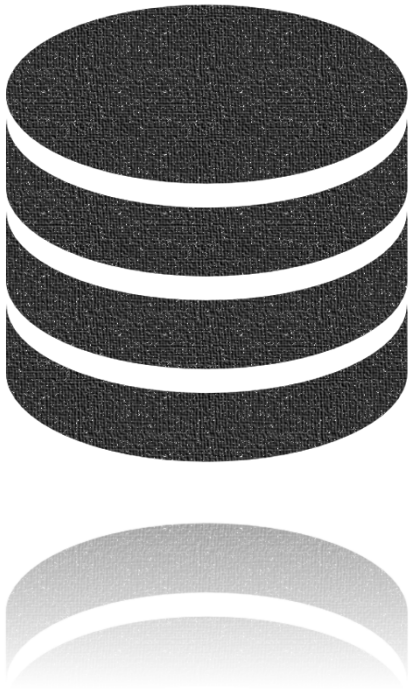
Ad – Non Ad Classifier

Introduction

- The project serves to classify the images displayed on a web page as either an Advertisement or as a regular image.

Dataset Overview

- Instances : 3279
- Features : 1558
- Discrete Features : 1555
- Continuous Features : 3
- Missing Data Percentage : 28%



Dataset Overview - 2

- Features from URL Terms : 458
- Features from Original URL Terms : 495
- Features from Anchor Text Terms : 472
- Features from Alternate Text terms : 111
- Features from Caption Terms : 19

```
<!DOCTYPE html>
<html lang="en">
 <head>
  .
 </head>
 <body>
  <a href="">..</a>
 </body>
</html>
```

Imputing Missing Data

- k Nearest Neighbor
  - Computing a similarity score for the test data
  - Assigning the test data's missing data with the most similar training example

Accuracy - Imputing

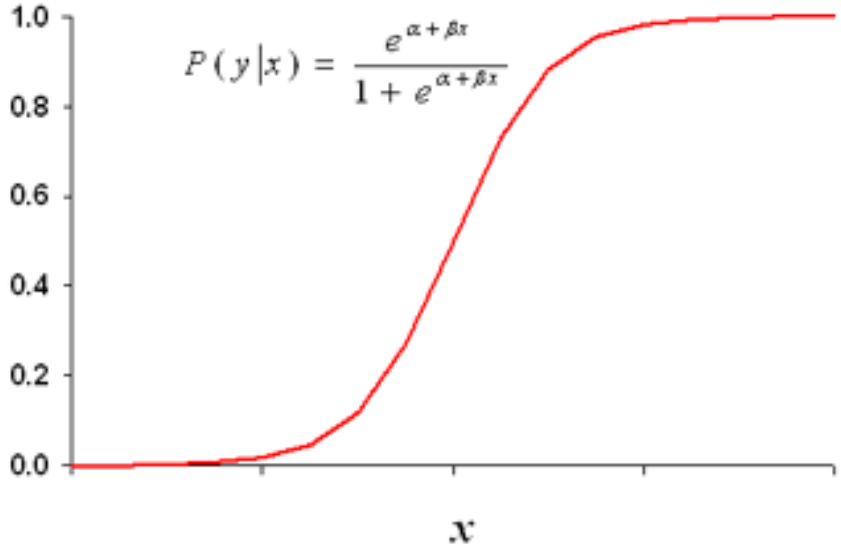
- k Nearest Neighbor :
  - 85-87% Accuracy – Calculated using cross validation techniques
  - The splits applied were 9:1 , 8:2 , 6:4

Discretization

- 3 Continuous Features
  - Height – 13 new binary features
  - Width – 13 new binary features
  - Aspect ratio – 6 new binary features
- Each of the features discretized into a vector of 0's and 1's

Prediction

- Logistic Regression



Accuracy - Prediction

- Logistic Regression :
  - Data-set broken down into training and testing
  - 94% accuracy achieved with 2:3 and also with 4:1 (training : testing) split

What did we learn ?

- Implement Linear and Logistic Regression
- Calculate the similarity measure
- Impute missing data
- Discretization from continuous feature space to binary feature space

Credits

- Dataset Used : https://archive.ics.uci.edu/ml/datasets/Internet+Advertisements
- Members:
  - Jairaj Singh Shaktawat (UIN: 675090156)
  - Pooja Donekal (UIN: 656819565)
  - Shvetha Suvarna (UIN: 663171839)
  - Sreejith Menon (UIN: 673420442)
  - Surbhi Arora (UIN: 660042750)

Thank You


Questions?