Fall 2018 Data Mining Homework 2 Due: Oct 17, 2018

1. Consider the training examples shown in Table 1 for a binary classification problem. [20 pt]

Table 1: Data set for Exercise 1.

Movie ID	Format	Movie Category	Class
1	DVD	Entertainment	C0
2	DVD	Comedy	C0
3	DVD	Documentaries	C0
4	DVD	Comedy	C0
5	DVD	Comedy	C0
6	DVD	Comedy	C0
7	Online	Comedy	C0
8	Online	Comedy	C0
9	Online	Comedy	C0
10	Online	Documentaries	C0
11	DVD	Comedy	C1
12	DVD	Entertainment	C1
13	Online	Entertainment	C1
14	Online	Documentaries	C1
15	Online	Documentaries	C1
16	Online	Documentaries	C1
17	Online	Documentaries	C1
18	Online	Entertainment	C1
19	Online	Documentaries	C1
20	Online	Documentaries	C1

(a) Compute the Entropy for the overall collection of training examples.

(b) Compute the Entropy for the **Movie ID** attribute.

(c) Compute the Entropy for the **Format** attribute.

(d) Compute the Entropy for the **Movie Category** attribute using multiway split.

(e) Which of the three attributes has the lowest Entropy?

(f) Which of the three attributes will you use for splitting at the root node? Briefly explain your choice.

2. Consider the decision tree shown in Figure 1, and the corresponding training and test sets in Tables 2 and 3 respectively. [20 pt]

(a) Estimate the generalization error rate of the tree using both the optimistic approach and the pessimistic approach. While computing the error with pessimistic approach, to account for model complexity, use a penalty value of 2 to each leaf node.

(b) Compute the error rate of the tree on the test set shown in Table 3.

(c) Comment on the behavior of training and test set errors with respect to model complexity. Comment on the utility of incorporating model complexity in building a predictive model.

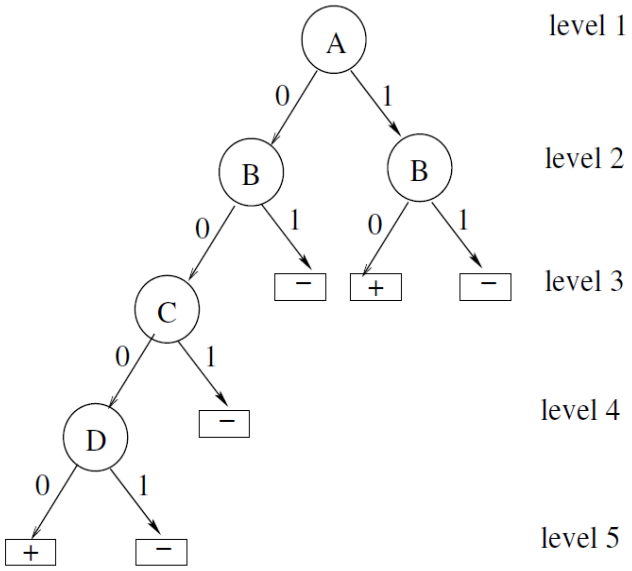


Figure 1: Decision tree for Exercise 2.

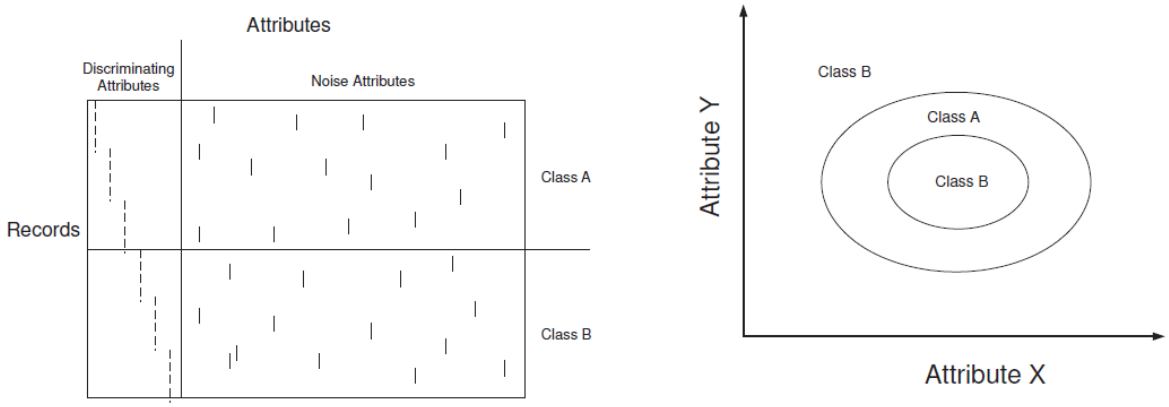
A	B	C	D	Number of + instances	Number of - instances
0	0	0	0	4	0
0	0	0	1	0	1
0	0	1	0	0	1
0	1	0	1	0	1
1	0	1	0	3	0
1	1	0	1	0	5

Table 2: Training set for Problem 2

A	B	C	D	Number of + instances	Number of − instances
0	0	0	1	4	0
0	0	1	1	3	0
0	1	0	0	0	1
0	1	1	0	0	2
1	0	0	0	2	0
1	0	0	1	3	0

Table 3: Test set for Problem 2

3. Given the data sets shown in Figure 2, explain how the decision tree and k-nearest neighbor (k-NN) classifiers would perform on these data sets. [20 pt]



(a) Synthetic data set 1. (b) Synthetic data set 2.

Figure 2: Data sets for Question 3

4. [20 pt] Consider the problem of predicting if a given person is a defaulted borrower (DB) based on the attribute values:

- Home Owner = Yes, No
- Marital Status = Single, Married, Divorced
- Annual Income = Low, Medium, High
- Currently Employed = Yes, No

Suppose a rule-based classifier produces the following rules:

- Home Owner = Yes → DB = Yes
- Marital Status = Single →DB = Yes
- Annual Income = Low →DB = Yes
- Annual Income = High, Currently Employed = No→DB = Yes
- Annual Income = Medium, Currently Employed = Yes→ DB = No

- Home Owner = No, Marital Status = Married → DB = No
- Home Owner = No, Marital Status = Single → DB = Yes

Answer the following questions. Make sure to provide a brief explanation or an example to illustrate the answer.

(a) Are the rules mutually exclusive?

(b) Is the rule set exhaustive?

(c) Is ordering needed for this set of rules?

(d) Do you need a default class for the rule set?

5. [20 pt] Consider the problem of predicting whether a movie is popular given the following attributes: Format (DVD/Online), Movie Category (Comedy/Documentaries), Release Year, Number of world-class stars, Director, Language, Expense of Production and Length. If you had to choose between RIPPER and a k-nearest neighbor classifier, which would you prefer and why? Briefly explain why the other one may not work so well?