ASI assessed exercise 2015/2016

22nd April 2016

Introduction and Instructions

In this work you will analyze two datasets on wine quality available from the UCI machine learning repository:

http://archive.ics.uci.edu/ml/datasets/Wine+Quality

Listed below are various exercises to undertake. Note that in each case you should implement the algorithms yourselves - you may not use other implementations - and should submit all of your code.

Submission

You are free to use any language of your choosing but it is your responsibility to ensure that we can compile (if necessary) and run your code. We recommend you use either Matlab or Python. Please submit either:

- Your code (including instructions for running - there should be one script that answers all the questions) and a .pdf report documenting your answers to the exercises…

- Or (preferably) a single iPython notebook or Matlab Script that we can run. If you take this route, please *also* submit a .pdf output of the script (iPython: print the html to pdf; Matlab publish the script to html and then print the html to pdf). Your notebook or script should include any text descriptions required in the answers. (Matlab's cell mode allows you to produce nice text and latex etc; iPythons markdown cells allow you to add text)

Submission will be through the ASI collaborative space.

The deadline is Friday 27th May 2016 at 4:00pm.

Exercises

Note (code) and (words) before each task show whether the corresponding part is coding, or writing.

1. (code) Download the red and white wine .csv files and import them. The first 11 columns are features and the final column is the target. [3]

2. (code) Plot bar-plots of the number of examples with each target value for the two datasets. [3]

3. (words) Comment on these distributions. How might they effect the analysis? [4]

4. Linear regression:

    (a) (code) We will now concentrate on the red wine data. Randomly split the data into a training and test set with 70% of the examples in the training and 30% in the test. [5]

    (b) (code) Fit a linear regression to the training data. [5]

(c) (code) Make a scatter plot the predictions versus the true targets for the test set and compute the mean squared error on the test set [5]

(d) (words) Suggest a benchmark that you could use to decide if this mean squared error value is good. [5]

(e) (code) Implement your benchmark. [5](f) (words) Briefly discuss the linear regression performance with respect to the benchmark. [5]

5. Regularized linear regression:

(a) (code) Implement regularized least squares and make a plot of the test performance versus the regularization parameter. [10]

(b) (words) Describe why this is not a good way of determining the value of the regularization parameter. [5]

(c) (code) Implement a 10-fold CV on the training data and use this to determine the value of the regularization parameter. Quote the optimal value, and the performance at this value on the test set. [10]

(d) (words) Compare the performance with the standard linear regression case, discussing possible reasons for any change in performance. [5]

6. Classification

(a) (words) Describe one limitation of using regression for this particular task? [2]

(b) (words) Pick either Naive Bayes or KNN. Describe a positive and a negative feature of your classifier with respect to this class. [2]

(c) (words) Describe any data pre-processing that you suggest for this data and your chosen classifier. [2]

(d) (code) Implement your classifier and optimize its parameters. Make sure your optimization is clearly commented. Use classification accuracy as your figure of merit [15]

(e) (code) Display the confusion matrix on the text data [5]

(f) (words) Discuss the performance and suggest a way in which they could be improved [4]

7. Bonus question

(a) (words) The data was originally published in *Modeling wine preferences by data mining from physicochemical properties*, Cortez et al, Decision Support Systems 47(4). Write a review of this paper, focusing on the Machine Learning methods chosen and Assumptions taken. Include a description of how you might advance the work. This should be submitted as a separate PDF file.

Numbers at the end of each section are the number of marks available.

Be concise - a complete solution should be around 5 pages (including figures) and certainly no more than 10.