**COMP 4331 Data Mining: Assignment 2**

Name: YEUNG, Man Yin Michael

SID: 20603418 / mymyeung

**Environment:**

OS: Windows 10 家用版

CPU: Intel Core i5-7200U

RAM: 8.00 GB

Model: ASUS UX410UQK

**Running Time:**

ID3 Decision Tree: approximately 1 second

C4.5 Decision Tree: approximately 1 second

Naïve Bayes Classifier: approximately 0.5 second

**Outputs (Predictions) of Different Classifiers**

|  |  |  |
| --- | --- | --- |
| **ID3 Decision Tree**  *max\_height = 5* | **C4.5 Decision Tree**  *max\_height = 5* | **Naïve Bayes Classifier** |
| Test Instance 1 : not\_recom  Test Instance 2 : spec\_prior  Test Instance 3 : priority  Test Instance 4 : very\_recom  Test Instance 5 : very\_recom  Test Instance 6 : priority  Test Instance 7 : not\_recom  Test Instance 8 : spec\_prior  Test Instance 9 : very\_recom  Test Instance 10 : very\_recom | Test Instance 1 : not\_recom  Test Instance 2 : spec\_prior  Test Instance 3 : spec\_prior  Test Instance 4 : spec\_prior  Test Instance 5 : spec\_prior  Test Instance 6 : spec\_prior  Test Instance 7 : not\_recom  Test Instance 8 : spec\_prior  Test Instance 9 : spec\_prior  Test Instance 10 : spec\_prior | Test Instance 1 : not\_recom  Test Instance 2 : spec\_prior  Test Instance 3 : priority  Test Instance 4 : priority  Test Instance 5 : very\_recom  Test Instance 6 : priority  Test Instance 7 : not\_recom  Test Instance 8 : spec\_prior  Test Instance 9 : priority  Test Instance 10 : very\_recom |

**Difference Between Predictions and Reasons, with Predicted Output**

Observations on the results:

1. After observing the predictions of the three classifiers, it is clear that the predictions of ID3 Decision Tree are close. 8 out of 10 test instances returned the same prediction. The remaining 2 test instances with different predictions are highlighted above in green.

2. The C4.5 Decision Tree generally has a different prediction with the other 2 classifiers. 6 out of 10 test instances are different from all other classifiers, while the remaining 4 instances are same among all classifiers. We can see that the 6 different instances all have prediction “spec\_prior”, and these instances have predictions “very\_recom” or “priority” in ID3 Decision Tree / Bayes Classifier.

Summarizing the above, there are only a three types of test instances, they have predictions:

(a) “not\_recom” or “spec\_prior” in all three classifiers

**(instances 1,2,7,8)**

(b) the same in ID3 and Bayes with “very\_recom” or “priority”,

but result “spec\_prior” in C4.5 **(instances 3,5,6,10)**

(c) “very\_recom” in ID3, “spec\_prior” in C4.5, and “priority” in Bayes **(instances 4,9)**

Actually (b) and (c) can be grouped as the same type, which is:

“very\_recom” or “priority” in ID3 or Bayes, “spec\_prior” in C4.5

There are possible reasons for the occurrence of the difference in result:

1. First, we discuss about the drawbacks in each classifier:

ID3 Decision Tree: The information gain measure of this classifier is biased towards attributes with a larger number of values. In this case, this program may bias towards the “has\_nurs” attribute as it has the most values.

C4.5 Decision Tree: This classifier tends to prefer unbalanced splits in which one partition is much smaller than the others.

Naïve Bayes Classifier: This classifier made an assumption that all attributes are independent (class conditional independence). However, this is probably not the case, leading to loss of accuracy. There are always dependencies among attributes (variables), for example in this case it is reasonable (but may not be evident) that:

1. People with convenient finance has more children.

2. People with inconvenient finance has less convenient housing.

The aforementioned issues may affect the accuracy of the predictions, biasing the predictions towards a particular type or set of attributes or intermediate decisions (eg. splitting of nodes in Decision Tree).

2. The C4.5 Decision Tree program has generally different predictions towards the other two. Assuming that the program is not defective, it is possible that such differences are caused by the bias that is mentioned above: that C4.5 Decision Tree tends to prefer unbalanced splits in which one partition is much smaller than the others. This affects the node-splitting mechanism, and thus providing a different result compared to ID3 Decision Tree.

