Logistic Regression

→ It is also called "Classification model" and help to make predictions in cases where the output is a Categorical variable.

→ Logistic Regression is the most easily interpretable of all classification model.

Logarithmic Functions

→ log with base 10 has following property

$$6 = \log 10^{\log 6}$$

$$20 = \log 10^{\log 20}$$

$$6^x = 20$$

$$(10^{\log 6})^x = 10^{\log 20}$$

$$10^{x \log 6} = 10^{\log 20}$$

$$x \log 6 = \log 20$$

$$x = \frac{\log 20}{\log 6} \approx 1.67$$

1. Product Property

$$\boxed{\log_b ac = \log_b a + \log_b c}$$ (a, b, c positive no., $b \neq 1$)

2. Second the Quotient property

$$\boxed{\log_b \frac{a}{c} = \log_b a - \log_b c}$$ (a, b, c positive, $b \neq 1$)

3. Third the power property

$$\boxed{\log_b a^c = c \log_b a}$$

c = real no., a & b positive no., $b \neq 1$

Univariate Logistic Regression

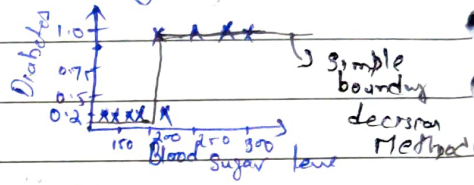
→ Binary Classification
→ Sigmoid function
→ Likelihood function
→ Building a logistic regression model in python
→ Odds and log odds

Binary Classification

→ It is a most common use of logistic regression model.

Ex:-

1. Customer will default or not
2. Spam/ ham example.



→ In this classification we have only 2 category

Note:- Simple boundary decision method would not work as some of the datapoint will become mis-classifying few data points.

→ Target variable has only 2 possible values.

Sigmoid Curve

→ Now the issue of Simple boundary decision method can be some how solved by the use of probability.

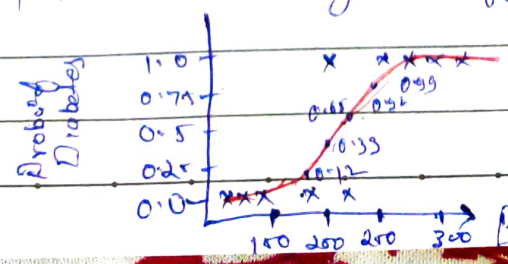
→ In method we assign low probability for the point below 150 and high probability for point above 250.

→ ~~We draw a~~ One possible curve to draw is called "Sigmoid Curve".

Eqn :-

$$Y(\text{Probability of Diabetes}) = \frac{1}{1+e^{-(B_0+B_1 x)}}$$

By changing B_0 & B_1 we can get different Sigmoid Curve.



$B_0 = -15$
$B_1 = 0.065$

M T W T F S S

Page No.:

Date: 13/01/22

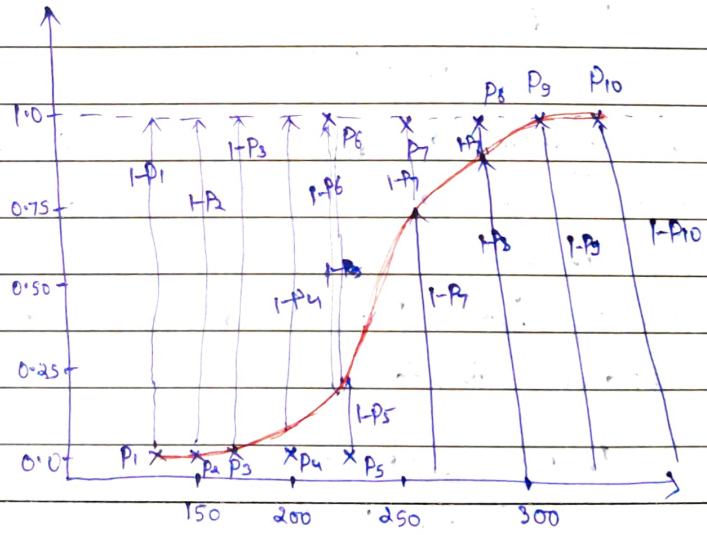
YOUVA

Note:- The main reason of not using straight line is that it is not steep enough. In the Sigmoid Curve at starting we have low probability value which raises suddenly and after that we have a lot of high values.

In the straight line the value rise from low to high in a very uniform and hence "boundary" region where the probability transition from high to low is not present.

* Finding the Best Fit Sigmoid Curve - I

→ It is a process of finding the best fit Sigmoid curve means finding B_0 & B_1 which best fits the dataset.



We need to maximise the product of these probability to find the best fitted line.

Product = $(1-P_1)(1-P_2)(1-P_3)(1-P_4)(1-P_5)(P_6)(P_7)(P_8)(P_9)(P_{10})\ldots$

↓

This product is called likelihood function.

[The $[(1-P_i)(1-P_i)\ldots$ for all non-diabetes $\ldots]$ * $[(P_i)(P_i)\ldots$ for all diabetes $\ldots]$

This process where we vary the betas until we find the best fit curve for the probability of diabetes is called logistic regression

MLE MAP

M T W T F S S
Page No.:
Date:
YOUVA

Point No.	1	2	3	4	5	6	7	8	9	10
Diabetes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes

$(1-P_1)(1-P_2)(1-P_3)(1-P_5)(1-P_7)(P_4)(P_6)(P_8)(P_9)(P_{10})$

Finding the Best Fit Sigmoid Curve - II

→ Now we can randomly choose the value of β_0 & β_1 and find the respective Probability value.
→ Now calculate the likelihood.
→ Keep on keep on doing the same until the best fit line is not found. (max value of likelihood)

* Odds and Log Odds

→ Equation of Logistic Regression

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

→ But it is difficult to understand the trend bcz P and x.
→ As we can't predict P when the value of x changes by few interval.
→ The relationship bcz P and x is too complex to see any apparent trends.

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

$\frac{P}{1-P}$ = odd ⇒ Ratio of Probability of an event occuring to the Probability of the event not occuring.

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

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

$\log\left(\frac{P}{1-P}\right)$ = Log odds.

(P = probability of diabetic 1)
(1-P = prob. of not diabetic)

Ex:- $\frac{P}{1-P} = 4$

$P = 4 - 4P$

$P = 4/5 = 0.80$

P(Diabetes) = 4 * P(No Diabetes)

Note:-
When the value of x changes linearly (by const value) then odd becomes multiple of previous odd value.

Other form of the Logistic Regression Equation

1. Sigmoid Curve (logit equation)

$$P = \frac{1}{1 + e^{-(B_0 + B_1 x)}}$$

Extremely low at start, extremely high at end & intermediate at middle

2. Probit form of logistic regression

$$P = \Phi^{-1}(B_0 + B_1 x)$$

3. Cloglog of logistic Regression

$$P = \log(-\log(1 - (B_0 + B_1 x)))$$

=> Last 2 eqn can also be used in logistic regression, as their graph also gives the same trend.

Note:-

links = sm.families.links.logit()
logm = sm.GLM(ytrain, (sm.add_constant(Xtrain)), family = sm.families.Binomial(link = links.logit)).fit().summary()

Probit → family = sm.families.Binomial(link = links.probit)) logm.fit().summary()

Cloglog → family = sm.families.Binomial(link = links.cloglog)) logm.fit().summary()

M T W T F S S

Page No.

Date: 14/01/22

YOUVA

Maximum likelihood Cost function

→ In case of logistic regression we use Sigmoid function where $P = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$

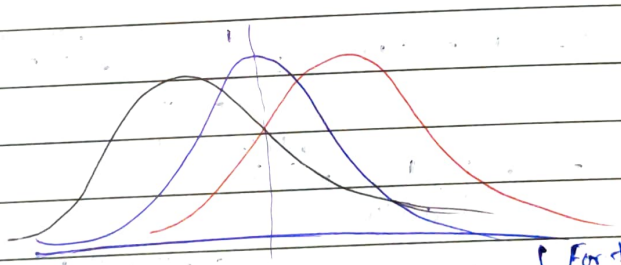
→ We need to find the optimal value of β_0 and β_1 such that the likelihood function is maximised.

