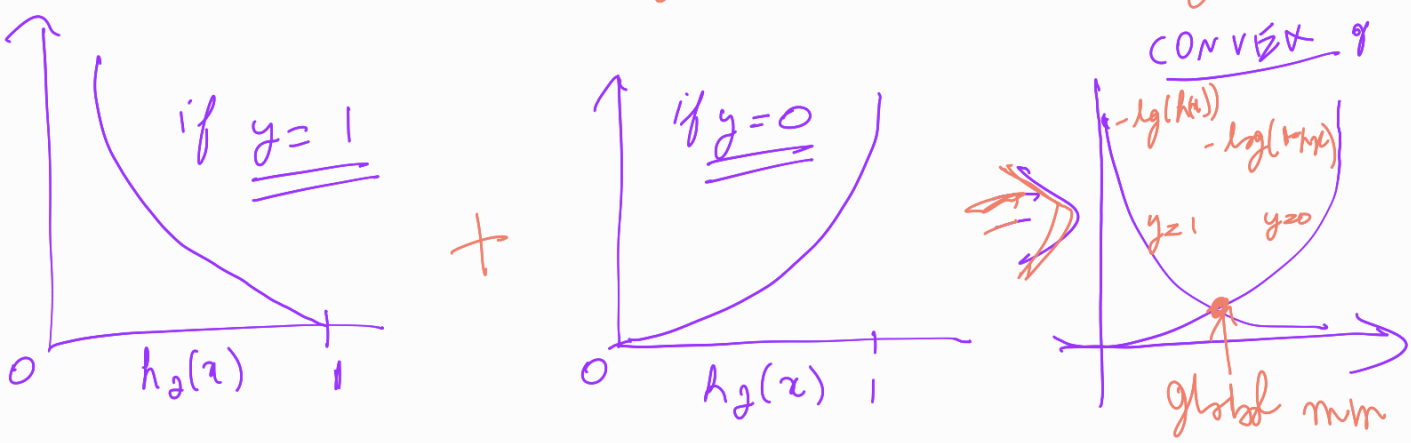
GOAL : minimize $J(\theta_0, \theta_1)$

Simplify : $h_\theta(x) = \theta_1(x)$

$J(\theta_1) \rightarrow$ function of the parameter θ_1, it controls the slope.

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

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



Cost function $\rightarrow J(\theta_0, \theta_1) \rightarrow$ Squared error function

$$\boxed{J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2}$$

Cost function for Logistic Reg

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

The entire cost function can be written as

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

GRADIENT DESCENT

To minimize the cost it is used.

→ Sigmoid derivatives are easy to calculate.

→ Iterative optimization technique to find minimum value.

GRADIENT Descent

GOAL: minimize $J(\theta_0, \theta_1)$

Random Sampling / Initialization of θ_0, θ_1

$\theta_0 = 0, \theta_1 = 0$

Keep changing θ_0 & θ_1 to reduce $J(\theta_0, \theta_1)$ until we end up at a minimum.

Repeat until convergence {

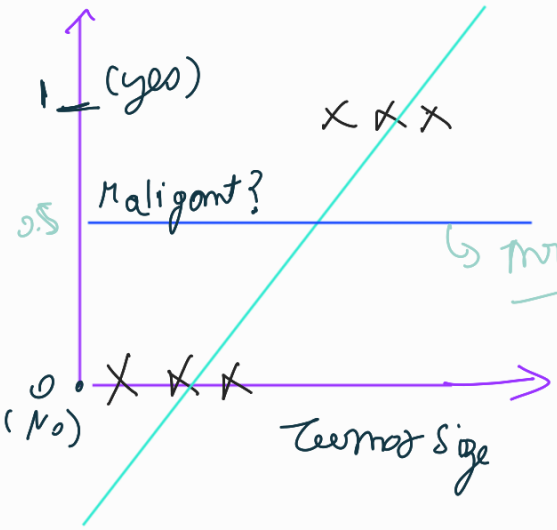
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

