DragonStar-2011 Solutions to Homework 2

1. [18 points: 3 for each part] Consider the training examples shown in Table 1 for a binary classification problem.

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 Gini index for the overall collection of training examples.
Answer: Gini $= 1 - 2 \times 0.5^2 = 0.5$.

(b) Compute the Gini index for the `Movie ID` attribute.
Answer: The gini for each `Movie ID` value is 0. Therefore, the overall gini for `Movie ID` is 0.

(c) Compute the Gini index for the `Format` attribute.
Answer: The gini for `DVD` is $1 - 0.25^2 - 0.75^2 = 0.375$. The gini for `Online` is 0.4444. Therefore, the overall gini for `Format` is $0.4 \times 0.375 + 0.6 \times 0.4444 = 0.4166$.

(d) Compute the Gini index for the `Movie Category` attribute using multiway split.
Answer: The gini for `Entertainment` movie is 0.375, `Comedy` movie is 0.2188, and `Documentaries` movie is 0.375. The overall gini is 0.3125.

(e) Which of the three attributes has the lowest Gini index?
Answer: `Movie ID`

(f) Which of the three attributes will you use for splitting at the root node? Briefly explain your choice.
Answer: `Movie Category`. Although `Movie ID` has the lowest Gini index, it is clearly just an identification attribute and a decision tree using `Movie ID` would not generalize at all.

2. [23 points: 4+4+7+8] Consider the decision tree shown in Figure 1, and the corresponding training and test sets in Tables 2 and 3 respectively.

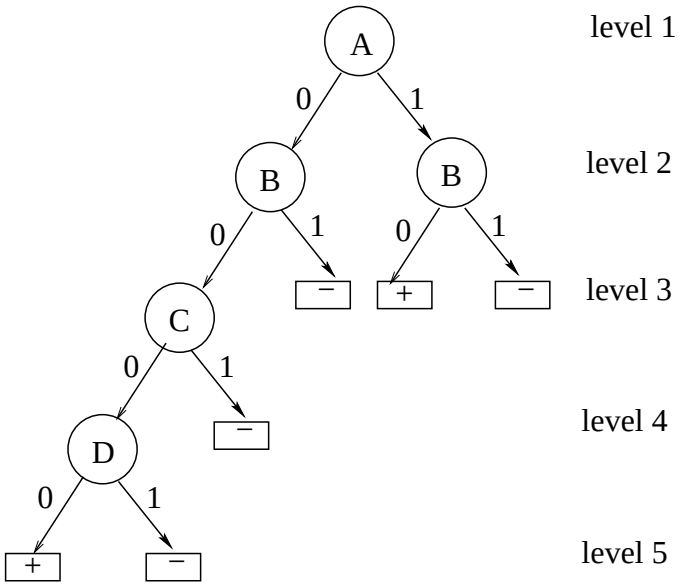


Figure 1: Decision tree for Exercise 2.

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

(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) Compute the optimistic generalization error on the training set and the error on the test

set (Table 3) by pruning each level from level 4 up to level 2, including the case where no pruning is done. Note that the root node is level 1.

(d) Comment on the behaviour of training and test set errors with respect to model complexity. Also, using a penalty value of 2, compute the generalization error rate using the pessimistic approach on the pruned version of the original tree that has lowest error rate on the test set. Comment on the utility of incorporating model complexity in building a predictive model.

Answers:

(a) $Error_{optimistic} = 0$; $Error_{pessimistic} = \frac{0+6*2}{15} = 0.8$

(b) $Error = \frac{4+3+0+0+0+0}{15} = 0.47$

(c) The answers are in the table below:

Level pruned	$Error_{train}$	$Error_{test}$
No Pruning	0	$7/15 = 0.467$
4	$1/15 = 0.066$	$3/15 = 0.2$
3	$2/15 = 0.133$	$0/15 = 0$
2	$6/15 = 0.4$	$8/15 = 0.53$

(d) As the tree is pruned more and more, the optimistic generalization error on the training set increases, but the error on the test set decreases to 0 when level 3 is pruned. This illustrates the overfitting problem. However, pruning further leads to an increase in both training and test error, since the model becomes too simple.

$Error_{pessimistic} = \frac{2+4*2}{15} = 2/3$

The original decision tree suffers from overfitting, which is reflected in the fact that its generalization error on the test set is significantly higher than that on the training set. Smaller trees are less likely to overfit the data. Indeed, the pruned tree's error rate on both the training and test set are very similar. The incorporation of model complexity through the pessimistic error rate allows a simpler (pruned) tree to be selected, even though it has a higher error rate on the training set.

3. [15 points: 3+3+3+6] Answer the following questions:

(a) Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 25%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student?

Answer:

Given $P(S|UG) = 0.15$, $P(S|G) = 0.25$, $P(G) = 0.2$, $P(UG) = 0.8$. We want to compute $P(G|S)$.

According to Bayes' Theorem,

$$P(G|S) = \frac{0.25 \times 0.2}{0.15 \times 0.8 + 0.25 \times 0.2} = 0.2941. \quad (1)$$

(b) Given the information in part (a), is a randomly chosen college student more likely to be a graduate or undergraduate student?

Answer:

An undergraduate student, because $P(UG) > P(G)$.

