Classification

1. Predicting category ( input features( independent) , predictive(dependent) output
   1. Binary Classification ( Two classes)
      1. Eg : Loan approval ( yes or no )
      2. Student pass or fail ( pass or fail )
      3. Predict image (Kishan or not )
   2. Multi class classification ( one feedback can be assigned to only one category)
      1. Feedback -> content, faculty, understanding , duration , support, overall ( eg feedback one from student1 )
         1. Kishan classes are interactive (faculty)
         2. he take long duration and short breaks( duration )
         3. he takes us outside the context and content ( content,)
         4. probably good if the notebook is given well in advance( support)
   3. Multi Label classification :One single message/feedback can contain info about all categories.( one single feedback is assigned to 4 labels )
      1. Eg- ( eg feedback one from student1 ) Kishan classes are interactive but he take long duration and he takes us outside the context and content and probably good if the notebook is given well in advance (faculty)
      2. Eg- ( eg feedback one from student1 ) Kishan classes are interactive but he take long duration and he takes us outside the context and content and probably good if the notebook is given well in advance ( duration )
      3. Eg- ( eg feedback one from student1 ) Kishan classes are interactive but he take long duration and he takes us outside the context and content and probably good if the notebook is given well in advance ( ( content,)
      4. Eg- ( eg feedback one from student1 ) Kishan classes are interactive but he take long duration and he takes us outside the context and content and probably good if the notebook is given well in advance ( ( support)

Classiifcation

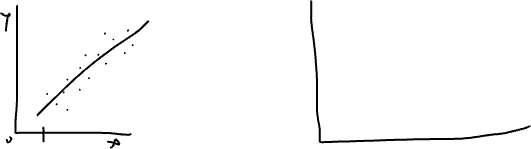
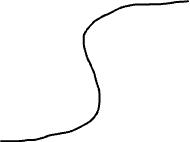
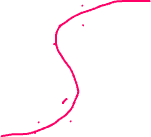
1. Logistic regression
   1. If dependent variable is not a number ,
   2. but still have all independent variables
   3. classify the dependent variable into classes / groups ( 0,1), true or false ) = binary
   4. eg: bank approval
      1. build model whether to give loan or not
         1. relationship between variables ( affect )
            1. linear regression = y – mx + c



* + - 1. threshold of the relationship = 0.5
      2. Logistic sigmoid function is then applied to get probabilities whether he can approved or not



* + - * 1. Logistic regression = p = 1/1+e-y



* + - 1. Log of the probability of YES vs log of proabability of NO
      2. Pr(y)= event occurring = YES or 1
      3. Pr(~y) =non-event = NO or 0
      4. Goal of Log Regresion is to get the best relationship between independent vs the dependent( maximum likelihood)to fit sigmoid curve
      5. Advantages
         1. Affect/influence of independent , on dependent
      6. Disadvantages
         1. Works for binary
         2. Large number of categorical features
         3. Assumes tht missing value handling is done

Bayes theorem( conditional proabablities ) ( probability of one event occurring given that the another event already occurred )

In ML classification – NB- Naïve Bayes Classification

Condition in stats , in ml its Data , more data we have we get better accurate results

Eg: student ages in class

Data : evidence. = X = 18,24,26,35,38,40, 46,47,48,50,56,57,51

Independence between every pair of features

Age group vs no of cards

Age group vs no of loans

Age groups vs no defailters

Adv

1. Very fast

Metrics

TP( true positive )

TN( true negaitive )

FP ( false positive ) = model predicted wrongly positive

FN( false negaitive) = model predicted wrongly negative

Confusion Matrix – performance of my model – visualue class

Records -3019

|  |  |  |
| --- | --- | --- |
|  | Predicted class NO | Predicted class YES |
| Actual class NO | TN(True Negative ) 1000 | FP( False positive) type 1 error 9 |
| Actual class YES | FN(False negative ) type 2 error. 10 | TP( True positive )2000 |

False positive rate = FP / FP + TN

False negative rate = FN/FN+TP

CLASSIFICATION METRICS

1. Accuracy = TP +TN / TP+ TN+FP+FN
2. Precision = for imbalanced datasets. = TP/ TP +FP, how I Predicted my model correctly ( MORE HIGH )
3. Recall = True positive rate = TP/TP + FN( MORE HIGH )



1. F1Score : Combine precision + recall
   1. F1 = 2\*( presicion \*recall )/( precision + recall)
2. AUC – area under curve

Do u have Labeleld data ( yes)

1. If Sample <100k
   1. Yes
      1. Linear svc
         1. If this is not working go for
            1. Naïve bayes with the test data
         2. If the test data also not working
            1. K neighbors

If this also not working

SVC ensemble classifier

* 1. NO
     1. SGD classifier
        1. Kernel approximation