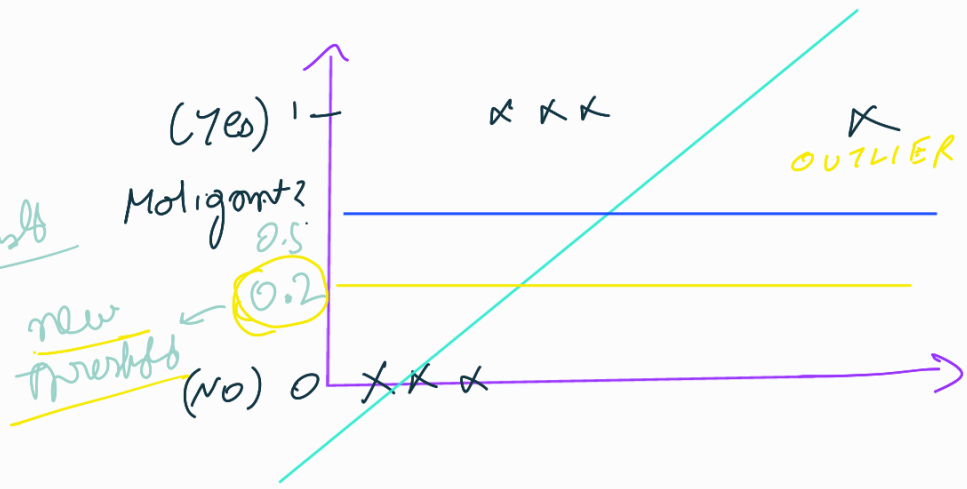
}

Q. Why do we need LOGISTIC REGRESSION, why can't we use LINEAR REGRESSION to classify Binary problems.

Ans 1) LINEAR REGRESSION is made for primarily for prediction and for CONTINUOUS DATA.

2) If we add an outlier in our dataset the best fit line in linear regression SHIFT TO FIT THAT POINT (Prone to outliers)





But in Logistic regression bcz of Sigmoid we have S-Shaped curve which is not sensitive to outliers

3) In Linear Regression result can be out of Range also i.e $\boxed{h(x) > 1 \text{ or } h(x) < 0.}$ (out of Range)

By model is primarily a regression problem

→ So logistic regression converts the straight line to an S-curve using sigmoid which ALWAYS gives value b/w 0 & 1.

Q. How logistic function squeezes the output of linear reg to 0 & 1.

Answer: Sigmoid fun $P(Y=1|x) = \frac{1}{1+e^{-z}}$

$e^{-z} \Rightarrow$ always +ve $\Rightarrow$ Denominator +ve always

$1/+ve(0 \text{ to } 1) \Rightarrow$ b/w 0 to 1.

$z \approx$ large $\rightarrow e^{-z} \approx 0 \rightarrow \frac{1}{1+e^{-z}} \approx 1$

$z \approx$ small $\rightarrow e^{-z} \approx 1 \rightarrow \frac{1}{1+e^{-z}} \approx 0$

$\boxed{z = \theta_0 + \theta_1 x}$

DERIVATION FOR SIGMOID FUNCTION

$$\boxed{P = \beta_0 + \beta_1 x}$$

$$\boxed{\frac{P}{1-P} = \beta_0 + \beta_1 x} \quad (0, +\infty)$$

Odds are nothing but the ratio of the prob. of success and prob of failure.

The problem is that range is restricted and we don't want a restricted range because if

we do so then our correlation will decrease.

By restricting the range we are actually decreasing the number of data points and of course, if we decrease our data points, our correlation will decrease. It is difficult to model a variable that has a restricted range. To control this we take the LOG OF ODDS which has range $(-\infty, \infty)$.

$$\boxed{\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x}$$

Now we want a function of p because we want to predict prob right?