→ The optimisation method used for to do so is called Maximum likelihood estimation or MLE

→ A random variable is a variable whose possible values are numerical outcomes of a phenomenon or event

→ There are 2 types of random variable

1. Discrete → head-tail, dice
2. Continuous → length, time, weigh



θ black (lower)

θ red (higher)

θ green (lower)

$P(x; \beta_0, \sigma) = P(x; \theta)$

$L(\theta) = P(x; \theta) = P(x_1, x_2, \ldots x_N; \theta)$

For the given value of θ the ~~product~~ $P(x_1), P(x_2) \ldots P(x_N)$ will have to maximum value

Assumption:

Independence $= \prod_{i=1}^{N} P(x_i; \theta) = L(\theta)$ (Product of $P(x_1)$-$P(x_N)$

$l(\theta) = \log L(\theta) = \sum_{i=1}^{N} \log P(x_i; \theta)$

Note:- Taking log is allowed because log is a monotonic function If we have a maxima or minima at $L(\theta)$, then if we will take log then $(\log L(\theta))$ will get hold at the same value of maxima & minima of also derivates also.

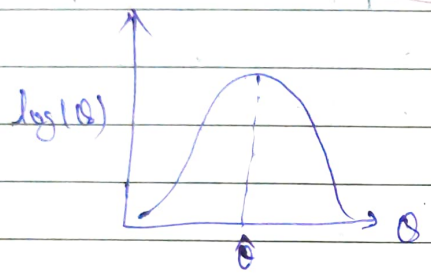
MLE For Continuous Distributions

→ MLE is basically a technique to find the parameters that maximize the likelihood of observing the data point assuming they were generated through a given distribution.

Eg:- For normal (Gaussian) distribution the parameter are the mean μ and the standard deviation σ.

Probability Density Function of normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



$$l(\theta) = \sum_{i=1}^{N} \log\left(P(x_i;\theta)\right)$$

→ Suppose this is a gaussian distribution

$$P(x_i;\theta) = P(x_i;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$\log\left(P(x_i;\theta)\right) = \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$\log\left(P(x_i;\theta)\right) = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \log\left(e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}\right)$$

$$l(\theta) = J(\mu,\sigma) = \sum_{i=1}^{N} \left[\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

Now we need to find $\hat{\mu}$ & $\hat{\sigma}$

① For iterative approach we use Gradient Descent

② Close form

$$J(\mu,\sigma) = -\sum_{i=1}^{N} \log(\sigma\sqrt{2\pi}) - \frac{(x_i-\mu)^2}{2\sigma^2}$$

constant

$$\frac{dJ}{d(\mu)} = 0 \Rightarrow \frac{d}{d\mu}\left(-\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{-1}{2\sigma^2}\sum_{i=1}^{N} 2(x_i-\mu)(-1) = 0$$

$$= \sum_{i=1}^{N} x_i = \sum_{i=1}^{N}\hat{\mu} \Rightarrow \left(\hat{\mu} = \frac{\sum x_i}{N}\right)$$

$$\hat{\mu} = \sum_{i=1}^{N}\frac{x_i}{N} \Rightarrow \text{Sample Mean}$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\hat{\mu})^2}{N}}$$

(MLE) Maximum Likelihood Estimation for Discrete Distributions

→ Let use Bernoulli Distribution for discrete distribution.

→ Implement MLE by finding the parameters of a Bernoulli Distribution.

Note:- Binary Logistic Regression is also a ~~bad~~ Bernoulli Distribution.

N observation (Discrete)

Bernoulli Distribution (p) $\{H, T\}$

$$J(p) = p(Y;\underline{p}) = p(y_1, y_2 \cdots y_N; p) = \prod_{i=1}^{N} p(y_i; p)$$

Maximise the probability of certain 'Y' vector, for the given parameter 'p'.

$p(y_1)$ = probability of y_1 and so on.

All events are mutually independent hence product is used.

$$l(p) = \log(L(p)) = \sum_{i=1}^{N} \log(p(y_i;p))$$

y_i = H or T = 1 or 0

$$P(y_i; p) = p^{y_i}(1-p)^{1-y_i}$$

• $y_i = 1$ Head $\Rightarrow (p)$
• 0 Tail $= (1-p)$

$$l(p) = \sum_{i=1}^{N} \log\left[p^{y_i}(1-p)^{1-y_i}\right]$$

$$= \sum_{i=1}^{N}\left[\log\left(p^{y_i}\right) + \log\left((1-p)^{1-y_i}\right)\right]$$

$$l(p) = \sum_{i=1}^{N}\left[y_i \log(p) + (1-y_i)\log(1-p)\right]$$

$$\frac{dl(p)}{d(p)} = 0 \Rightarrow \sum_{i=1}^{N}\left[\frac{y_i}{p} + \frac{(1-y_i)(-1)}{(1-p)}\right]$$

$$= 0 = \sum_{i=1}^{N}\frac{y_i}{p} - \sum_{i=1}^{N}\frac{(1-y_i)}{(1-p)}$$

$$\Rightarrow \sum_{i=1}^{N}\frac{y_i}{p} = \sum_{i=1}^{N}\frac{(1-y_i)}{(1-p)}$$

$$\Rightarrow \frac{1}{p}\sum_{i=1}^{N} y_i = \frac{1}{1-p}\sum_{i=1}^{N}(1-y_i)$$

• Now: $\sum_{i=1}^{N} y_i = N_H$ & $\sum_{i=1}^{N}(1-y_i) = N_T$

$$\Rightarrow \frac{N_H}{p} = \frac{N_T}{1-p}$$

$$\Rightarrow \frac{p}{(1-p)} = \frac{N_H}{N_T}$$

$$\Rightarrow \boxed{p = \frac{N_H}{N}}$$

$$\boxed{\hat{p} = \sum_{i=1}^{N}\frac{x_i}{N}}$$

Page No.: Date: YOUVA

Maximum Likelihood Estimation for Logistic Function

→ Logistic classification model is also known as logit model or logistic regression.

$$P(y;\beta) = \prod_{i=1}^{N} P(y_i;\beta)$$

$$l(\beta) = \sum_{i=1}^{N} \left[y_i \ln(\beta) + (1-y_i) \ln(1-\beta) \right]$$

$$= \sum_{i=1}^{N} y_i \ln[P(\text{Female}|x_i)] \quad \left| \beta = P(x_i) \right.$$

$$+ (1-y_i) \sum_{i=1}^{N} \ln(1 - P(\text{Female}|x_i))$$

y_i = Male/Female

x_i = length of hair

$$P(\text{Female}|x_i) = \sigma(\beta_0 + \beta_1 x_{i,1})$$

$$= \sigma(\beta^T x_i)$$

$$= \frac{1}{1+e^{-\beta^T x_i}}$$

$$P(\text{Female}|x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$$

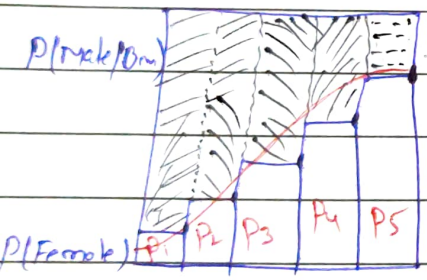
$$P(\text{Male}|x_i) = 1 - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$$

$$= \frac{1}{1+e^{\beta^T x_i}}$$

$$P(\text{Male}|x_i) = \frac{e^{-\beta^T x_i}}{1+e^{-\beta^T x_i}}$$

$P(\text{Male}|\text{Bin}=5)$

$P(\text{Female}|\text{Bin}=5)$



$$l(\beta) = \sum_{i=1}^{N} y_i \log P(\text{Female}|x_i) + (1-y_i) \log(1 - P(\text{Female}|x_i))$$

$$= \sum_{i=1}^{N} \left[\log(1 - P(\text{Female}|x_i)) + \sum_{i=1}^{N} y_i \log \frac{P(\text{Female}|x_i)}{(1 - P(\text{Female}|x_i))} \right]$$

$$= \sum_{i=1}^{N} \log \frac{1}{1+e^{\beta^T x_i}} + \sum_{i=1}^{N} y_i \beta^T x_i$$

$$l(\beta) = -\sum_{i=1}^{N} \log(1+e^{\beta^T x_i}) + \sum_{i=1}^{N} y_i \beta^T x_i \quad \rightarrow \text{Cost Function}$$

