Soft Computing:ISCO630E
Report

Assignment 5.1

Anant Chaturvedi(IIT2016506)

Question Description

1. Using Naive Bayesian classifier predict where a given mail is spam or not. Use the data set provided for this purpose. (structured data set)

Introduction

We are given a dataset comprising of email body text followed by the label depicting wether the email was a spam mail or ham mail.

Concepts Used

Bayesian Formula

P(A | B) = **P(B | A) * P(A) / P(B)**

Formula Definition

P(ham | bodyText) = Probability that email is ham given that it contains document- bodyText (lets say bodyText = content of email)

P(spam | bodyText) = Probability that email is spam given that it contains document- bodyText

P(ham | bodyText) = (P(ham) * P(bodyText | ham)) / P(bodyText)

P(spam | bodyText) = (P(spam) * P(bodyText | spam)) / P(bodyText)

P(ham | bodyText) = (P(ham) * P(bodyText | ham)) / P(bodyText)

P(spam | bodyText) = (P(spam) * P(bodyText | spam)) / P(bodyText)

P(ham) = no of documents belonging to category ham / total no of documents

P(spam) = no of documents belonging to category spam / total no of documents

P(bodyText | spam) = P(word1 | spam) * P(word2 | spam) * ...

P(bodyText | ham) = P(word1 | ham) * P(word2 | ham) * ...

P(word1 | spam) = count of word1 belonging to category spam / total count of words belonging to category spam.

P(word1 | ham) = count of word1 belonging to category ham / total count of words belonging to category ham.

RESULTS

75% of the data was used for training while the remaining was used for testing the predictor.

The Accuracy on the Training data is around :- 98.82051282051282%

The Accuracy on the Testing data is around :- 97.188995215311%