ECM3420 Learning from Data (Year 3 only)

Course Work 1: Supervised Learning

Submission Deadline: 12:00 midday on Wednesday 21st October (Week 5)

Weight: 20%

Data and Problem Description

In this coursework, your task is to develop a machine learning classifier using “decision trees” and “k-nearest neighbors”. Your model is to support clinicians in identifying patients who are likely to have Diabetes. The dataset has 9 attributes in total including the “target/label” attribute. The full dataset is available on ELE under assessment coursework 1. The dataset consists of the following:

Attributes:

* Number of times pregnant
* Plasma glucose concentration 2 hours in an oral glucose tolerance test
* Diastolic blood pressure (mm Hg)
* Triceps skin fold thickness (mm)
* 2-Hour serum insulin (mu U/ml)
* Body mass index (weight in kg/ height in m2)
* Diabetes pedigree function
* Age (year)

Target:

* Class variable (0 or 1)

Your task

1. **Exploratory data analysis:** Before applying any machine learning models, it is important to understand the data. In this task you should:
   1. Draw two charts that should help with understanding the data.
   2. Perform correlation analysis between 2 input attributes.
2. **Classification:** using n-fold cross-validation (with n=5)
   1. Build a Decision Tree (DT) classifier for this dataset and provide the “testing” for each fold and average total. Compare the accuracies of the experiment above when you change the criterion from Gini impurity (“gini”) to information gain (“entropy”).
   2. Build a K-Nearest Neighbors (KNN) classifier and provide the “testing” for each fold and average total. Compare the accuracies of this experiment when you use K = 1, 3 and 5.
3. **Classification parameters DT:** using a split dataset (70% training and 30% testing)
   1. How does increasing the minimum number of samples required to split an internal node parameter in the DT algorithm (i.e. min\_samples\_leaf = 2, 3, 4 and 5) affect the accuracy on the test set? Show your results in a tablet and using a suitable chart.
   2. How does increasing the maximum depth of the decision tree parameter (i.e. max\_depth = 3, 4, 5 and 6) affect the accuracy on the test set? Show your results in a tablet and using a suitable chart.
4. Reflection
   1. Write around 600 words reflection on the results. You are expected to cover the following:
      * Interpretation/explanation of the experiments and results.
      * Reflection on the results, advantages and limitation of the methods used.
      * Justification of the correlation attributes selection.
      * Possible improvements.

Submission

You are expected to complete this assessment using a Python 3 Jupyter notebook and the packages Sklearn, Numpy, Matplotlib and Pandas.

You are expected to submit the following items:

* Use Python 3 and submit a single Jupyter notebook (please use the attached template).
* You can write your justification section at the end of the Jupyter file or in separate file and to be submitted as a separate PDF file.
* PDF version of the Jupiter notebook file.

Please submit your BART assignment here: <https://bart.exeter.ac.uk/>

It is vital that you check the time and date of your deadline on BART prior to your submission. Assignments that need to be submitted through BART must be done so using the link above. BART will then guide you through the steps you need to complete for your submission. Failure to do so will result in your mark being capped for a late submission.  [Full details are on the ELE page]

Marking Scheme

* [20%] Exploratory data analysis.
  + 10% correctness of the correlation analysis
  + 10% Quality and selection of the charts
* [30%] Classification
  + 20% on the correctness and quality of the Python code and results
  + 10% Quality and selection of the charts and tables.
* [20%] Classification parameters
  + 20% on the correctness and quality of the Python code and results
  + 10% Quality and selection of the charts and tables.
* [30%] Justification:
  + Interpretation of the experiments and results.
  + Reflection on the results, advantages and limitation of the methods used.
  + Justification of the correlation attributes selection.
  + Possible improvements.
  + Presentation style and references quality.