Using Gradient Descent Method to find parameter values.

$$\frac{d\,l(\beta)}{d\beta_1} = -\sum_{i=1}^{N} \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} x_{i,1} + \sum_{i=1}^{N} y_i x_{i,1}$$

$\beta^T x_i = \beta_0 \cdot 1 + \beta_1 x_{i,1}$

$x_{i,j}$
i = Sample
j = feature

$$\frac{d\,l(\beta)}{d\beta_j} = \sum_{i=1}^{N} \left(y_i - P(\text{Female}|x_i)\right) x_{i,j}$$

y_i: actual; $P(\text{Female}|x_i)$: estimated

$e_i = y_i - P(\text{Female}|x_i)$

$$= [e_1 e_2 \cdots e_N] \begin{bmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{n,j} \end{bmatrix} = [x_{1,j}, x_{2,j} \cdots x_{n,j}] \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$$\nabla_\beta l(\beta) = \begin{bmatrix} \frac{d\,l(\beta)}{d\beta_0} \\ \frac{d\,l(\beta)}{d\beta_1} \end{bmatrix} = \begin{bmatrix} x_{1,0} x_{2,0} \cdots x_{N,0} \\ x_{1,1} x_{2,1} \cdots x_{N,1} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} = X^T e$$

$X \to N \times (d+1)$

Read ⇒ Obtaining the parameter : Newton's Method
(Addition reference on "log odds & odds")

Obtaining the Parameters: Newton's Method

→ Newton-Raphson method is a second order technique.

→ By using this method we will compute the value of betas.

$\frac{d l(\beta)}{d\beta_j}$ Grad Dest: $\beta_j^{new} = \beta_j^{old} - \eta \frac{d l(\beta)}{d\beta_j}$

Newton: $\beta_j^{New} = \beta_j^{old} - H^{-1} \frac{d l(\beta)}{d\beta_j}$ (H^{-1} = Hessian matrix, it is second order derivative terms.

$$\frac{d l(\beta)}{d\beta_j} = \sum_{i=1}^{N} (y_i - P(Female|x_i)) x_{ij}$$

$P(Female|x_i) = \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}$

$$\frac{d l(\beta)}{d\beta_k d\beta_j} = -\sum_{i=1}^{N} \frac{(1+e^{\beta^T x_i}) e^{\beta^T x_i} x_{ik} - e^{\beta^T x_i} e^{\beta^T x_i} x_{ik}}{(1+e^{\beta^T x_i})^2} x_{ij}$$

$$= \sum_{i=1}^{N} P(Female|x_i) + (P(Female|x_i))^2$$

$$\boxed{\frac{d l(\beta)}{d\beta_k d\beta_j} = \sum_{i=1}^{N} (P(Female|x_i)(1 - P(Female|x_k))}$$

Multivariate Logistic Regression (Model Building)

→ When only one independent variable is not enough to get a better result then we need to add more feature.

Univariate Logistic Regression

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

Multivariate Logistic Regression

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_n x_n)}}$$

** Build a multivariate logistic regression model in Python
* Conduct feature selection for logistic regression
  * Automated methods: RFE - Recursive Feature Elimination
  * Manual Method: VIF & p-value check

Steps to Build a logistic Regression Model

Two most commonly used metrics to evaluate a Classification Model

1. Sensitivity
2. Specificity

* Sensitivity:- (True Positive Rate) (Recall)

It is defined as the ratio of correctly predicted by total no. of actual predicted positive output of a model.

Ex

Actual / Predicted	Predicted: Not Churn	Predicted: Churn
Not Churn	3289	966
Churn	595	692

From the above table we can say that the model has actually predicted 595+692 no. of Churn but out of which the correctly predicted is 692.

Actual / Predicted	Not Churn	Churn
Not Churn	True Negative	False Positive
Churn	False Negative	True Positive

$$\text{Sensitivity} = \frac{\text{No. of actual Yeses Correctly Predicted}}{\text{Total No. of actual Yeses}}$$

$$\text{Sensitivity} = 100\% \left(\frac{692}{692+595} \right) = 53.768\%$$

$$\text{Sensitivity} = \frac{TP}{FN+TP}$$

Probability of an actual 'yes' cases is predicted Correctly.

Event Rate | Imbalance data

M T W T F S S
Page No.: YOUVA
Date:

* Specificity

Specificity = No. of actual Nos Correctly Predicted / Total Number of actual Nos

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{Specificity} = \left(\frac{3269}{3269+366}\right) \times 100 = 89.931\%$$

* Accuracy

$$\text{Accuracy} = \frac{TN+TP}{FN+FP+TN+TP}$$

* Positive Rate :- Predicted +ve when it was -ve

$$\text{Positive Rate} = \frac{FP}{TN+FP}$$

(False positive Rate) (1 - Specificity)

* Positive Predictive Value (Precision)

$$\text{Positive Prediction Rate} = \frac{TP}{TP+FP}$$

True +ve / Total Predicted +ve

Probability that Predicted 'yes' is actually 'yes'

* Negative Predictive value

$$\text{Negative Prediction Rate} = \frac{TN}{TN+FN}$$

Note:- A Good model is the one in which TPR is high and FPR is low.

(Harmonic Mean)

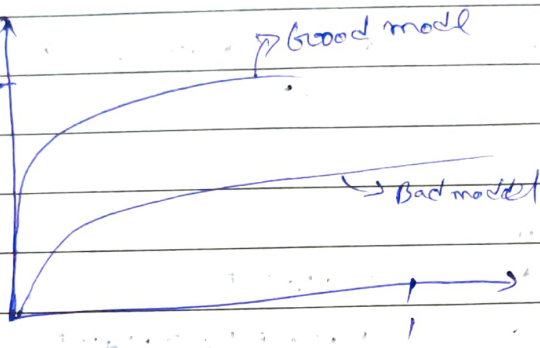
$$F = \frac{2 \cdot \text{Precision} \cdot \text{recall}}{\text{Precision} + \text{recall}}$$

High recall, low precision → optimistic Model

M T W T F S S
Page No.
Date
YOUVA

ROC Curve :
Receiver Operating Characteristics

→ Used in Radar
→ Developed by Electrical Engineer in 1950's.
→ It is used in all the classification model.
→



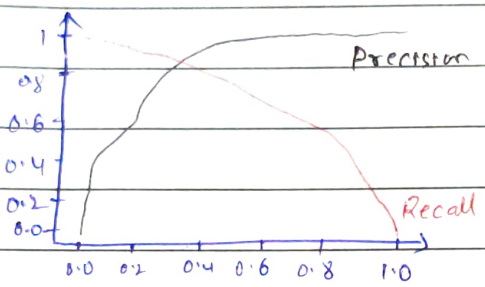
→ Good model has higher area under curve and vice versa.
→ Good ROC touches the upper left corner of the graph.
→ There is a trade off betn the True Positive Rate and False Positive Rate, or Simply, a tradeoff betn Sensitivity and Specificity.
→ When we plot TPR against FPR we get a graph which shows the trade-off betn them and this curve is known as **ROC.**

Note :- 1. If TPR increases then FPR also increases.
2. For a completely random model, the ROC Curve will pass through the ~~upper~~ 45° line.
3. When Sensitivity is increasing, (1- Specificity), and Since (1- Specificity) is increasing, it simply means that Specificity is decreasing.
4. Cutoff point is taken from X-axis

Read (FD)

M T W T F S S
Page No.:
Date:
YOUVA

→ When we draw a curve of beta accuracy, sensitivity, specificity from different value of cut-off (prob) we choose the point where all the metrics (accuracy, sensitivity & specificity) are fairly decent and almost equal.



→ AUC is near to 1 then model is working satisfactorily
→ AUC is low (around 0.5) the the model is not working properly and just guessing randomly.

→ If we want to created a balance model, without business need, then we can use an ROC curve to see the tradeoff betn sensitivity and specificity and accordingly choose an optimal trade on optimal cut off point where both these values along with accuracy are decent.

M T W T F S S

Page No.

