Christopher McClaskey
CSCI 4502 - Homework 1
SID: 810937948
christopher.mcclaskey@colorado.edu

Honor Code Pledge

On my honor, as a University of Colorado at Boulder student, I have neither given nor received unauthorized assistance on this work.

Question 1

Resource 1:

(a) http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

(b) Aggregation of socio-economic data, law-enforcement data, and crime data, by region. It has 1,994 objects, and 128 attributes (ie state, county, pctHSGrad, etc).

(c) Knowledge that can be mined would be patterns and associations between socio-economic status, region, and crime.

(d) This knowledge would be useful in mapping areas that need attention, whether that attention be better a better education system, economic stimulus, or law-enforcement improvement. This map could help in the decision to divvy out resources whether it be money or man-power.

Resource 2:

(a) http://archive.ics.uci.edu/ml/datasets/Census+Income

(b) Selection of tax and income data, and includes some socio-economic data. It has 48,842 objects, and 14 attributes (age, workclass, sex, etc).

(c) Knowledge that can be mined would be patterns and associations among socio-economic status and income.

(d) This knowledge would be useful in economic models. It could help judge the state of the economy, the employment situation, and see if there are any problems that could be fixed. Perhaps low income areas all have something in common that could be addressed.

Question 2 (extra credit)

(a) KDD'11. "Multiple Domain User Personalization" by Amr Ahmed, Yucheng Low, Mohamed Aly, Vanja Josifovski, Alexander J. Smola
http://dl.acm.org/citation.cfm?id=2020433&CFID=176070325&CFTOKEN=86241249

(b) The problem being addressed is the "Cold Start Problem". This problem is specific to websites that try to make recommendations to the user. Typically these websites require some user input before they can do anything. For example, there isn't a way for Netflix to know what kind of movies a user likes without asking the user or looking at their previously watched movies. The idea is to make recommendations from the start, without user input, but this is very difficult.

(c) The authors' proposed solution is to combine information from across multiple domains. A user's google searches or facebook likes can provide insight into a user's preferences elsewhere. The author's do this by constructing a Bayesian model which integrates multiple different recommendation systems.

(d) The solution is presented and evaluated with some statistical theory, implementation of the statistics with code, and testing. The test results were presented with metrics.

Question 3

Source Code and Output (Ruby):

```ruby
# Christopher McClaskey
# CSCI 4502
# Homework 1
# Question 3 parts (a) and (b)

require 'csv'

wind = Array.new
headers = true
CSV.foreach("forestfires.csv") do |row|
        #skip the column name row
        if headers
                headers = false
                next
        end
        #add the wind value to my array
        wind.push(row[10].to_f)
end

#Calculating Mean
mean = wind.inject { |sum, i| sum + i } / wind.size

#Calculating Standard Deviation
std_dev =  Math.sqrt( wind.inject { |sum, i| sum + (i-mean)**2 } /
wind.size )

#Calculating median and quartiles
wind = wind.sort
median_id = wind.size/2
lower = wind[0..median_id-1]
upper = wind[median_id+1..wind.size-1]
q1 = lower[lower.size/2]
median = wind[median_id]
q3 = upper[upper.size/2]
iqr = q3-q1

puts "Min:        #{wind.min}"
puts "Max:        #{wind.max}"
puts "Mean:       #{mean}"
```

```ruby
puts "Std Dev:    #{std_dev}"
puts "Q1:         #{q1}"
puts "Median:     #{median}"
puts "Q3:         #{q3}"
puts "IQR:        #{iqr}"

# -OUTPUT-
# Min:          0.4
# Max:          9.4
# Mean:         4.017601547388782
# Std Dev:      1.789651406176351
# Q1:           2.7
# Median:       4.0
# Q3:           4.9
# IQR:          2.2
```

Scatter Plot of Attributes 10 (RH) and 11 (Wind):