

PESIT Department of Computer Science and Engineering

Course: Data Mining
Semester: 2016 Spring (January – May)
Instructor: BNR (Dr. B. Narsing Rao)

Assignment: 12
Topic: Outlier Detection
Due by: Midnight on Tuesday, April 19, 2016
Method: See below for details; email to bnrao@pes.edu

For this assignment use the provided file **winequality-white.csv** which has details about the quality of 4898 white wines. For an explanation of the attributes see the accompanying file **winequality.names**.

Use the Weka API to perform the following tasks:

1. List, for each attribute, the number of outliers as determined by the following expression:

Outlier if $\text{value} > Q3 + OF * IQR$
or $\text{value} < Q1 - OF * IQR$

where OF is the outlier factor. OF should be an input parameter (use 3.0)

The output should be a table with the attribute name and number of outliers (low, high, total), sorted in descending order of total.

2. List the number of outliers that have a given number of attributes in the outlier range, in descending order of number of attributes, and generating a set of descriptive labels for each outlier
3. Record your observations on what may be regarded as the general characteristics of outliers for this dataset. What is the effect of changing OF? Can you make any general statement about the quality of outliers?

Submit the following using standard naming conventions:

- Your program
- Text file containing outputs from tasks listed above
- PDF file containing observations and answers to questions