Date 17/1/22

YOUVA

Logistic Regression

1. Types of Logistic Regression
   a. Multinomial logit
   b. Binary logit

Classification Techniques

1. Support Vector Machine
2. Neural Network
3. Random forest
4. Gradient ~~Descent~~ boosting
5. Deep Learning

Note:- Logistic regression is most common classification techniques instead of other method, because

1. Easy to understand and intuitive variable explanation In other classification techniques it becomes very complex to make understand the features of the model to the managers.
2. Linear Relationship With log of odds.

As probability increases the log of odds linearly increases and linearity is easy to understand in comparison to other complex relationship.

2. Logistic Regression Nuances

a. Sample Selection (Most important)

1. Seasonal or Cyclical fluctuations populations (Big Billion Day, Diwali)
2. Representative population (used sample base on recent Business Strategy)
3. Rare incidence population (fraud detection)

b. Segmentation.

We are buildy different models based on different nature

Read more from additional source regarding segmentation

M T W T F S S

Page No.

Date.

YOUVA

of the 6 inputs which are behaving differently which is made on different segment of HPs
→ We combine all these models to make to a single model

Note:- Building Segmentation for logistic regression is different from other Segmentation as it is because the Segmentation of logistic regression is build in order to ensure that the overall predicted power of Segmented System is higher than the predictive power of a Single model.
→ We need to make sure that the Segmentation of variable is done in such a way that the predictor are slightly different or has different power across the different Segment.

c. Variable Transformation

→ How we transform a variable before making a model.

1. Dummy Variable Transformation

Adv:- Dummy Variable becomes stable in of continuous variables makes model stable.

Disad:- Changing Continuous variables to dummies, all the data will be compressed into few categories and that might result in data clumping

2. Weight of Evidence (WOE) transformation

→ It is generally used in finance sector.

$$WOE = \log\left(\frac{\text{good in bucket}}{\text{Total Good}}\right) - \log\left(\frac{\text{bad in bucket}}{\text{Total Bad}}\right)$$

$$WOE = \log\left(\frac{\% \text{ of Good}}{\% \text{ of Bad}}\right)$$

→ Calculating woe values for the fine binning and coarse binning
→ The importance of woe for fine binning and coarse binning
→ The usage of woe transformation

Advantage of WOE

1. It reflects group identity (trend of distribution)
2. It helps in treating missing values logically for both types of variables.
3. Variable remains stable overtime.

Note:- The graph of WOE should follow increasing or decreasing trends across bins.

- If trends are not monotonic, then we need to combine the bucket/bins of that variable and then calculate the WOE again.

~~→ Disadvantage is small as dummy variable.~~

→ We may end doing some score clumping ~~doo~~ which decrease the model efficiency on small band.

→ Decrease in predictive power.

3. Information Value

→ It is an important indicator of predictive power

$$IV = WOE * \left(\frac{\text{Good in bucket}}{\text{Total Good}} - \frac{\text{Bad in Bucket}}{\text{Total Bad}} \right)$$

3. Continuous Variable Transformation Technique

1. Spline transformation
2. Interaction Variable
3. Mathematical transformation
4. Principal component transformation.

Page No.
Date: 18/01/22
YOUVA

Advantage of WOE

1. It reflects group identity (trend of distribution)
2. It helps in treating missing values logically for both types of variables.
3. Variable remains stable overtime.

Note:- The graph of WOE should follow increasing or decreasing trends across bins.

→ If trends are not monotonic, then we need to coarse the bucket/bins of that variable and then calculate the WOE again.

~~→ Disadvantage → small as dummy variable.~~

→ We may end doing some score clumping which decrease the model efficiency on small band.

→ Decrease in predictive power.

3. Information Value

→ It is an important indicator of predictive power

$$IV = WOE * \left(\frac{\text{Good in bucket}}{\text{Total Good}} - \frac{\text{Bad in Bucket}}{\text{Total Bad}} \right)$$

3. Continuous Variable Transformation Technique

1. Spline transformation
2. Interaction Variable
3. Mathematical transformation
4. Principle component transformation.

Logistic Regression

1. Types of Logistic Regression.
2. Logistic Regression Nuances
   a. Sample Selection
   b. Segmentation
   c. Variable transformation
      1. Dummy Variable transformation
      2. Weight of evidence (WOE) transformation
      3. Continuous variable transformation
      4. Interaction Variables
      5. Splines
      6. Mathematical transformation
      7. Principal component transformation

3. Model Evaluation (performance) measures
   - (a) Discrimatory Power
     - → KS Statistics
     - → Gini (Receiver operating characteristics)
     - → Rank ordering
     - → Sensitivity
     - → Specificity
   - (b) Accuracy
     - → Sensitivity
     - → Specificity
     - → Compare actual versus predicted by ...
   - (c) Stability
4. Model Validation
5. Model Tracking or model governance

Read about → KS Statistics
→ Rank ordering

Challenges in Logistic Regression:

1. Low event rate

It is a scenario in which we have very less true condition in the dataset.

Ex.:- Fraud in credit card use

Method to Fix it

(a) Use of Sampling where we create a biased sampling and increase the sampling rate and perform cross validation on it.

(b) Collect a data over wider time frame and get more of the 1's.

(c) Change the difinition of 1's. Suppose we quoise failure to increase machine failure chances.

Ex.:- Credit card default payment then in this case we can use multiple times credit card late payment as an early sign of default.

2. Missing values

We need to find that missing value is random or it's due to any pattern.

Method to be used

(a) Imputation using WOE

(b) Imputation using Mean

(c) Imputation using Median

(d) Imputation using predictive pattern

(e) Markov Chain Monte Carlo

(f) Expectation Maximisation

8. Truncated Data

Mean in this process we don't analyse how it would have behaved if accepted.

Ex:- Rejecting a credit card application based on a predefined parameter then in this case we don't know how that customer have performed after giving the credit card.

Methods:-

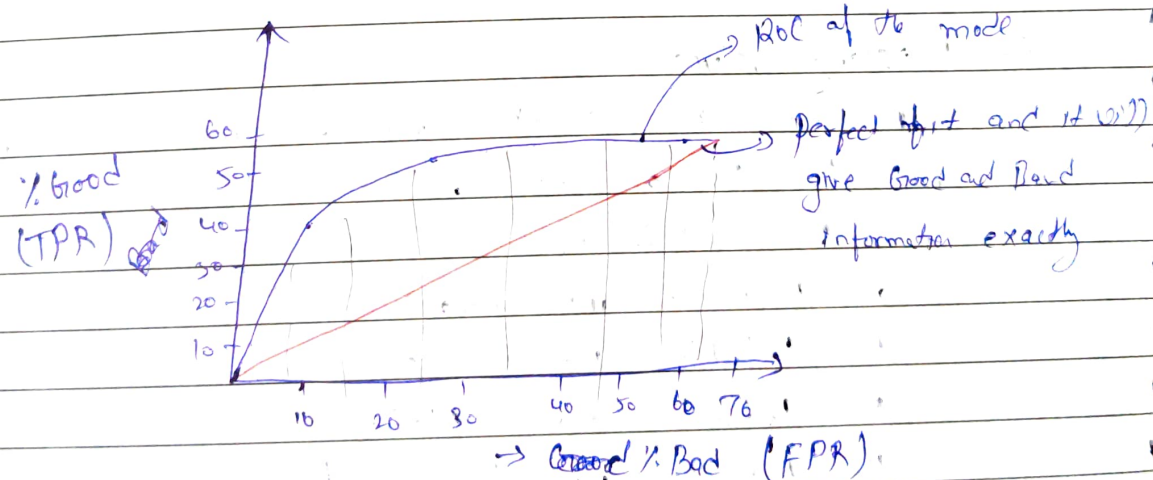
(a) Reject Inference

(b) Heatman correction Method

Model Evaluation Method.

1. Gini (AOC) (Receiver operating Characteristics)

→ It is used to find Good and Bad present in the model dataset.



→ Point on the ROC curve of a model gives us no. of Good and Bad at that particular point.

→ The area under ROC curve is called Gini of the model.

→ More steeper the ROC of the curve higher will be the Gini of the model.

