CIS 635 Data Mining

Homework 3

Objectives

- reinforce the concepts and characteristics of data
- practice exploratory analysis

Instructions

Part A – data characteristics

id	gender	ageGroup	pulse	temp	virus
101	M	1	67	96.1	Y
102	M	3	68	97.8	N
103	F	2	73	98.3	N
104	M	2	77	98.9	N
105	F	1	62	98.1	Y
106	F	3	81	99.1	N
107	M	1	70	96.9	N
108	F	2	81	97.5	Y
109	M	3	105	98.2	N
110	F	2	63	98.6	Y

The data in the table to the right represents a study of patients with respect to a specific virus. The id is just a number that is used to identify a patient. The gender is the patient's gender, ageGroup places the patient in one of three groups: 1=teenager, 2=young adult and 3=older adult. The pulse and temp are the heart rate and temperature of patient and the column with virus shows whether or not the patient has the virus. Answer the following:

1. assume that the heart rate for patient 103 is not valid (just for this question). What would be the replacement value using
   a) global average
   b) average for ageGroup
2. if we use 3 std dev as a way to identify outliers, list any outliers for heart rate
3. to reduce the dimensionality, suggest one or two columns to remove based on correlation or lack of generalization.

Part B – exploratory analysis

The data file, hw03data.txt has been posted on BB. Download this file and following the directions below.

1. using summary statistics, answer the following questions
   a) how many instances are in the data set?
   b) what is your estimate of the probability of having the virus?
   c) what is the probability of having a heart rate between 56 and 75?
2. histograms
   a) show histograms of both heart rate and temperature
   b) do they appear to be distributed according to the gaussian distribution?
3. boxplots:
   a) show boxplots of heart rate
   b) show boxplots of heart rate for just those with the virus
   c) show boxplots of heart rate for those with the virus and those without the virus, side-by-side
   d) show boxplots of temperature for those with the virus and those without the virus, side-by-side
   e) looking at the boxplots, which attribute (heart rate or temperature) appears to be best at predicting whether a person has a virus?
4. scatter plots
   a) show a scatter plot with hear rate vs temperature with dots for the instances. Give those with the virus a different color than those without the virus.
   b) does it appear that we could separate the majority of the instances by class?
5. tables
   a) use the table function to create a table of gender by virus
   b) use the table function to create a table of ageGroup by virus
   c) does it appear that gender or ageGroup are predictive of virus?