CS423: Assignment 1 (Due: 16 September)

Data collection

Introduction

In this assignment you will be collecting data and constructing a dataset which will be used in the subsequent data mining tasks through out the rest of the term. The data can come from any subject domain for examples vehicle, food or animal. There is no restriction on data acquisition method, except that you are not allowed to simply take a dataset from public domains, e.g., UCI respository.

Your data should
1. contain at least 10 features
2. contain at least 100 instances

(The more the better anyway)

Instructions

1. Write down a description of your dataset including short description of each of its features. State also the original type of the features (binary, ordinal …) .

2. Write down any preprocessing steps used to transform the data into a computable form. This includes discretisation and normalisation methods used.

3. Perform correlation analysis between features using a method of your choice to single out
    - The most correlated features
    - The least correlated features

   Present the results in an easily understandable form and make sense of your findings.

4. Perform an educated guess on what interesting information can be mined from this set of data. Justify your answer.

What to submit

1. A report.
2. The dataset in a csv format.

Submission method

Email to: jakramate.b@cmu.ac.th