Gini = 2 * Area Under ROC Curve - 1

4. Model Validation

→ Every sample dataset has it's own predictive power.
→ We split our dataset into Train and Test when ever we build any model.
→ The real test of model is when we test our model on Completely different set of data.

Method of Model Validation

1. In- Sample Validation

→ In this method we use 70% dataset for training and 30% dataset for testing the model.
→ We check wether our model gives the same performance on 30% of test dataset or not.

2. Out of time Validation

→ In this process Suppose we took a data from year 2015 to 2019 to train our model and we test our model on 2013, 2014, 2020, 2021 dataset and check how our model is working on it.

3. K-cross Validation (Bagging)

→ It is based on the Concept of Bagging and Boosting.
→ Suppose we don't have enough dataset and in this case the model can overfit by memorizing the data and its performance will fall on test dataset.
→ To fix this problem we can takes various 30% (many times) data from the overall sample and testing the model.
→ In this process it's not the same 30% dataset used for the validation but instead different set of 30% dataset and check model is working well on these diffent set of 30% dataset.
→ It helps us to understand the stability of the model across the different set of sample.
→ K-iteration mean k time different set of data will be used for train and test the model. (K-fold cross validation)

4. Stability

→ Model stability is most important part of a
→ If models are not stable the we will not be able to make right decision over-time.
→ Stability is as important than predictive power and in some cases stability is more important than predictive power.
→ Even if the predictive power of the model is not that good but need to be stable in order to work over time to take decision.

Factor on which Stability Depends

(a) Performance Stability

→ All the below factor like should have similar value on dev & test samples for a model.
- → In-Sample validation
- → Gini
- → Capture Rate
- → Sensitivity
- → Specificity

→ Result of in-sample validation approximately match those of ow-of-time validation.

(b) Variable Stability.

- → Variable distribution Stability

→ If a data set has 8 feature then all the feature should be present on train and test dataset.
- → Population Stability index (PSI)

→ Percentage of distribution of feature in the train and test dataset should be nearby.
→ The distribution of amount of feature should be stable beta train and test dataset.

Good Model

1. The sample used for model building hasn't changed too much and has the same general characteristics.

(C) Predictive Pattern

→ While calculating the WOE for different group we need to check the WOE of different groups on test and train set should be same.

Note:- 1. Sign of WOE should not change on test dataset

2. We calculate the PSI of WOE and if change in PSI of a WOE of group is lev than 0.1 on test set then we say it is Stable and if it is more than 0.25 it is not ~~accepted~~ Stable.

Note:- A Good model will be Stable when

1. Performance Stability:-
   Results of in-sample validation approximetly match those of out-of-time validation.
2. Variable Stability:-
   The sample used for model building hasn't changed too much and has the same general characteristics.

5. Model Tracking or Model Governance

→ We track the performance of a model over the time with new dataset to check whether our model is work fine or not.

→ If the performance decrease by sligh then we can do "recalibration" else we will have to build a model.

6. Model Recalibration

→ In this method we don't build the model again we just update the coefficient of the variables present in the model.

→ It is used when the Gini of the model is decrease by 0.1 or below and

→ If over the time Gini decreases by 0.25 then we need to rebuild the model.

GTL - ML - C34 - CC - G04

* Closed Form Soln.
→ It is a equation where we put value of β_0, β_1 and find value of y.

$$w^* = (X^T X)^{-1} X^T y$$

→ Inverse (X^T) is not always possible
→ N-Cube operation it takes time.

* Why Approximation
→ Because it is not possible to determine w using closed form soln.

$$E(w) = \frac{1}{2} \sum_{i=1}^{N} [f(x;w) - y]^2$$

~~Find~~ min E(w)

* Logistic regression brings y value from $-\infty$ to $+\infty$ between 0 to 1 and for that we used Sigmoid funtn.

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

* Accuracy = $\frac{\#\text{ Correct Prediction}}{\#\text{ Total Predictions}}$

Is accuracy a good metric? No.

M T W T F S S
Page No.
Date 23/11/21
YOUVA

Q1. Why can't linear regression is used in place of logistic Regression for binary classification? (BC)

→ 1. Distribution of Error term
- The distribution of data in the case of linear and logistic regression is different
- Linear regression assumes that error terms are normally distributed.
→ In case of binary classification, this assumption does not hold true.

2. Model Output
- In linear regression the output is a continuous.
→ In BC, linear regression may predict the value that can go beyond 0 and 1.
- If we want our output as probability then its output range is restricted to 0 to 1 and based on it the model output will be 0 and 1.

3. Variance of Residuals Errors.
- Linear regression assume that the variance of random error is constant and this assumption is also violated in case of logistic regression.

Q2. What are the outputs of logistic model and logistic funtn?

→ Logistic model output the logits. i.e log odds; and the logistic function output the probability.

$$\text{Logistic Model} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

$$\text{Logistic funtn} = f(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n)}}$$

Q3. How to interpret the result of a logistic regression model? Or, what are the meaning of the different betas in a logistic regression model.

→ B_0 is the baseline in a logistic regression Model.

→ It is log odds for an instance when all the attributes $(X_1, X_2, X_3, \ldots X_n)$ are zero.

→ B_0 is the log odds for an instance when none of the attributes is taken into consideration.

→ All the other Betas are the values by which the log odds change by a unit change in a particular attribute by keeping all other attributes fixed or unchanged (control variables).

Q4. What is odd ratio?

→ $OR_{x_1, x_0} = e^{\sum_{i=1}^{k} B_i (X_{1i} - X_{0i})}$

- X_1 and X_0 stand for two different groups for which the odds ratio needs to be calculated.
- X_{1i} stands for the instance 'i' in the group X_1.
- X_{0i} stands for the instance 'i' in the group X_0.
- B_0 stands for the coefficient of the logistic regression model
- Baseline is not included in the formula

Q5. What is the Maximum likelihood Estimator (MLE)?

→ It choses those set of unknown parameters (estimator) that maximise the likelihood function.

→ If find the MLE is to use calculus and setting the derivates of the logistic function with respect to an unknown parameter to zero and solving gives the MLE.

Note:- MLE and ordinary square estimation gives the same results for linear regression if the dependent variable is assumed to be normally distributed. MLE doesn't assume anything about independent variable

Q6. What are the different method of MLE and When each method preferred.

-> In case of logistic regression there are 2 approaches to MLE.

1. Unconditional Method
2. Conditional Method.

These are algorithms that we different likelihood funto.

Conditional Method	Unconditional Method
1. It don't estimate unwanted	1. It estimates the values of unwanted parameter also.
2. It can't be done with of joint probability.	2. It can directly be developed with joint probabilities.
3. It gives unbiased in such case.	3. If the no. of parameter is higher relative to the no. of instances, then it gives biased result.
4. Statisticians suggest that Conditional MLE is used when in doubt.	

Q.7. What is the Output of MLE program.

-> Maximised likelihood value.

-> Estimated likelihood value.

Q8. Why can't we use Mean Square Error (MSE) as a cost funtn for logistic regression.

→ In logistic regression we use Sigmoid funtn and perform a non-linear transformation to obtain the probabilities.

→ Squaring this non-linear transformation will lead to non-convexity with local minimums.

→ Finding the global minimum in such cases using gradient descent is not possible.

→ Due to this reason MSE is not suitable for logistic regression.

→ Cross-entropy or log loss is used as a cost funtn for logistic regression.

→ In cost funtn for logistic regression, the confident wrong predictions are penalised heavily and confident right predictions are rewarded less.

→ By optimising this cost funtn, convergence is achieved.

Note:- Accuracy gives equal importance to both False positive and False negatives which is not required in many cases.

M T W T F S S

Page No.

Date 24/1/22

YOUVA

Cost Function

* Cost Function
* Min/Max of a Cost Functo
* Differentiation of a funto
* Types of Minimisation
  - Constrained
  - Unconstrained
* Solution of Unconstrained minimisation
  - Closed Form
  - Gradient Descent

Cost Functions

-> It helps us to reach the optimal solution.

