

Course: CS-513

Instructions:

There are a total of six (6) problems, with point values noted for each. Use excel to complete all six problems; please show all of your calculations. Use excel to solve problems 1 to 5 and use R for the 6th problem.

Combine all of your answers/files into a single zipped file and post the zipped file to "Final Exam" in CANVAS.

All submissions are due **no later than 6.00 p.m. Wednesday, Dec 14, 2014.**

Use the following data descriptions for problems #1 and #2:

Predict student performance in a statewide exam (e.g. BAR), using the training dataset StateWide.csv (in the “raw data” module in CANVAS). It contains data for 50 students using the following variables. Show your work in an Excel file.

- ID: Student ID
- LSG: Last semester grade (A, B,C)
- CTG: Two preparation tests are conducted and the average of two tests are used to calculate marks. CTG is split into three categories:
 - Poor: = or < 40%
 - Average > 40% and < 60%
 - Good = or >60%.
- GP: General Proficiency seminar.
 - Yes: Student participated in seminar
 - No: Student did not participate
- Outcome: Outcome of statewide exam.

#1 (25 points)

Use StateWide.csv (CANVAS) and the CART Methodology to develop a classification model (Two levels).

#2 (25 points)

Use StateWide.csv (CANVAS) and the C4.5 Methodology to develop a classification model (Two levels).

#3 (10 points)

A neural network consists of 4 nodes in the input layer (1,2,3,4), two nodes in the hidden layer (A,B), and one node in the output layer (z).

Use the following input values, weights and learning factor to calculate the output of the network.

	$W_{xA}=.5$	$W_{xB}=.6$	$W_{XZ}=.4$
$x_1=.2$	$W_{1A}=.6$	$W_{1B}=.5$	$W_{AZ}=.9$
$x_2=.4$	$W_{2A}=.7$	$W_{2B}=.7$	$W_{BZ}=.9$
$x_3=.2$	$W_{3A}=.6$	$W_{3B}=.8$	
$x_4=.7$	$W_{4A}=.9$	$W_{4B}=.6$	
Learning Factor=.1			

4 (15 Points)

Using the single-linkage hierarchical clustering method and the Euclidian distance function, cluster the following normalized points. (Note: Do NOT normalize the data)

$A=(0.1,0.3)$, $B=(0.3,0.3)$, $C=(0.3,0.3)$, $D=(0.5,0.3)$, $E=(0.1,0.2)$

5 (15 Points)

Using the k-means clustering method ($k=2$), cluster the following normalized points (Do NOT normalize the data) and calculate BCV and WCV.

$A=(0.2,0.3)$, $B=(0.3,0.2)$, $C=(0.5,0.4)$, $D=(0.1,0.2)$, $E=(0.2,0.2)$, $F=(0.4,0.5)$, $G=(0.5,0.5)$.

Hint use D and F as your initial cluster centroids.

#6 (10 points)

Develop the following R program (do NOT normalize the data):

- a) Load the IRIS dataset into memory.
- b) Create a test dataset by extracting every third (3rd) row of the data, starting with the second row.
- c) Create a training dataset by excluding the test data from the IRIS dataset.