$$\exp\left[\log\left(\frac{p}{1-p}\right)\right] = \exp\left(\beta_0 + \beta_1 x\right)$$

$$e^{\ln\left[\frac{p}{1-p}\right]} = e^{(\beta_0 + \beta_1 x)}$$

$$\frac{p}{1-p} = e^{(\beta_0 + \beta_1 x)}$$

$$p = e^{(\beta_0 + \beta_1 x)} - p e^{(\beta_0 + \beta_1 x)}$$

$$p = p\left[\frac{e^{(\beta_0 + \beta_1 x)}}{p} - e^{(\beta_0 + \beta_1 x)}\right]$$

$$1 = \frac{e^{(\beta_0+\beta_1 x)}}{p} - e^{(\beta_0+\beta_1 x)}$$

$$p\left[1+e^{(\beta_0+\beta_1 x)}\right] = e^{(\beta_0+\beta_1 x)}$$

$$\boxed{p = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}}$$

Now dividing by $e^{(\beta_0+\beta_1 x)}$, we will get

$$\boxed{p = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}}, \quad \text{(Sigmoid function)}$$

$$p = \frac{1}{1+e^{-z}} \quad \forall \; \boxed{z = \beta_0 + \beta_1 x}$$

(MLE)

Q. What is the use of MAXIMUM LIKELIHOOD ESTIMATOR?

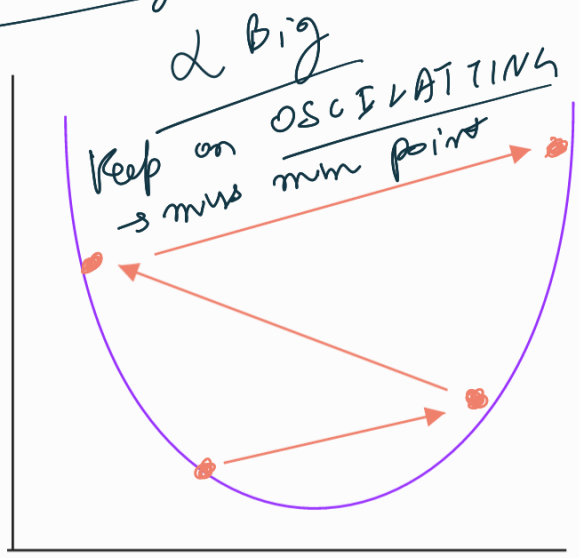
Answer. Used to identify parameter value that maximize the likelihood function. This function represents the joint probability density function (PDF) of our sample observations.

Summary: It multiplies the conditional probabilities

summary, it must pass ... for observing each example given the distribution parameters.

LOG LOSS VS DERIVED FROM MLE

$\alpha \approx$ Step Size



Cost function Derivation → (Analytic Vidhya article).

Derivative of Sigmoid function → which will be further used in COST function

$$\frac{d\,\sigma(x)}{dx} = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)$$

$$= \frac{d}{dx}\left(1+e^{-x}\right)^{-1}$$

$$= -\left(1+e^{-x}\right)^{-2} \times \frac{d}{dx}\left(1+e^{-x}\right)$$

$$= -\left(1+e^{-x}\right)^{-2} \times \left[0 + \frac{d}{dx}\left(e^{-x}\right)\right]$$

$$= -\left(1+e^{-x}\right)^{-2} \times \left[e^{-x} \times \frac{d}{dx}(-x)\right]$$

$$= -\left(1+e^{-x}\right)^{-2} \times -\left(e^{-x} \times 1\right)$$

$$= e^{-x}\left(1+e^{-x}\right)^{-2}$$

$$= \frac{e^{-x}}{\left(1+e^{-x}\right)^{2}} = \frac{e^{-x}+1-1}{\left(1+e^{-x}\right)^{2}}$$

$$= \frac{\cancel{\left(1+e^{-x}\right)}}{\left(1+e^{-x}\right)^{\cancel{2}}} - \frac{1}{\left(1+e^{-x}\right)^{2}}$$

$$= \frac{1}{\left(1+e^{-x}\right)} - \frac{1}{\left(1+e^{-x}\right)^{2}}$$

$$\boxed{\frac{d\,\sigma(x)}{dx} = \frac{1}{\left(1+e^{-x}\right)}\left[1 - \frac{1}{\left(1+e^{-x}\right)}\right]}$$

STEP 1

Deriving Cost function → using Chain Rule

$$-\frac{\partial LL(\theta)}{\partial \theta_j} = -\frac{\partial LL(\theta)}{\partial p} \cdot \frac{\partial p}{\partial \theta} \qquad \forall\, p = \sigma\left[\theta^T x\right]$$

$$= \boxed{-\frac{\partial LL(\theta)}{\partial p}} \cdot \boxed{\frac{\partial p}{\partial z}} \cdot \boxed{\frac{\partial z}{\partial \theta_j}} \qquad \forall\, z = \theta^T x$$