Rule of Differentiation

$y = ax^b$

$\frac{dy}{dx} = bax^{b-1}$

$y = \sin x$

$\frac{dy}{dx} = \cos x$

$y = \cos x$

$\frac{dy}{dx} = -\sin x$

$y = e^x$

$\frac{dy}{dx} = e^x$

$y = \ln x$

$y' = \frac{1}{x}$; for $x > 0$

Ex:- ① $y = 10x^2 \Rightarrow dy/dx = 2 * 10x^1 = 20x$

Slope at $x = 4 \Rightarrow \frac{dy}{dx} = 20 \times 4 = 80$

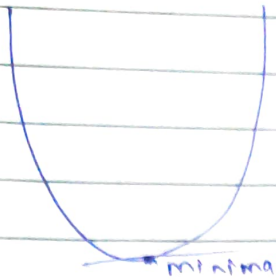
② $y = 10 = 10x^0$ ($x^0 = 1$)

$\frac{dy}{dx} = 0 \times 10x^{0-1} = 0$

③ $y = 2x^2 + 3x + 4$

$\frac{dy}{dx} = 2 * 2x^{2-1} + 1 * 3x^{1-1} + 0$

$\frac{dy}{dx} = 4x + 3$



x^2 $\Rightarrow$ $y = x^2$

$\frac{dy}{dx} = 2x = 0$ (Minima)

$2x = 0$

$x = 0$

$y = \theta_1^2 + \theta^2$

$\frac{dy}{d\theta_1} = 2\theta_1 = 0$

$2\theta_1 = 0$

$\theta_1 = 0$

$y = \theta_1^2 + \theta_2^2$

$\frac{dy}{d\theta_2} = 2\theta_2 = 0$

$2\theta_2 = 0$

$\theta_2 = 0$

* $J(m,c) = [y_1 - (mx_1 + c)]^2 + \cdots$

$\frac{dJ}{dm} = 0$

$\frac{d[y_1 - (mx_1 + c)]^2 + \cdots}{dm}$

$2[y_1 - (mx_1 + c)](-x_1)$

$= 0$

$\frac{dJ}{dc} = 0$

$d(y_1 - (mx_1 + c)^2)(-1) + \cdots / dc$

$2(y_1 - (mx_1 + c)^2(-1)$

$= 0$

Note :- To minimise a funto

$[y' = 0$ and $y'' > 0]$

To maximise a funto

$[y' = 0$ and $y'' < 0]$

→ For Straight line we find the sum of squared Errors and find the optimal value of 'm' & 'c' by differentiating the eqn of straight line w.r.t to 'm' & 'c'

Naive Bayes

→ It is another type of Supervised Classification.
→ It is a probabilistic classifier which returns the probability of a test point belonging to a class rather than the label of the test point.
→ It is a type of "Name Based" classifier.
Ex:- Spam or Ham
→ It is based on Bayes' Theorem

Conditional Probability

$$P(\text{Head}) = \frac{\#\text{ Favourable Outcomes}}{\#\text{ Total Outcomes}}$$

$$\begin{cases} P(\text{Play}) = \frac{20}{30} = 66\% \\ P(\text{Play} \mid \text{Rain}) = \frac{1}{10} = 10\% \end{cases}$$

$$P(\text{Spam}) = \frac{1}{5} = 20\%$$

$$P(\text{Ham}) = \frac{4}{5} = 80\%$$

$$P(\text{Spam} \mid \text{Word} = \text{VIAGRA}) = 70\%$$

Prior

$P(\text{Play}) = 66\%$
$P(\text{Spam}) = 20\%$

Posterior

$P(\text{Play} \mid \text{Rain}) = 10\%$
$P(\text{Spam} \mid \text{Word} = \text{Viagra}) = 70\%$

Conditional Probability

A and B are events
$P(A)$ = Probability of an email being spam
$P(B)$ = Probability of the word being VIAGARA
$P(A) = 20\%$
$P(A \mid B) = 70\%$
↳ to understand the relative probabilities.

Naive Bayes:

→ It is another type of Supervised Classification.
→ It is a probabilistic classifier which returns the probability of a test point belonging to a class rather than the label of the test point.
→ It is a type of "Name Based" classifier.
Ex:- Spam or Ham
→ It is based on Bayes' Theorem

Conditional Probability

$$P(\text{Head}) = \frac{\#\text{ Favourable Outcomes}}{\#\text{ Total Outcomes}}$$

$$P(\text{Play}) = \frac{20}{30} = 66\%$$

$$P(\text{Play} | \text{Rain}) = \frac{1}{10} = 10\%$$

$$P(\text{Spam}) = \frac{1}{5} = 20\%$$

$$P(\text{Ham}) = \frac{4}{5} = 80\%$$

$$P(\text{Spam} | \text{Word} = \text{VIAGRA}) = 70\%$$

Prior
$P(\text{Play}) = 66\%$
$P(\text{Spam}) = 20\%$

Posterior
$P(\text{Play} | \text{Rain}) = 10\%$
$P(\text{Spam} | \text{Word} = \text{Viagra}) = 70\%$

Conditional Probability

A and B are events
$P(A)$ = Probability of an email being Spam
$P(B)$ = Probability of the word being VIAGARA
$P(A) = 20\%$ $P(A|B) = 70\%$

→ It is need to understand the relative probabilities.

Bayesian estimate Vs Maximum likelihood estimate

M T W T F S S

Page No.:

Date: 26/1/22

YOUVA

Bayes' Theorem

	Indra Win (A)	(A) Indra lose	Total
(B) Sachin Scored a Century	10	2	12
Sachin Doesn't Scored a Century (B)	50	38	88
Total	60	40	100

Prior

$P(A) = \frac{60}{100}$ $P(B) = \frac{12}{100}$

Joint Probability

$P(A,B) =$ P(Indra Win and Sachin's Century)

$P(A \cap B) = \frac{10}{100} = P(B \cap A)$

Conditional Probability

$P(A|B) = \frac{10}{12} = \frac{10/100}{12/100} = \frac{P(A \cap B)}{P(B)}$ $\begin{cases} \cap = \text{intersection} \\ \cap = \text{and} \end{cases}$

$P(A \cap B) = P(A|B) \cdot P(B)$

$= \frac{10}{12} \times \frac{12}{100} = \frac{10}{100}$

$P(A \cap B) = P(B|A) \cdot P(A) = \frac{10}{60} \times \frac{60}{100} = \frac{10}{100}$

$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$

$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

→ Bayes Theorem

A and B are independent

$P(A \text{ and } B) = P(A|B) * P(B) = P(A) * P(B)$

Naive Bayes - With One Feature

→ It is a probability classifier.
→ It uses probability as a ~~cate~~ criteria to classify.
→ Common use is Spam & Ham filter.

Ex:-

Types of Mushroom	Cap-Shape	Cap Surface	Cap Color	Bruises	Odor
Edible	Convex	Smooth	White	Bruises	Almond
"	"	"	"	"	Anise
"	Bell	Scaly	"	"	Almond
"	"	"	yellow	"	Anise
Poisonous	Convex	Smooth	Brown	"	Pungent
Edible	"	"	Gray	No	None
Poisonous	"	Fibrous	"	"	Foul
"	"	Scaly	"	"	"
"	Flat	Smooth	White	Bruises	"
"	Flat	Scaly	Pink	"	None
Edible	Convex	Fibrous	Brown	No	"
Poisonous	Bell	Scaly	yellow	NO	"
Edible	Bell	Fibrous	White	No	"
Poisonous	Knobbed	Scaly	Cinnamo	No	Musty

$$P(C = Edible / x = Convex) = \frac{P(x = Convex / C = Edible) * P(C = Edible)}{P(x)}$$

$$P(C = poisonous / x = Convex) = \frac{P(x = Convex / C = Poisonous) * P(C = Poisonous)}{P(x)}$$

$P(x)$ is a common factor so nee not be considered

* $P(Edible) = 7/14 = 0.5$
* $P(Poisonous) = 7/14 = 0.5$

$P(\text{Cap-Shape} = \text{Convex} / \text{edible} = \text{Yes}) = 4/7$

$P(\text{Cap-Shape} = \text{Convex} / \text{edible} = \text{No}) = 3/7$

$$P(\text{edible} = \text{Yes} \mid x = \text{Convex}) = \frac{P(x = \text{Convex} / \text{edible} = \text{Yes}) * P(\text{edible} = \text{Yes})}{P(x)}$$

$$= \frac{4/7 \times 1/2}{P(x)} = \frac{4}{14 * P(x)}$$

$$P(\text{edible} = \text{No} \mid x = \text{Convex}) = \frac{P(x = \text{Convex} / \text{edible} = \text{No}) * P(\text{edible} = \text{No})}{P(x)}$$

$$= \frac{3/7 \times 1/2}{P(x)} = \frac{3}{14 * P(x)}$$

Note:- $P(x)$ is common and acts as a scaling factor and hence can be removed by comparing the two classes

$$\frac{4}{14 * P(x)} > \frac{3}{14 * P(x)} \Rightarrow \frac{4}{14} > \frac{3}{14}$$

$\Rightarrow P(C_i / x) = \frac{P(x / C_i)\, P(C_i)}{P(x)}$. C_i = Classes

X = Features of the data point

→ Probability is simply calculated by counting the no. of instances / occurance for categorical data.

