Danmarks Tekniske Universitet

Report 3

Course:
02450: Machine Learning

Author(s):

Anders H. Opstrup (s160148)
Gu Jinshan (s161944)
Huayu Zheng(s162077)

DTU Compute
Institut for Matematik og Computer Science

02450: Machine Learning
Tue Herlau

29 November 2016

Contents

1 Introduction

As described in the conclusion of the first report the group have had some concerns about the dataset (forest fires). Therefor the group decided to change the dataset. The dataset was too small to make any interesting analysis. The group have chosen the SPAM dataset from http://statweb.stanford.edu/ tibs/ ElemStatLearn/. This dataset has around 50 attributes, compared to 12 the forest fires dataset has and the SPAM dataset has around 5000 records where the forest fires dataset only has around 500.

This report covers the second assignment for course *02450: Introduction to Machine Learning and Data Mining*. The assignments purpose is to solve a relevant regression problem and classification problem for our dataset. The group decided to predict the variable $word_freq_credit$ which is the frequency of the word credit used in an email, which could be an indication of SPAM mail. For the classification problem the group decided to identify SPAM from not-SPAM.

2 Clustering

3 Outlier detection/Anomaly detection

4 Association mining

5 Conclusion