Lab sheet N°10

Classification evaluation and association rules

Exercise 1:

	Accuracy	
Data set	T_{10}	T_{100}
A	0.86	0.97
B	0.84	0.77

Consider a labeled data set containing 100 data instances, randomly partitioned into sets ***A*** and ***B***, each containing 50 instances. We use ***A*** as the training set to train two decision trees, ***T10*** with 10 leaf nodes and ***T100*** with 100 leaf nodes. The accuracies of the two decision trees on data sets ***A*** and ***B*** are shown in the table.

1. Based on the accuracies shown in the table, which classification model would you expect to have better performance on unseen instances?
2. Now, you tested ***T10*** and ***T100*** on the entire data set (A + B) and found that:
   - The classification accuracy of ***T10*** on the data set (A+B) is 0.85
   - The classification accuracy of ***T100*** on the data set (A + B) is 0.87.

   Based on this new information and your observations from the table, which classification model would you finally choose for classification?

Exercise 2:

This exercise highlights one of the limitations of the leave-one-out evaluation procedure.

Let's consider a data set containing 50 positive and 50 negative instances, where the attributes are purely random and contain no information about the class labels. Hence, the generalization error rate of any classification model learned from this data is expected to be 0.5.

Let's consider a classifier that assigns the majority class label of training instances (ties resolved by using the positive label as default class) to any test instance, irrespective of its attribute values. We can call this approach the majority inducer classifier.

Determine the error rate of this classifier using the following methods:

1. Leave one out.

2. 2-fold stratified cross-validation, where the proportion of class labels at every fold is kept the same as that of the overall data.
3. From the results above, which method provides a more reliable evaluation of the classifier's generalization error rate?

Exercise 3:

Consider the data set shown in the following Table:

Customer ID	Transaction ID	Items Bought
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}

1. Compute the support for itemsets {e}, {b,d}, and {b,d,e} by treating each transaction ID as a market basket.
2. Use the results in question 1 to compute the confidence for the association rules {b,d} → {e} and {e} → {b,d}. Is confidence a symmetric measure?
3. Repeat question 1 by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise).
4. Use the results in question 3 to compute the confidence for the association rules {b,d} → {e} and {e} → {b,d}.
5. Suppose ***s1*** and ***c1*** are the support and confidence values of an association rule ***r*** when treating each transaction ID as a market basket. Also, let ***s2*** and ***c2*** be the support and confidence values of ***r*** when treating each customer ID as a market basket. Discuss whether there are any relationships between ***s1*** and ***s2*** or ***c1*** and ***c2***.