→ The class assigned to the new test point is the class for which $P(C_i / x)$ is greater.

→ Naive theorem would properly in text classification

Conditional Independence in Naive Bayes

→ When we have multiple feature then we multiply each feature with Class of different feature.

$P(\text{edible} \mid x = (\text{convex}, \text{smooth}))$

$\propto P(x = (\text{convex}, \text{smooth}) \mid \text{edible}) * P(\text{edible})$

$= P(\text{smooth} \mid \text{edible}) * P(\text{convex} \mid \text{edible}) * P(\text{edible})$

Note:- Convex and Smooth are independent events.

→ When we divide the equation on separate feature like below is called Naive assumption and hence called Naive Bayes.

$$P(C \mid x(x_1, x_2 \cdots)) = P(x_1 \mid C) * P(x_2 \mid C) \cdots * P(C)$$

→ In the word Naive Bayes follows an assumption that the variable are conditionally independent.

Ex:- $P(A \text{ and } B \mid C)$

If $P(A \mid C)$ is same for all values of B.

$P(B \mid C)$ is same for all values of A

then there is Conditional independence bet A and B given C.

This is when $P(A \text{ and } B \mid C) = P(A \mid C) * P(B \mid C)$ imply that A is not conditioned on B or vice versa

When we divide the numerator with denominator in Bayes theorem then we get Base probability

M T W T F S S
Page No.:
Date:
YOUVA

Deciphering Naive Bayes

* $P(edible|x) = P(x|edible) * P(edible)$
* $P(poisonous|x) = P(x|poisonous) * P(poisonous)$ } Bayes Theorem

$P(edible|x) > P(poisonous|x)$

If $P(C_1|x) > P(C_2|x)$

→ x is classified as C_1

→ This is called "Maximum Aposteriori Classification Rule" (MAP)

* $P(C_i|x) = \frac{P(x|C_i) * P(C_i)}{P(x)}$

$P(C_i)$ = Prior probability $[P(C_i=0), P(C_i=1)]$

→ It is defined as the probability of occurance of an event before collection of new data or feature is called as Prior

→ Prior probability highly influences the class of new test point

$P(x|C_i)$ $[P(x|C_i=1), P(x|C_i=0)]$

↓

Likelihood Function

→ It maximises probability of observing data

Ex:-

$P(x=convex|edible) = 0.85$

If edible is yes, then 85% chance mushroom is convex.

→ It tells the likelihood of a point occurring in a class.

→ The conditional independence assumption is leveraged while computing the likelihood probability.

Note:- 1. If prior is neutral (50%) then the likelihood may largely decided the outcome.

2. If prior is too powerful, then likelihood often barely affect the result.

$P(x_1, x_2 \ldots | C_i) = P(x_1|C_i=0) \times P(x_2|C_i=0) \cdots \times P(C_i=0)$

Likelihood Calculation

Page No.:
Date:
YOUVA

$P(C_i|x)$ [$P(C_i=0|x)$, $P(C_i=1|x)$]

Posterior Probability

→ It is calculated after incorporating the feature of the data point.

→ It combines prior belief and case-specific information.

→ It is balanced outcome of the prior and the likelihood

Naive Bayes Assumption

1. Variables are independent given class
2. Or variables are conditionally independent
3. Decreases Computational time of algorithm.

Q. How point 3 happens?

M T W T F S S

Page No.: YOUVA

Date: 27/1/22

Introduction: Naive Bayes for Text Classification

→ It is widely used for
- Document Classification
- Natural language processing (NLP)
- Sentiment Analysis
- Spam, Ham Classification

Document Classifier ⇐ Pre Processing Steps

Test Dataset

Document	Feature Represent →	Class
0. upgrad is a education institute		Education
1. Educational greatness depends on ethics		Education
2. A story of great ethics and educational greatness		Education
3. Sholay is a great cinema		Cinema
4. Good movie depends on good story		Cinema

Dictionary before stop word removal

0: and
1: Upgrad
2: is
3. a
4. institute
5. of
6. Educational
7. Cinema
8. Movie
9. Story
10. good

Stop Words

0: and
2: is
3: a
5: of

Dictionary After stop word removal

0: Upgrad
1: Institute
2: Educational
3: Cinema
4: Move
5: Story
6: Good

Dictionary / vocabulary

Document	Cinema	depends	educational	ethics	good	great	upgrad	movie	
1	0	0	0	1	0	0	1	0	
2	1	0	1	1	0	0	1	0	
3	2	0	0	1	1	0	1	1	14
4	3	1	0	0	0	0	1	0	
	4	0	1	0	0	2	0	0	

Bag of Word Representation

The sentences are broken down into words and the ordering doesn't matter anymore as if it is put in a bag and shuffled.

$D^{education}$ = Cinema, ... upgrad, movie

$$\begin{pmatrix} 0,0,1,0,0,1,0,1 \\ 0,1,1,1,0,0,1,0,0 \\ 0,0,1,1,0,1,1,0,0 \end{pmatrix}$$

Prior
$P(E) = 3/5$
$P(Cinema) = 2/5$

$$D^{Cinema} = \begin{pmatrix} 1,0,0,0,1,0,0,1 \\ 0,1,0,0,2,0,0,1 \end{pmatrix}$$ 7

$P(Edu | w_1, w_2 \dots w_n)$ } Posterior
$P(Cin | w_1, w_2 \dots w_n)$ }

$$P(Cinema | w_1, w_2, \dots w_n) = \frac{P(w_1, w_2, \dots | Cinema)}{P(w_1, w_2, \dots w_n)}$$

$$= \frac{P(w_1|C_i)\, P(w_2|C_i) \dots P(w_n|C_i)}{P(w_1, w_2, \dots w_n)}$$

		Cinema	depends	educational	ethics	good	great	upgrad	movie
Cinema	2/7	1	0	0	3	0	0	2	
	1/11	1/11	1/11	1/11	1/11	1/11	1/11	1/11	
Edu	3/7	2	1	1	0	1	1	0	
	1/11	2/11	1/11	1/11	1/11	1/11	1/11	1/11	

Ex:- Text Document = "Great Story"

$P(Ed|Text\ Doc) = P(Ed|Great)\, P(Ed|Story)\, P(Edu)$

$= (3/11) \times (1/11) \times 3/5$

$P(Cin|Text\ Doc) = 0 \times 1/7 \times 2/5$

Text Doc = "Great great Story"

$P(Ed|TD) = (3/11)^2 (1/11)(3/5)$

$P(C|TD) = (0)^2 \propto (1/7)(2/5)$

Laplace Smoothing

→ Some time we see "zero probability problem" - probability of a word which has never appeared in a class. (it may appeard in the dataset of another class) is 0.

→ This issue can be solved using "Laplace Smoothing"

→ We need to ignore the word which doesn't belong to any class as a feature.

