Homework 1

DSE 220: Machine Learning

Due Date: 13 April 2017

1 Instructions

The answers to the questions should be submitted on gradescope and the code should be submitted on github by 13 April 2017. You don't need to explain your approach (unless specified) so please be concise in your gradescope submission. To obtain full marks for a question, both the answer and the code should be correct. Completely wrong (or missing) code with correct answer will result in zero marks. Please make sure that your code is clean and well segmented for each question.

2 Data

Download the 'wine_modified' and 'wine' (train, validation and test) data from the github. Use the 'wine_modified' data for the data preprocessing section and the wine train, validation and test data for the other two sections.

3 Data Preprocessing

The questions in this section are sequential steps. So use the data obtained after Question 1 for Question 2 and so on.

Question 1: Remove the rows with missing labels ('class') and rows with more than 7 missing features. Report the remaining number of rows. (1 mark)

Question 2: Remove features with $> 50\%$ of missing values. For other features with missing values fill them with the mean of the corresponding features. Report the removed features (if any) and standard deviation of features with missing values after filling. (2 marks)

Question 3: Detect and remove rows with any outliers/incorrect values in features 'alcohol' and 'proline' (if any). Clearly state the basis of your removal. (1 mark)

4 Decision Trees

Note: When predicting for the test data, you should train the model again using train + validation data.

Question 4: Train Decision Tree model on train data for criterions = {'gini', 'entropy'} and report the accuracies on the validation data. Select the best criterion and report the accuracy on the test data. (1 mark)

For information on gini criterion, you can refer:
`http://statweb.stanford.edu/~jtaylo/courses/stats202/restricted/notes/trees.pdf`

Question 5: Use the criterion selected above to train Decision Tree model on train data for min_samples_split={2,5,10,20} and report the accuracies on the validation data. Select the best parameter and report the accuracy on the test data. (2 marks)

Question 6: Use the parameters selected above (Q4 and Q5) to train Decision Tree model using the first *20, 40, 60, 80* and *100* samples from train data. Keep the validation set unchanged during this analysis. Report and plot the accuracies on the validation data. (2 marks)

5 Nearest Neighbor

Normalize Data: Normalize features such that for each feature the mean is 0 and the standard deviation is 1 in the train+validation data. Use the normalizing factors calculated on train+validation data to modify the values in train, validation and test data.

Question 7: Train k-nn model on train + validation data and report accuracy on test data. Use Euclidean distance and k=3. (1 mark)

Question 8: Train the model on train data for distance metrics defined by ℓ_1, ℓ_{inf}, ℓ_2. Report the accuracies on the validation data. Select the best metric and report the accuracy on the test data for the selected metric. Use k=3. (1 mark)

Question 9: Train the k-nn model on train data for k=1,3,5,7,9. Report and plot the accuracies on the validation data. Select the best 'k' value and report the accuracy on the test data for the selected 'k'. Use Euclidean distance. (2 marks)

Question 10: Instead of using full train data, train the model using the first *20, 40, 60, 80* and *100* data samples from train data. Keep the validation set unchanged during this analysis. Report and plot the accuracies on the validation data. Use Euclidean distance and k=3. *Note: Don't shuffle the data and use only the 'first n samples', otherwise your answers may differ.* (2 marks)