STEP 2: Find derivative of log likelihood w.r.t p

We know,

$$LL(\theta) = y(\log(p)) + (1-y)\log(1-p) \quad \forall p = \sigma[\theta^T x]$$

Log loss

$$\boxed{\frac{\partial LL(\theta)}{\partial p}} = \frac{y}{p} + \frac{1-y}{1-p}$$

STEP-3: Find derivative of 'p' w.r.t 'z'.

$$p = \sigma(z)$$

$$\boxed{\frac{\partial p}{\partial z}} = \frac{\partial[\sigma(z)]}{\partial z}$$

we know the derivative of sigmoid function is

$$\boxed{\sigma[\theta^T x][1-\sigma(\theta^T x)]}$$

$$\Rightarrow \boxed{\frac{\partial p}{\partial z}} = \boxed{\sigma[z][1-\sigma(z)]}$$

Step 4 :- Find derivative of z w.r.t θ

$$z = \theta^T x$$

$$\boxed{\frac{\partial z}{\partial \theta_j}} = x_j$$

Step 5 :- Put all the derivatives in equation 1

$$-\frac{\partial LL(\theta)}{\partial \theta_j} = -\frac{\partial LL(\theta)}{\partial p} \cdot \frac{\partial p}{\partial z} \cdot \frac{\partial z}{\partial \theta_j}$$

$$-\frac{\partial LL(\theta)}{\partial \theta_j} = -\left[\frac{y}{p} + \frac{(1-y)}{(1-p)}\right] \cdot \sigma(z)(1-\sigma(z)) \cdot x_j$$

$$= -\left[\frac{y}{p} + \frac{(1-y)}{(1-p)}\right] \cdot p[1-p] \cdot x_j \quad \forall\, p = \sigma[z]$$

$$= -[y(1-p) - p(1-y)] \cdot x_j$$

$$= -[y-p] \cdot x_j$$

$$= [p-y] \cdot x_j = [\sigma(\theta^T x) - y] \cdot x_j$$

Hence the derivatives of our cost function is

$$\boxed{\theta_{new} = \theta_{old} - \alpha[\sigma(\theta^T x) - y] \cdot x_j}$$

So, if slope is -ve (Downwards) → then our gradient Descent will add (+) some value to our new value → directing it towards MIN POINT.

if slope +ve (upwards) → then G.D will — some value to direct it towards MIN.

LOGISTIC REGRESSION IN SIMPLE TERMS

→ It is a statistical method for binary classification.

→ It models the relationship b/w dependent (yes/No) (0/1) and independent variables, by estimating the PROBABILITY of the BINARY OUTCOME.

→ It is uses a LOGISTIC FUNCTION to TRANSFORM LINEAR combinations of independent variables, making it suitable for predicting & classifying data into two categories based on THRESHOLD.

Q. What is better than Logistic Reg?
Ans, Random Forest & gradient Boosting are more efficient in CLASSIFICATION tasks.

Q Why it is called Logistic fun?
Ans Bcz it uses Logistic i.e Sigmoid func, which converts Linear to S-shaped curve.
And also squeeze data b/w range of 0 to 1 which is ideal for model probabilities.

Applications

1) Qualify leads :- Segment users to diff category. (MARKETING)
Help to predict which customer will convert from freemium to premium.

2) Recommend Products (CLASSIFICATION)

3) Anticipate Rare customer behaviour :- Loan defaults or churning. (FINANCE)

4) Predict prob of Heart Attack (HEALTHCARE)

Q. When to use L.R.
Ans → when your DATA is binary 0/1, T/F
→ when you need Prob result

→ when your data is LINEARLY SEPERABLE

→ when you need to understand impact of feature.

Advantage of L.R

→ Ease of use

→ Interpretability (Easy to interpret model)

→ Scalability

→ Real time predictions.