(c) Repeat part (b) assuming that the student is a smoker.

Answer:

From (a), $P(G|S) = 0.2941$, and thus $P(UG|S) = 1 - P(G|S) = 0.7059$. Thus, a smoker student is more likely to be an undergraduate since $P(UG|S) > P(G|S)$.

(d) Suppose 40% of the graduate students live in a dorm but only 10% of the undergraduate students live in a dorm. If a student smokes and lives in the dorm, is he or she more likely to be a graduate or undergraduate student? You can assume independence between students who live in a dorm and those who smoke.

Answer:

First, we need to estimate all the probabilities.

$P(D|UG) = 0.1$, $P(D|G) = 0.4$.

$P(D, S|G) = P(D|G) \times P(S|G) = 0.4 \times 0.25 = 0.1$ (using independence between students who live in a dorm and those who smoke.)

$P(D, S|UG) = P(D|UG) \times P(S|UG) = 0.1 \times 0.15 = 0.015$.

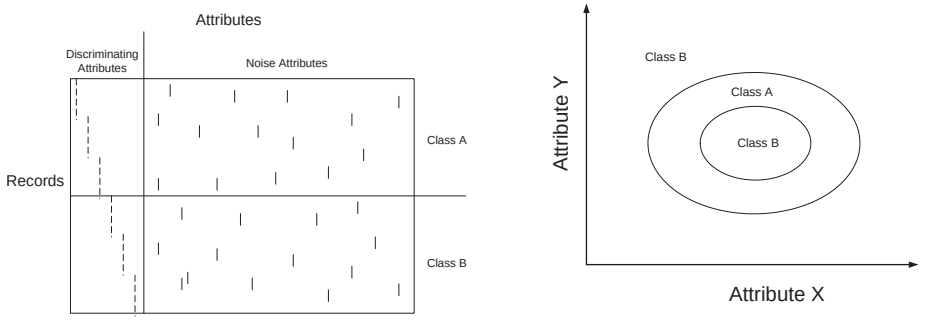
We need to compute $P(G|DS)$ and $P(UG|DS)$.

$$P(G|D, S) = \frac{P(D, S|G)P(G)}{P(D, S)} = \frac{0.1 \times 0.2}{P(D, S)} = \frac{0.02}{P(D, S)} \tag{2}$$

$$P(UG|D, S) = \frac{P(D, S|UG)P(G)}{P(D, S)} = \frac{0.015 \times 0.8}{P(D, S)} = \frac{0.012}{P(D, S)} \tag{3}$$

Since $P(G|D, S) > P(UG|D, S)$, he/she is more likely to be a graduate student.

4. [18 points: 9 for each part] Given the data sets shown in Figure 2, explain how the decision tree, naïve Bayes (NB), and k-nearest neighbor (k-NN) classifiers would perform on these data sets.



(a) Synthetic data set 1.

(b) Synthetic data set 2.

Figure 2: Data sets for Question 4

Answer:

(a) Both decision tree and NB will do well on this data set because the distinguishing attributes have better discriminating power than noisy attributes in terms of entropy gain and conditional probability. k-NN will not do as well due to relatively large number of noise attributes, which will adversely effect the computation of similarity between two examples.

(b) k-NN will work the best due to the proximity of the examples of the same class to each other. NB does not work well for this data set since the attributes that determine the class boundaries are not independent. Decision tree will have to be large in order to capture the circular decision boundaries, and thus is not the ideal solution.

5. [12 points: 3 for each part] 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 $\rightarrow$ DB = Yes
   - Marital Status = Single $\rightarrow$ DB = Yes
   - Annual Income = Low $\rightarrow$ DB = Yes
   - Annual Income = High, Currently Employed = No $\rightarrow$ DB = Yes
   - Annual Income = Medium, Currently Employed = Yes $\rightarrow$ DB = No
   - Home Owner = No, Marital Status = Married $\rightarrow$ DB = No
   - Home Owner = No, Marital Status = Single $\rightarrow$ 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 ?

   Answers:

   (a) No. The instance {Home Owner = Yes,Marital Status = Single} will trigger the first two rules.

   (b) No. The instance { Marital Status = Divorced, Home Owner = No, Annual Income = High, Currently Employed = Yes } is not covered by any of the rules.

(c) Yes because a record can match two or more rules that give conflicting predictions about the class. For example, the instance HomeOwner=Yes, MaritalStatus=Divorced, AnnualIncome=Medium, CurrentlyEmployed=Yes will trigger rule 1 (prediction: DefaultBorrower=Yes) and rule 5 (prediction: DefaultBorrower=No). If you do not tell the system to prefer one rule to another (i.e., order them), the system will not know how to classify the instance.

(d) Yes, since the rules are not exhaustive.

6. [14 points] Consider the problem of predicting whether a movie is popular or not, 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?

Answer: RIPPER is preferred over k-NN in this problem, since the variables are diverse in nature and thus are expected to be of varying relevance for differentiating between the "popular" and "not popular" classes. k-NN's performance can be adversely impacted by irrelevant attributes, as they can unduly influence the similarity function. RIPPER, like most other rule-based models, as well as decision tree classifiers, performs variable selection using measures such as the Gini index that, although not perfect, allow irrelevant variables to be discarded.