3. As we found that:

1. the predictions of ID3 Decision Tree and Naïve Bayes Classifier are similar, and

2. the ID3 Decision Tree and Bayes Classifier’s programming algorithm is independent,

we can conclude, or make a reasonable guess, that:

1. the bias of C4.5 on the given data set is comparatively large, thus providing comparatively inaccurate predictions (which make predictions far from other two classifiers)

2. the bias of ID3 and Naïve Bayes Classifier is comparatively small, thus providing accurate predictions (which makes them provide similar predictions)

4. By the conclusion above, we can thus proceed into comparing predictions of ID3 and Naïve Bayes Classifier only:

|  |  |
| --- | --- |
| **ID3 Decision Tree**  *max\_height = 5* | **Naïve Bayes Classifier** |
| Test Instance 1 : not\_recom  Test Instance 2 : spec\_prior  Test Instance 3 : priority  Test Instance 4 : very\_recom  Test Instance 5 : very\_recom  Test Instance 6 : priority  Test Instance 7 : not\_recom  Test Instance 8 : spec\_prior  Test Instance 9 : very\_recom  Test Instance 10 : very\_recom | Test Instance 1 : not\_recom  Test Instance 2 : spec\_prior  Test Instance 3 : priority  Test Instance 4 : priority  Test Instance 5 : very\_recom  Test Instance 6 : priority  Test Instance 7 : not\_recom  Test Instance 8 : spec\_prior  Test Instance 9 : priority  Test Instance 10 : very\_recom |

By comparing predictions of only these two classifiers, it is reasonable for me to believe that the predicted output are as follows for data points (instances) 1, 2, 3, 5, 6, 7, 8, 10:

Test Instance 1 : not\_recom

Test Instance 2 : spec\_prior

Test Instance 3 : priority

Test Instance 5 : very\_recom

Test Instance 6 : priority

Test Instance 7 : not\_recom

Test Instance 8 : spec\_prior

Test Instance 10 : very\_recom

because the two classifiers return the same predictions. However, for test instances 4 and 9, the two classifiers return different predictions probably due to their (different) bias:

|  |  |
| --- | --- |
| **ID3 Decision Tree**  *max\_height = 5* | **Naïve Bayes Classifier** |
| Test Instance 4 : very\_recom  Test Instance 9 : very\_recom | Test Instance 4 : priority  Test Instance 9 : priority |

It is mentioned before that ID3 is biased towards attributes with a larger number of values; while Naïve Bayes Classifier is inaccurate when the attributes depend on each other.

In this case, the attributes has 2 to 5 values, while most of them has 3 to 4 values. We can conclude that the amount of values is not very large, and its range is not very wide. I believe the bias (degree causing accuracy) ID3 is comparatively small. However, it is probable that the attributes are dependent on each other, for example many attributes are believed to be dependent on the attribute “finance” (financial status) because money is always a determining factor for decisions (eg. choosing where you reside, deciding give birth to how many children etc). I believe that the above dependencies has led to a comparatively large degree of inaccuracy, and thus led to comparatively inaccurate result.

As there are two classifiers to compare at this stage, and that “very\_recom” and “priority” are neighboring class values:

1. we cannot take average between the predictions of the two classifiers

(because there are no other class values between two neighboring class values)

2. we can only:

(a) believe the result of either classifier (choosing one); or

(b) take a weighted average between the predictions of the two classifiers

which they actually have the same result (“higher weight = chosen” in this case).

As we discussed before above, I believe it is more reasonable that Naïve Bayes Classifier making an inaccurate prediction. Thus, to summarize, I believe that the predictions of ID3 Decision Tree (using this prediction solely) is the more accurate and thus I regard it as the “predicted output”.

|  |
| --- |
| **Predicted Outout** |
| Test Instance 1 : not\_recom  Test Instance 2 : spec\_prior  Test Instance 3 : priority  Test Instance 4 : very\_recom  Test Instance 5 : very\_recom  Test Instance 6 : priority  Test Instance 7 : not\_recom  Test Instance 8 : spec\_prior  Test Instance 9 : very\_recom  Test Instance 10 : very\_recom |