DV2599

Alexander Jamal

Arlind Iseni

Assignment 1 Report

Introduction

The subject of study for this assignment is a concept learner-algorithm applied on a ham-or-spam classification problem. We were given a data set of aggregated statistical summaries for each e-mail in a big cohort of e-mails. The data type for most of these features were of the continuous type and an appropriate data-preprocessing with discretization as the main operation was to be conducted. As will be described in the method section of this report we used ordinal binning to solve this problem.

Method

Firstly, we imported all modules needed for this assignment. We used *KBinsDicretizer* function from the module *sklearn.preprocessing* to turn our continues data values into discrete values with binning. We then imported the data and saved it in a dataframe and check if the required columns are in the dataframe. We continued with exploring the data by plotting the frequency of different columns then checked the ham and spam rate. After that, we did some data munging and made the values discrete, so we continued by plotting the distribution of all columns. Finally using concept learning algorithm, we were able to find out the accuracy of the model.

Results

From our analysis and implementation we find an accuracy of roughly 60% with 10 bins and the uniform binning algorithm. All other binning algorithms (k-means, quantiles) and bin amounts (ranging from 2 to 20) resulted in lower accuracy. We can conclude that a concept learner-algorithm gives us a semi-reliable (> 50%) estimate of whether an e-mail is spam or not however it is a very inaccurate tool and with some configurations a coin flip would have given us a more robust prediction.