TUTORIAL 2 - MACHINE LEARNING

DONE BY:

ALBERTO CASTELLANO MACIAS with NIA: 100414242

ALEX OSWALD with NIA: 100479018

NARIM WOO with NIA: 100478751

TUTORIAL 2 - MACHINE LEARNING

EXERCISE 1

1. How many input attributes does the file have? What type are they?
The file has two input attributes: name and class, both of type nominal.

2. Could a machine learning algorithm identify a function able to predict the class of an instance with the current data? Why?
This would not be possible with the current data because you only have the attribute name to determine the class. However, by including more attributes such as the number of vowels, consonants, dots, etc, a machine learning algorithm could definitely work.

EXERCISE 2

1. On the Classify tab, select the classifier trees/ID3 1 . On Test options select Use training set and press the Start button to generate the model. How good are the results?
We can see that we have a 100% accuracy and that we have not misclassified anything, so the results seem perfect.

EXERCISE 3

1. Consider at least 6 attributes that you think may be relevant to solve this problem. These attributes should be extracted just from the previous attribute name. Write and explain them. Why have you chosen these attributes?
From the attribute name we can probably extract the following attributes:

- number of dots in the whole name
- number of spaces in the whole name
- number of vowels in the whole name
- number of consonants in the whole name
- length of the surname
- whether the number of letters in the whole name is even or odd

2. Open the file badges1.arff with Weka. How many input attributes does the file have? What type are they?
The file has 9 input attributes:

- dots: numeric type
- spaces: numeric type
- words: numeric type
- consonants: numeric type
- length: numeric type
- even_odd {0,1}: nominal type
- class {+,-}: nominal type
- first_char_vowel: nominal type
- name: nominal type
- class: nominal type

TUTORIAL 2 - MACHINE LEARNING

3. Which type of statistical information is shown? Press the "Visualize All" button. What do you see?
We can see different plots for each attribute, each of them with a different distribution. For example, we see that the attribute "class" is evenly distributed as there are 144 names of class "-" and 150 of class "+". On the other hand, the attribute "spaces" is not evenly distributed as there are many more names with 1 space than 2 spaces.

4. Generate a classifier using tree/ID3. What happened? What can you do to avoid this problem with ID3?
It does not allow us to perform a test. This is because ID3 only works with nominal attributes and here we also have numeric ones. To avoid this error we can discretize the numeric ones and turn them into nominal.

EXERCISE 4

1. Select the filter Filter/unsupervised/attribute/Discretize, fix the number of bins to 5, and apply the filter to the data set. What effect does this filter have?
This filter converts the numeric attributes into nominal ones.

2. How many instances are correctly classified? What is the percentage?
There are 236 correctly classified instances, which is 80.27% of the total number.

3. What do you think the confusion matrix shows?
It shows how many instances are correctly classified as + or -, which are the values in the diagonal (93 for - and 143 for +), and the other 2 values are the incorrectly classified instances.

4. How many instances of each type are misclassified?
7 instances for - and 51 instances for +

5. Click the More Options button and select the Output Predictions (PlainText) option. Classify it again and check the results. Which is the first misclassified instance? Why?
The first misclassified instance is number 7 as it is classified as + and actually belongs to -. We know it is misclassified because it shows an error in the prediction of its class.

6. How would the instance "Donald Trump" be classified? Which are the attributes of this name? What happens to the values of this instance if you use the previous filter?
The instance 'Donald Trump' has:

- length: 12
- consonants: 8
- dots: 0
- even_odd: 0 (even)
- spaces: 1
- first_char_vowel: 0 (no)

Therefore, it is of class +.
With the previous filter, the values would have been different as some of the attributes were numeric.

TUTORIAL 2 - MACHINE LEARNING

EXERCISE 5

1. What model does ZeroR generate?
ZeroR generates a model which predicts that every value would take the value of the class with more instances, which in this case is +

2. What is the success rate of this model?
The success rate was 51.02%.

3. How would the instance "Donald Trump" be classified?
It would be classified as + as well as all instances based on this model.

EXERCISE 6

1. How many leaves does the tree generated by J48 have?
21 (20)

2. How many instances are correctly classified?
229 instances (287)

3. What is the success rate?
77.89% (97.619%)

4. How many instances of each type are incorrectly classified?
15 misclassified instances for - and 50 for +
7 misclassified instances (2.381%) were yielded: 4 misclassified as + and 3 misclassified as – according to the confusion matrix.

```
   a   b   <-- classified as
 140   4 |   a = -
   3 147 |   b = +
```

5. How would the instance "Donald Trump" be classified?
It would be classified as + too.

6. Would you prefer this model or the one generated by ID3? Why?
We would prefer this model generated by J48 as it has a higher accuracy and less error than with id3.