Evaluation Metrics

1.) Accuracy

2.) AUC ROC → Area under Reciver operating characterstic Curve.

describe relationship b/w True Pve rate (TPR) i.e % of sample correct predicted are RIGHT

3.) BIC AIC stats

Q. How to improve

Ans → Remove OUTLIERS, MULTI COLLINEARITY, (remove HIGHLY CORELATED FEAT)

→ ASSERT LINEAR ASSUMPTIONS i.e LOG Transform

→ ASSERT NORMAL DISTRIBUTION

→ FEATURE SCALING

→ Feature Selection

→ REGULARIZATION

→ NEWTON-RAPHSON METHOD

uses gradient & Hessian Matrix.

L1, L2 Regularization prevent OVERFITTING
Lasso, Ridge and improve GENERALIZATION

L1 Regularization → sets weight to 0.

L2 " → Reduce the predictor or lower its weights towards 0, making feature less impactful

8. Assumptions of Logistic Reg.

Answer 1) Binary outcome → Dependent var should be 0 or 1.

2) LINEARITY → Log odds of the outcomes is assumed to be in a LINEAR COMB of predictors.

3) INDEPENDENCE OF ERRORS :- LITTLE or not auto correlation in RESIDUALS

4) NO MULTI COLLINEARITY :- Independent variables should be uncorrelated with each other.

5) LARGE SAMPLE SIZE :- Large dataset for good results.

Check Assumptions by plotting RESIDUAL plots, multicollinearity.

6) Problem with extreme outliers :- Assumed not extreme outliers in data.
can be analysed by using COOK'S DISTANCE (D_i) for each observation

BIC / AIC

BIC (Bayesian Information Criterion)

AIC (Akaike Information Criterion)

These are STATISTICAL measures used for MODEL selection.

→ They help in comparing different models by balancing goodness of fit with model complexities, penalizing models that are TOO COMPLEX TO PREVENT OVERFITTING.

$$\boxed{BIC = 2 \times \log \text{likelihood} + K \times \ln(n)}$$

K → # of parameters in the model

n → sample size

BIC penalizes complex model more strongly than AIC, as it includes a term that is proportional to the # of parameters in the model

$$BIC \propto K$$

[Lower BIC ↓ → BETTER MODEL FIT]

$$\boxed{AIC = -2 \times \log \text{likelihood} + 2K}$$

AIC is less aggressive than BIC.

AIC is more imp when the emphasis is on prediction accuracy rather than the model selection

↓ Lower AIC ↑ Better model fit

How they are used in Logistic Reg.

1) **Model Selection**: Create multiple L.R models with diff parameters, compare all the models using BIC / AIC measures & select model of ↓BIC & ↓AIC.

2) **REGULARIZATION**: BIC or AIC can be used to select optimal parameter for Regularization (Lasso & Ridge) use param that minimizes BIC & AIC

3) **ASSESSMENT OF FIT**:
Use to ASSESS the fit of model
→ provide adequacy of chosen model.

Ridge Regularization (L2)

REGULARIZATION
To prevent overfitting

J = Mean squared error

$$J = \frac{\sum(\hat{y}-y)^2}{n} + \text{Sum of squares of all coefficients}$$

penalty parameter / value

starts penalizing if we move towards higher coefficients values

So our model will not increase more features or overfit the model

$$J = \frac{\sum(\hat{y}-y)^2}{n} + \frac{\lambda}{n}\sum \beta_j^2$$

→ Sum of squares of coef.

→ Strength of Regularizer to apply

If $\lambda = 0$, $J = MSE$
Standard Cost function
Similar to Linear Reg.

$\lambda \to \infty$, $\Rightarrow \beta \to 0$ High strength Regularization
Sensitivity to param gone down.

LASSO Regularization (L1 Regz)

J = MSE + Abs value of coefficients

$$J = MSE + \frac{\lambda}{n} \sum |\beta| \rightarrow$$ Sum of Abs value of coef.

it give use less # of features which are significant.

