Coursework 8

Fortunat Mutunda

March 30, 2016

```
library("rpart")
```

1. Read - http://www.r2d3.us/visual-intro-to-machine-learning-part-1/ What is the quality of the classifier? Can you understand when it works well and when not?

2. Use this small data example and build a decision tree (manually, explaining all steps/choices).

ord.	Outlook	Temp	Humidity	Windy	Play
1.	Sunny	Hot	High	FALSE	No
2.	Sunny	Hot	High	TRUE	No
3.	Overcast	Hot	High	FALSE	Yes
4.	Rainy	Mild	High	FALSE	Yes
5.	Rainy	Cool	Normal	FALSE	Yes
6.	Rainy	Cool	Normal	TRUE	No
7.	Overcast	Cool	Normal	TRUE	Yes
8.	Sunny	Mild	High	FALSE	No
9.	Sunny	Cool	Normal	FALSE	Yes
10.	Rainy	Mild	Normal	FALSE	Yes
11.	Sunny	Mild	Normal	TRUE	Yes
12.	Overcast	Mild	High	TRUE	Yes
13.	Overcast	Hot	Normal	FALSE	Yes
14.	Rainy	Mild	High	TRUE	No
15.	Overcast	Cool	High	FALSE	No

Providing that there is mild, overcast, high humidity and high wind weather - should one play tennis or not?

3. Use the Cars data set and apply decision trees for classification. Describe the tree. (you can use R, or Weka (install Weka from here), or python...). Compare the decision tree approach to the association rules derived from the same data.***To make your life easier, we recommend you remove observations with two infrequent classes - good and v-good. You can get the resulting dataset here in R, you can use library rpart to build the trees and rpart.plot to visualize them***

4. Use the same cars data set. Apply decision trees and Naive Bayes classifiers on the same data. Can you confirm that one method is better than the other in some way? Perform 10-fold cross-validation. Provide final results as 2x2 tables of TP. FP, FN, TN and some measures - accuracy, precision, recall.

5. Use the Titanic data set - compare your classifiers learned from Titanic data - decision trees, Bayes rules, association rules - and try to characterise the rules observed in data using these approaches. How can they be interpreted against each other?

6. (Bonus 1p) How to detect and avoid overfitting? What is the good (optimal?) size of the decision tree classifiers? Use the above Cars data, and for comparison use one of the two data sets - the Mushroom (LINK) or the Connect 4 (LINK).