7. Have we found the perfect function to label the instances? Why?
Even though it has a high accuracy, no used model has yet fit our dataset to perfection.

EXERCISE 7

1. Go back to the preprocess window and select the filter Filter/unsupervised/attribute/AddExpression to generate a new attribute that computes the number of vowels.
a2 - (a5+a6+a7)
length - (num consonants + num spaces + num dots)

2. Store the data set badges-Ej7.arff.
done.

TUTORIAL 2 - MACHINE LEARNING

3. Could you tell which is the most common range of vowels in the provided file?
The most common range was of names with 4 vowels, and it contained 94 names. The model represented this bin as (3.455,4.182].

4. Generate a new classifier with J48 and the previous data set. In this case you have to select the class in the drop-down menu of the Classify tab.
done.

5. Write down the percentage of correctly classified instances and the confusion matrix.
100% (294) Correctly Classified Instances

```
a   b
144   0 |   a = -
  0 150 |   b = +
```

6. Right click on the generated model that appears in Result list. Visualize the tree by clicking on Visualize Tree. What does the number on the leaves show?
The number on the leaves shows the number of instances when the value of the 'number of vowels' is less than 4 and greater. It displays the same value as the confusion matrix.

7. Given these results, what features do you think attributes should have to maximize the success of machine learning algorithms?
The most valuable attributes to classify a set of data are those that are positively correlated with other aspects of the model and of the classifications. In this example, by adding the number of vowels as a classifier, we're inclining the model toward name length and also adding correlations to other types of names that may include those beginning with a vowel as well. Additionally, according to the experiments conducted here, attributes with numerical data tend to produce the best results over strings which may be too hard to process and other categorical data types that may have too many categories.

EXERCISE 8

1. Load in Weka the file adult-data.arff.

2. How many input attributes does the file have? How many training instances?
It has 15 input attributes and 32561 training instances.

3. Execute the classifier J48. Select the option "Cross-validation" in Test Options. What results appear? Explain the result.
We obtained an accuracy of 86.21% by using a 10-fold cross-validation on the attribute 'salary'. This means that 28071 instances were correctly classified while 4490 were incorrectly classified. In addition, the output tree has 564 leaves.

4. Now, evaluate the classifier only in instances from the adult-test.arff file. To do so, select "Supplied test set" in Test Options. What results appear? Are these results comparable with the previous ones? Why?
We obtained 85.84% of accuracy, which means that there are 13977 correctly classified instances and 2304 incorrectly classified ones out of a total of 16281.
Here, we obtained a slightly worse accuracy but it is very much comparable with the previous results as they are really similar.

TUTORIAL 2 - MACHINE LEARNING

5. Go back to Preprocess and click the attribute salary. What proportion of the data is from each class? Do you think this percentage is suitable for a machine learning algorithm to properly learn?
There is 24% of the data belonging to the class of >50k while there is 76% of the data belonging to the class of <=50k, approximately.
In principle, the data should be more balanced in order for a machine algorithm to properly work. However, this proportion might be enough.

6. Now we are going to modify the training instances to have a similar proportion of each class. To do so, select supervised/instance/Resample changing the parameter biasToUniformClass to 1.0 What happens with the salary attribute? What does happen with the number of training instances?
The proportion of the data of each class of the salary attribute changed to 50% each.
The number of training instances changed to 32560, the closest even number, so only one instance was removed.

7. After applying this filter, evaluate again with cross-validation and supplied test set the algorithm J48. What is the result? Is it better or worse?
Using cross-validation the accuracy of the results increased about 1%, to 87.18%.
However, using the supplied test algorithm, the accuracy of the results decreased noticeably, to 80.56%. This makes it worse.

8. Finally apply the standardization filter unsupervised/instance/Normalize for the numeric attributes. What are the results?
As the name of the filter suggests, these attributes were normalized, so now they take values between 0 and 1.

9. After applying several filters to the data as you have just done, do you think this helps the learning process? Why?
This definitely helps the learning process as we have learned how to play with our data and to distinguish the different methods to achieve a better accuracy in our results.

10. What are the best results you have obtained? Justify it.
The best results we obtained, based on a higher accuracy of our results, is the one using a 10-fold cross-validation and using the Resample filter. However, it wasn't by too much.

11. Save the data set as badges-Ej8.arff.

CONCLUSIONS

During this tutorial, we have gained an insight into the machine learning program Weka. We have learned how to play with the different filters, classifiers, etc, in order to modify our data and obtain much better results. We have also realized that there are still more ways to maximize the accuracy in our results and that bid datasets can be understood very easily by using this application. We believe it has true potential and will be very useful for us in the future.