CS 5525: Data Analytics I

Homework 2

Due Date: September 27th, 2016 (4:00PM) **Total: 100 Points**

Problem 1. Covariance Matrix

(10 Points)

Consider the following survey data about users who joined an online community. The sample covariance between the user's height (in mm) and number of years being a member of the community is C (a positive number). Answer the following two questions. To obtain full credit, you must prove your answer by showing the computations clearly.

(i) Suppose the height attribute is re-defined as height above the average for all users who participated in the survey. For example, a user who is 1650 mm tall has a height value of -50 mm (assuming the average height of all users is 1700 mm). Would the covariance between the re-defined height attribute and number of years in the community be greater than, smaller than, or equal to C?

(ii) If the measurement unit for height is converted from mm to inches (where 1 inch = 25.4 mm), will the covariance between height (in inches) and number of years in the community be greater than, smaller than, or equal to C?

Problem 2. Document similarity **(20 Points; 10 + 10)**

A sample Term-Document matrix (8 terms and 4 documents) is provided below to perform vector space retrieval:

	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6	Term 7	Term 8
D1		2	3	4		4		5
D2		3	2		4	1	3	
D3	3	1	5	1	1	3	5	
D4	2	3	2			5	3	9

(a) Build the *weight matrix* in which each element corresponds to TF * IDF. (note: For IDF, please use natural logarithm)

(b) Using the weight matrix computed above, compute the rank order of the documents that would be found for the unweighted query **Term4 Term8** using the vector space retrieval method with cosine similarity measure. Report the corresponding similarity scores for all the documents in the ranked list.

Problem 3. Minkowski Distance Measures

(10 Points; 5 + 5)

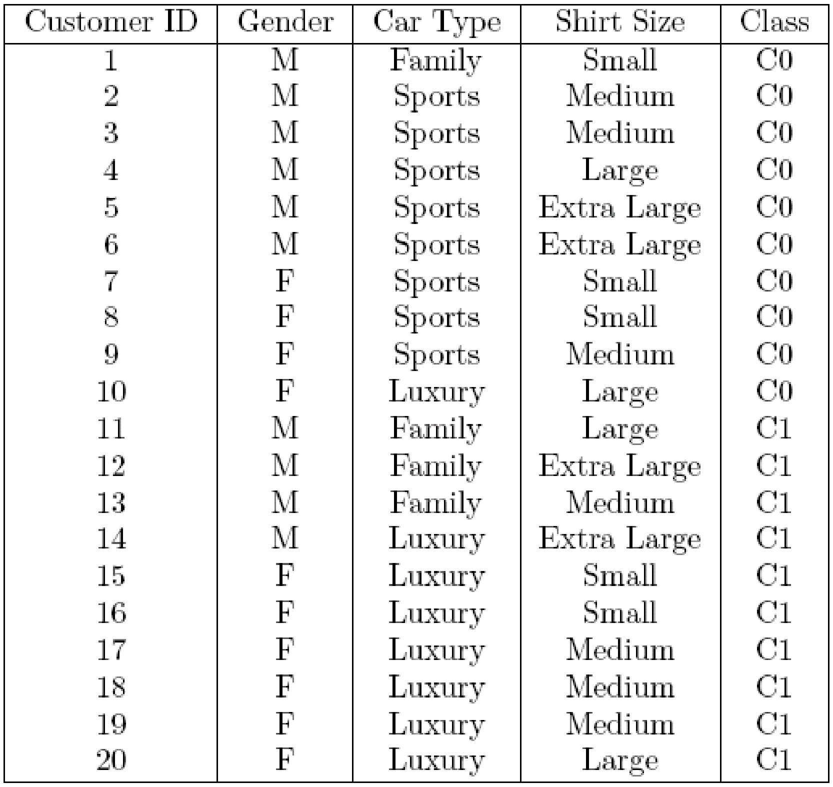
(a) What is the relationship between the distances obtained from the minkowski distance measures when r=1, r=2 and r=infinity? (Which one is smaller and which one is greater?)

(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?

Problem 4. Decision Tree based on Gini Index

(20 Points)

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



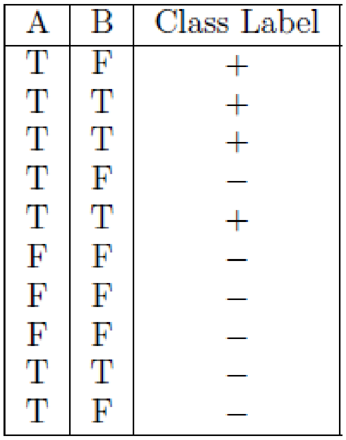
Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	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

Compute the Gini Index value of the parent and the child nodes obtained when the split was made on the following attributes (i) Gender and (ii) Car Type. Show all your calculations including the Gini of the parent, contingency tables, Gini of the individual children.

Problem 5. Decision Tree based on Entropy

(20 Points)

Consider the following data set for a binary class problem.



A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

(a) Calculate the information gain (gain in the entropy) when splitting on A and B.

(b) Which attribute would the decision tree induction algorithm choose?
Show all your calculations including the entropy of the parent, contingency tables, entropy of the individual children and the joint entropy of the children.

Problem 6. Decision Trees based on Misclassification Error

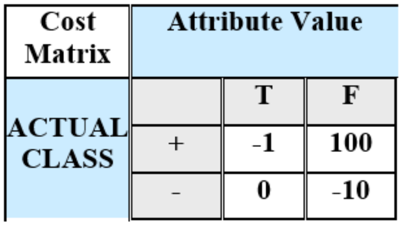
(20 Points; 10 + 10)

(a) The following table summarizes a data set with two attributes *A*, *B* and two class labels +, −. The original dataset had 5 attributes, but only 2 of the most informative attributes were selected for further analysis.

A	**B**	**Class Label**	
		+	**-**
T	T	0	20
T	F	20	10
F	T	15	0
F	F	0	35

(1) According to the classification error rate, which attribute would be chosen as the first splitting attribute?

(2) Use the following cost matrix and decide the first splitting attribute. The total cost is the metric for splitting



Cost Matrix	**Attribute Value**		
ACTUAL CLASS		**T**	**F**
	+	**-1**	**100**
	-	**0**	**-10**