Logical Rhythm Problem Statements

CyberQuest, Avishkar 2016, MNNIT Allahabad

1 General Guidelines

- A team can comprise of 3 members, all of same year.
- A modular design will be appreciated. (E.g. Use of functions for performing repetitive tasks)
- You have to attempt atleast one problem.
- You can use any platform.
- If the work is found to be plagiarized, it would lead to immediate disqualification.
- *Data set for the problems can be found at* `http://172.31.9.68`

2 Problem Statements

2.1 Problem 1 - Grade Calculator

A study was conducted in two well-known private institutes to determine the factors affecting grades of students. The description of the attributes noted is as below :
Format:

$$< FeatureName - Description(< type >:< Setof PossibleValues >) >$$

- school - student's school (binary: 'NITA' - National Institute of Technology, Allahabad or 'IIITA' - Indian Institute of Technology, Allahabad)
- sex - student's sex (binary: 'F' - female or 'M' - male)
- age - student's age (numeric: from 15 to 22)
- address - student's home address type (binary: 'U' - urban or 'R' - rural)
- famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
- Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
- Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
- Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
- guardian - student's guardian (nominal: 'mother', 'father' or 'other')
- reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
- traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- studytime - weekly study time (numeric: 1 - < 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

- activities - extra-curricular activities (binary: yes or no)
- nursery - attended nursery school (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)
- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)
- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- absences - number of school absences (numeric: from 0 to 93)

These grades are related with the course subject:

- G1 - first period marks(numeric: from 0 to 20)
- G2 - second period marks (numeric: from 0 to 20)
- G3 - final marks (numeric: from 0 to 20, output target)

Your task is to predict the values of these three marking components using the training set provided. Even though the grades G1, G2, G3 in the training set are discrete integer values, you are expected to predict grades as a continuous real value in the range [0 - 20].

2.2 Problem 2 - Something's Phishy

In todays world where almost all user data is online, phishing has become a major menace. We want to classify websites as phishing websites based on some attributes.

Numbers against the attributes show the possible categories for that attribute. The descriptions for the features have been given only to fulfill your curiosity.

$$< FeatureName - SetofPossibleValues > -[Description(ifNecessary)]$$

- Has an IP Address in URL $\{-1, 1\}$
- URL length $\{1, 0, -1\}$
  1 - Short length URL ($len < 54$); 0 - Medium length URL ($54 < len < 75$); -1 - Long length URL ($len > 74$).
- URL Shortening Service Used $\{1, -1\}$
  URL Shortening services (like TinyURL) are used sometimes to redirect to phishing websites.
- Having @ Symbol $\{1, -1\}$
  Using '@' symbol in the URL leads the browser to ignore everything preceding it, and thus can be used for phishing.
- Uses Redirection $\{-1, 1\}$
  The '//' symbol can be used in a URL path to redirected to another website, and thus can be used for phishing.
- Has Prefix/Suffix $\{-1, 1\}$
  Phishers tend to add prefixes or suffixes using the dash$(-)$ symbol, which is rarely used in legitimate URLs.
- Having Sub-Domain $\{-1, 0, 1\}$
  1 - Having no sub domain; 0 - Having one sub domain; -1 - Having multiple sub domian.
- Uses HTTPS $\{-1, 1, 0\}$
  1 - uses https with trusted certificate; 0 - uses https with untrusted certificate; -1 - Otherwise.
- Domain Registration Length $\{-1, 1\}$
  -1 - Domain expires in one year; 1 - Otherwise.
- Has Favicon $\{1, -1\}$
  -1 - Favicon Loaded From External Domain; 1 - Otherwise.
- Uses other Ports $\{1, -1\}$
- Uses HTTPS Token $\{-1, 1\}$
  The phishers may add the "HTTPS" token to the domain part of a URL in order to trick users.
- Requests other URL $\{1, -1\}$
  1 - % of external requests $< 22\%$; 0 - $22\% <$ % of external requests $< 61\%$); -1 - Otherwise.

- URL of Anchor $\{-1, 0, 1\}$
  1 - % of URL Of Anchor $< 31\%$; 0 - $31\% <$ % of URL Of Anchor $< 67\%$); -1 - Otherwise.
- Links used in tags $\{1, -1, 0\}$
  1 - % of Links in meta, script and link tags $< 17\%$; 0 - $17\% <$ % of Links in meta, script and link tags $< 84\%$); -1 - Otherwise.
- SFH $\{-1, 1, 0\}$
  -1 - SFH is "about: blank" ; 0 - SFH "refers to" A Different Domain;1 - Otherwise .
- Submitting something to Email $\{-1, 1\}$
- Abnormality in URL $\{-1, 1\}$
- Website Forwarding $\{0, 1\}$
  1 - Redirected pages < 1; 0 - $2 <$ Redirected pages $< 4\%$); -1 - Otherwise.
- Performs actions on Mouseover event $\{1, -1\}$
- Tampers with context menu $\{1, -1\}$
- Has popup window $\{1, -1\}$
  Legitimate websites hardly have pop-ups in them.
- Uses Iframe $\{1, -1\}$
  Phishers can make use of the "iframe" tag to display an additional webpage into one that is currently shown.
- Age of Domain $\{-1, 1\}$
  -1 - website life $< 6 months$; 1 - Otherwise
- DNS Record Present $\{-1, 1\}$ Most Phishing websites do not have a DNS record.
- Amount of Web Traffic $\{-1, 0, 1\}$
  -1 - If domain has no traffic and no record in Alexa Database; 1 - Ranked among the top 100,000 websites; 0 - Otherwise.
- Page Rank $\{-1, 1\}$
  PageRank is a value ranging from "0" to "1", which aims to measure how important a webpage is on the Internet. The greater the PageRank value the more important the webpage. In our datasets, most phishing webpages have no PageRank.
- Google Index found $\{1, -1\}$
  This feature examines whether a website is in Googles index or not.
- Pages linked to this page $\{1, 0, -1\}$
  1 - More than 2 external links pointing to the webpage ; -1 - No external links pointing to webpage ; 0 - Otherwise.
- Statistical Report Present $\{-1, 1\}$
- Whether this is a phishing site or not $\{-1, 1\}$ (Expected as Output)

Based on this information, train a classifier to judge whether a given site is a phishing website or not.