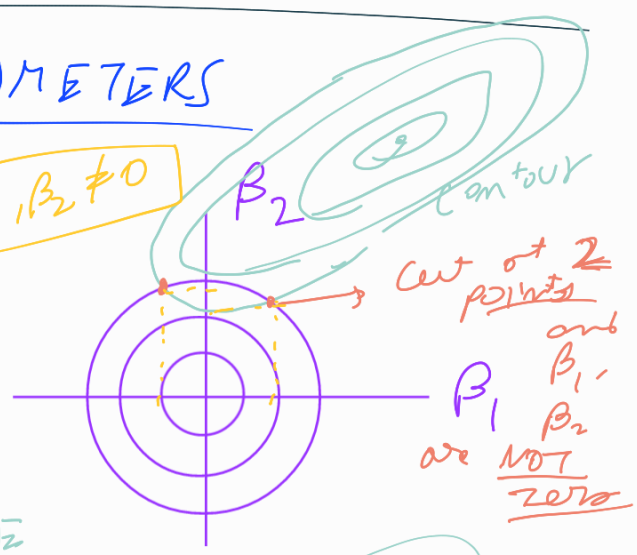
Ridge	Lasso
→ Nullified coef are still used.	→ It can make the power of coef 0,
→ It reduces but as not neglect	→ It can neglect the feature
→ Difficult to interpret	

SHRINKAGE PARAMETERS

⇒ **RIDGE**

$$\frac{\lambda}{n} \sum_{j=1}^{n} \beta_j^2 \rightarrow \beta_1^2 + \beta_2^2$$

$\beta_1, \beta_2 \neq 0$



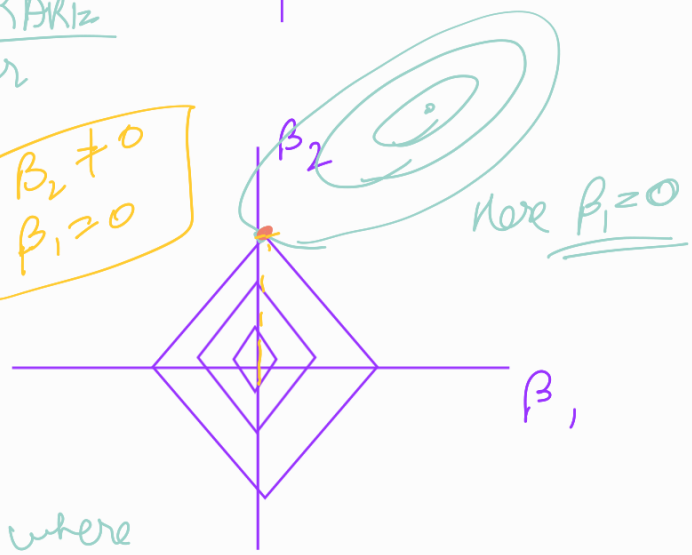
Here the point is intersect tangentially is very RARE to intersect at 0, so it never reaches 0 value.

⇒ **LASSO**

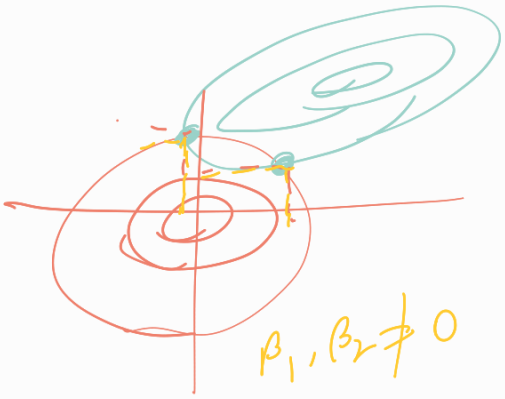
$$\frac{\lambda}{n} \sum |\beta_j| \rightarrow |\beta_1| + |\beta_2|$$

Shape Differ bcz of Abs

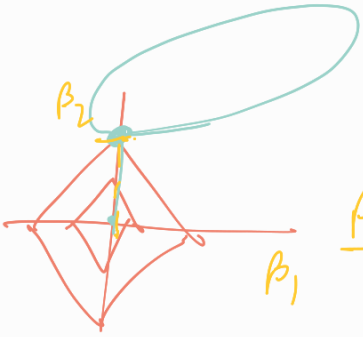
$\beta_2 \neq 0$
$\beta_1 = 0$



In LASSO you can find the intersection where one of the coefficients can be zero (0).



$\beta_1, \beta_2 \neq 0$



$\beta_2 =$ some value

$\beta_1 = 0$