CS 584 – Data Mining
Problem Solving HW 2
Due October 31, 2019

1. Consider an information retrieval task to find the text document from a collection of documents, which best matches a search keyword. The collection of text documents D1-D3 are given in the table, represented by bag-of-words (the features are words and values are the word counts in a document). (10 pts)

	baby	music	game	food	street
D1	10	0	0	5	0
D2	1	4	0	5	3
D3	0	2	10	5	3

The search keyword is *baby*.

(a) Represent the search keyword as a document in bag-of-words format.

(b) Using Euclidean distance, which document will be retrieved for this search?

(c) Using cosine similarity, which document will be retrieved?

(d) Based on the above observations, which proximity measure is more suitable for matching text documents, and why?

2. Consider the following set of training examples from two classes C1 and C2.

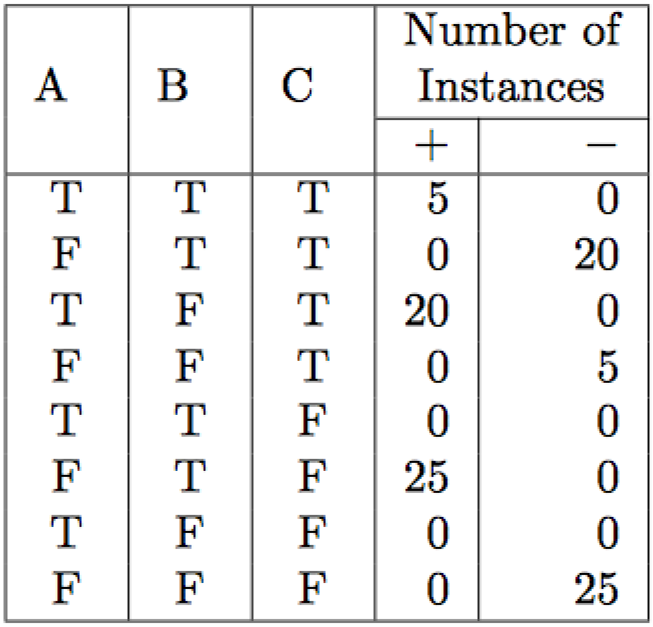
X	Y	Z	Number of C1 examples	Number of C2 examples
0	0	0	15	0
0	0	1	25	10
0	1	0	10	5
0	1	1	20	25
1	0	0	15	0
1	0	1	0	25
1	1	0	0	35
1	1	1	15	0

Compute a two-level decision tree. Use the classification error index as the criterion for splitting. What is the overall error rate of the decision tree? (30 pts)

3. Repeat question 2 using Z as the first splitting attribute and then choose the best remaining attribute for splitting at each of the two successor nodes. What is the overall error rate?

Compare the results of question 2 and 3. Comment on the suitability of the greedy heuristic used for splitting attribute selection. (20 pts)

4. The following table summarizes a data set with three attributes A, B, C and two class labels +, −. Build a two-level decision tree, using GINI index as the splitting criterion.



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 GINI index, which attribute would be chosen as the first splitting attribute?

(b) Compute the training error rate of the decision tree.

(c) If the cost associated with each leaf node is 0.5, compute the generalization error estimate. (20 pts)

5. Consider the following data with three binary attributes (A, B, C)

Record	A	B	C	Class
1	0	0	0	P
2	0	0	1	N
3	0	1	0	N
4	0	1	1	N
5	0	0	1	P
6	1	0	0	P
7	1	0	1	N
8	1	0	1	N
9	1	1	1	P
10	1	0	0	P

Using Naive Bayes model, predict the class label for instance (A=1, B=1, C=0). Detail the steps to arrive at your answer. (20 pts)