CSC 5800 : Intelligent Systems

Homework 2

Due Date: October 15th, 2018 **Total: 100 Points**

Problem 1. PCA + Visualization

(10 Points; 5 + 5)

(a) Perform Principal Component analysis (PCA) on the iris data and reduce it into two dimensional data. Plot this new 2D data. (show different classes with different symbols).

(b) Show the scatter plot between petal width and petal length features for the iris data. Which of the above two plot show the data that is relatively well separated? Explain your reasoning.

Problem 2. Distance measures

(15 Points; 4 + 5 + 6)

(a) What is the relationship between the distances obtained from the minkowski distance measures when r=1, r=2 and r=infinity? (Which one is smaller and which one is greater?)

(b) Let (x1=0, y1=0) and (x2=5, y2=12) be two points on a two-dimensional plane. Find the values of the minkowski distance between these two points when r=1, r=2, r=4 and r=8? Do you observe any trend in these values? Based on this observation, what can you conclude about higher values of r?

(c) Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object vector has an L2 length (magnitude) of 1. (NOTE: your final answer should be independent of the original vectors).

Problem 3. Visualization

(20 Points; 7 + 7 + 6)

(a) For various plots, fill-in the table shown below: (NOTE: try to use one of the following: Attributes, data objects, attribute values, count, percentile, or others.)

plot	x-axis	y-axis
Histogram		
Box Plot		
Percentile Plot		
Scatter Plot		
Data Matrix Plot		
Correlation Matrix Plot		
Parallel Coordinates Plot		

(b) Two Walmart stores (one in Chicago and the other in Detroit) made a combined profit of $1000 by selling Apples and Bananas. It is also known that, in the Detroit store,

Bananas got double the profit compared to Apples. The total profit made in the Chicago store is $400 and the total combined profit made by selling Bananas in both the stores is $700. Construct the data-cube for this problem.

(c) Construct a data cube from the following Table

Product ID	Location ID	Number Sold
1	1	10
1	3	6
2	1	5
2	2	22

Problem 4. **Decision Tree based on Gini Index**

(15 Points; 2 + 1 + 2 + 4 + 4 + 2)

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Consider the training examples shown in the above Table for a binary classification problem.

(a) Compute the Gini index for the overall collection of training examples.
(b) Compute the Gini index for the Customer ID attribute.
(c) Compute the Gini index for the Gender attribute.
(d) Compute the Gini index for the Car Type attribute using multiway split.
(e) Compute the Gini index for the Shirt Size attribute using multiway split.

(f) Which attribute is better, Gender, Car Type, or Shirt Size?

Problem 5. Decision Tree based on Entropy

(20 Points; 2 + 8 + 8 + 2)

Consider the training examples shown in the following Table for a binary classification problem.

(a) What is the entropy of this collection of training examples?
(b) What are the information gains of $a1$ and $a2$ relative to these training examples?
(c) For $a3$, which is a continuous attribute, compute the information gain for every possible split.
(d) What is the best split (among $a1$, $a2$, and $a3$) according to the information gain?

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

Problem 6. Decision Tree based on Classification Error Rate

(20 Points; 7 + 10 + 3)

The following Table summarizes a data set with three attributes A, B, C and two class labels +, −. Build a two-level decision tree.

A	B	C	Number of Instances	
			+	−
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

(a) According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.
(b) Repeat for the two children of the root node after splitting.
(c) How many instances are misclassified by the resulting decision tree?