Ex:-

Text Document = "very good great Educational"
(ignored)

P(Edu | Text Docum) = 0/11 × 3/11 × 3/11

P(Cinema | Text Docum) = 3/7 × 0 × 0

→ Word "good" doesn't appear for education so we add 1 to all the feature of both the class so total atleast it appear 1 time.

	reduced	P(w/c)	n cinema	P(w/c) = Cm
Cinema	0+1	(1/(1+8)) = 1/19	1+1 = 2	2/7+8 = 2/15
Story	1+1	= 2/19	1+1 = 2	= 3/15
Education	3+1	= 4/19	0+1 = 1	= 1/15
ethics	2+1	= 3/19	0+1 = 1	= 1/15
good	0+1	= 1/19	3+1 = 4	= 4/15
great	3+1	= 4/19	0+1 = 1	= 1/15
unfgood	0+1	= 1/19	0+1 = 1	= 1/15
Movie	1+1	= 2/19	2+1 = 3	= 3/15

~~Similarly we can do it again as we don't have it~~

P(e | Text Doc) = 1/19 × 4/19 × 4/19

P(Cm | Text Doc) = 4/15 × 1/15 × 1/15

$$\frac{a}{a} \qquad \boxed{\frac{a+}{|v|+|w|*a}}$$

Quick Introduction to Bernoulli Naive Bayes

→ Prior to this we had gone through "multinomial way of classifying documents".

→ Now we will go through "Bernoulli Naive Bayes"

Bernoulli Theorem : Introduction

Oneone - - - ... ubgrad

0 0 1 0 0 1 0 1 0 0 0 1
D = 0 1 1 1 0 0 1 0 0 0 0 0
0 0 1 1 0 1 1 0 0 0 1 0
1 0 0 0 0 1 0 0 0 1 0 0
0 1 0 0 2 0 0 0 1 0 1 0 . 5th Document.

↑ This word appear 2 times in the 5th document

→ In "Bernoulli's Naive Bayes" we don't care about how many times a word occurs, but we care about whether a word occurs or not.

→ Now the 5th Document will be replaced by following value and other field will remains same.

5th Doc = 0 1 0 0 (1) 0 0 0 1 0 1 0
↳ Become 1 from 2.

Note:- Bernoulli Naive Bayes is concered only with whether the word is present or not in a document, whereas the Multinomial Naive Bayes counts the no. of occurance of the words as well.

M T W T F S S
Page No
Date 29/01/22
YOUVA

ML-CO34-SG12 004

Linear Regression

$x^{(i)}$ → input features

$y^{(i)}$ → o/p target variable

Train = $\{(x^{(i)}, y^{(i)}) \; i = 1, \ldots N\}$

Machine learning Model

(distribution of data) X = space of i/p sample

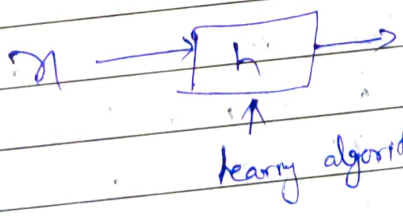
Y = space of o/p value

Goal of Supervised learning ~~Algo~~ is: Given

$h: x \to y$

hypothesis function

such a way that $h(x)$ is a good



linear reg

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \ldots$

θ_i → parameter or weights

↳ that control how you map function $h(x)$ x to y

$h_\theta(x) = \sum_{i=0}^{d} \theta_i x_i$ ($x_0 = 1$) (d = no. of i/p variable)

$$h_\theta(x) = \sum_{i=0}^{d} \theta^T x$$

(Bedroom, Bathroom, Squarefeet)

5
6
7

Choos θ such that minimise $J(\theta)$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

m = no. of sample

Least Mean Square (LMS) → gt

→ Choos θ such that minimise $J(\theta)$

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

$\frac{d}{d} x^2 = 2x \frac{dx}{dx}$

For Sigle term

$$\frac{J(\theta)}{d\theta} = 1 (h_\theta(x) - y) \times (h_\theta x_i - 0)$$

$$= (h_\theta(x) - y) x_j$$

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}$$

Two Way

1. Batch gradient descest
2. Stochastic " "

Batch Gradut descet

Repeat until Converges {

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}$$

(for every j)

→ Assum you learing rate is not too large, the above eqn Converges to global minima.

→ For every data point based on.

See all Sample

Batch GD (Slow but reliable)

See 1 Sample of update

Stochastic GD (fast)

SGD get θ close to minima but never reaches reachable the global minima.

→ Iterative approach (BGD & SGD)

⇒ Non-iterative approach

find the optimal set of 'weight' in a single-setup

$$y = mx + C$$

$$X = \begin{bmatrix} — & x^{(1)T} & — \\ — & x^{(2)T} & — \\ & \vdots & \\ — & x^{(n)T} & — \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

$$X\theta - y \begin{bmatrix} h\theta(x^{(1)}) - y^{(1)} \\ \vdots \\ h\theta(x^{(m)}) - y^{(n)} \end{bmatrix}$$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h\theta(x^{(i)}) - y^{(i)})^2$$

$$Z^T Z = \sum z_i^2$$

$$J(\theta) = \frac{1}{2} (X\theta - y)^T (X\theta - y)$$

M T W T F S S
Page No:
Date:
YOUVA

Matrix derivation

$f: A \begin{bmatrix} A_1 & \cdots & \\ & & A_n \end{bmatrix}$

$\nabla_A f(A) = \begin{bmatrix} \partial A_{11} & \cdots & \partial A_{1n} \\ & & \end{bmatrix}$

grad in nxn matrix

$J(\theta) = \frac{1}{2}(X\theta - y)^T (X\theta - y)$

$\nabla_\theta J(\theta) = \frac{1}{2}\nabla_\theta [(X\theta)^T X\theta - (X\theta)^T y - y^T X\theta - y^T y]$

$(AB)^T$

$= \frac{1}{2}(\theta^T(X^TX)\theta - y^T(X\theta) - y^T X\theta)$

$= \frac{1}{2}\nabla_\theta (2(X^TX)\theta - 2(X^Ty)^T\theta)$

$(ab)^T = a^T b^T$

$= (X^TX)\theta - X^Ty$

Equate the first derivative to zero

$X^T X\theta = X^T Y$

linear regression has closed for loop

$\theta = (X^TX)^{-1} X^Ty$

X^T is not possible to find anytime

Matrix need to be Singular to find X inverse.

M T W T F S S | Page No: | YOUVA | Date:

$X_{(n \times d)} \rightarrow X^{-1}_{(d \times n)}$

$d \times d$ $n \times n$

Logistic Regression

$h_\theta(x) \rightarrow \theta^T x \rightarrow$ Sigmoid funtn

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

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

As $z \to \infty \quad g(z) \to 1$

$z \to -\infty \quad g(z) \to 0$

$$\theta^T x = \theta_0 + \sum_{j=1}^{d} \theta_j x_j$$

$$= \Sigma$$

$$\frac{dg(z)}{dz} = \frac{-1}{(1+e^{-z})} \frac{d}{dz}(1+e^{-z})$$

$$= \frac{e^{-z}}{(1+e^{-z})^2}$$

$\frac{d}{dx} f(x) = 1/x$

$df(x) = \frac{1}{x^2} dx$

derivative of Sigmoid

$$g'(z) = g(z)(1-g(z))$$

$h_\theta(x)$ $\quad y=1, y=0$

likelihood way of writing

$$\begin{cases} P(y=1|x;\theta) = h_\theta(x) \\ P(y=0|x;\theta) = 1 - h_\theta(x) \end{cases}$$

$$P(y|x;\theta) = (h_\theta(x))^y (1-h_\theta(x))^{1-y}$$

Assuming training examples are generated independent we can write down the likelihood parameters as

$$L(\theta) = P(y|x;\theta)$$

$$\prod_{i=1}^{n} P(y^{(i)}|x^{(i)};\theta)$$

$$= \prod_{i=1}^{n} (h_\theta(x^{(i)}))^{y^{(i)}} (1-h_\theta(x^{(i)}))^{1-y^{(i)}}$$

log is monotonically increasing

$$\log L(\theta) = \sum_{i=1}^{n} y^{(i)}$$

Maximum likelihood using gradient descent

$$\theta := \theta + \alpha \nabla \theta$$

• Tensor flow Pytorch ---> Auto diff

M T W T F S S
Page No.
Date
YOUVA

$$\frac{\partial}{\partial \theta} J(\theta) = \left(y \frac{1}{g(\theta^T x)} - (1-y) \frac{}{1 - g(\theta^T x)} \right.$$

$$\left(\frac{y}{g(\theta^T x)} - \frac{(1-y)}{1 - g(\theta^T x)} \right) g(\theta^T x)(1 -$$

find Cost func then find derivate