Practical Skills Assessment 2

Due: 12:00, 06 January 2022

Contribution towards coursework marks: **60%**
Submission: COM761 area in Blackboard

Keep a copy of all submitted coursework – i.e. your computer files.

Learning Outcomes:

- Evaluate and compare supervised and unsupervised learning algorithms on problems involving real datasets
- Diagnose and rectify common problems that affect the performance of machine learning algorithms
- Design machine learning experiments and justify the procedures employed
-

This assessment is to be completed individually

When submitting your assessment, you are agreeing to the following statement:

I declare that this is all my own work and does not contain unreferenced material copied from any other source. I have read the University's policy on plagiarism and understand the definition of plagiarism. If it is shown that material has been plagiarised, or I have otherwise attempted to obtain an unfair advantage for myself or others, I understand that I may face sanctions in accordance with the policies and procedures of the University. A mark of zero may be awarded and the reason for that mark will be recorded on my file.

The University policy on **plagiarism** is available at
http://www.ulster.ac.uk/academicservices/student/plagiarism.pdf

Requirements

This assessment requires you to design and analyse machine learning experiments, making use of appropriate Python libraries. These experiments will include components such as algorithm selection, data pre-processing, hyperparameter tuning, evaluation of algorithm performance, and feature attribution. The experiments will cover both supervised and unsupervised learning problems. It will be assessed via a Jupyter notebook, which will be submitted electronically through Blackboard on 6th January 2022. Individual feedback will be provided within 20 working days of submission.

You are required to document the experiments, results and discussions. Any external sources used in the design or implementation process should be appropriately cited and referenced.

Coding requirement

All code should be completed using Python as the programming language. You should use Scikit Learn, NumPy and Pandas. You are free to use imported graphical libraries such as MatPlotLib or Seaborn (for generating graphs). You are also free to import Scikit-Learn contribution packages such as Imbalanced Learn. If you wish to use other external libraries please check with me in advance.

Your code should have a logical structure and a high level of readability and clarity. Please comment your code and put all code into functions. Your code should be efficient and should avoid duplication.

Part 1 Bayesian Machine Learning (20%)

In this question, you may reuse any of the solution code provided for the Week 6 labs on Gaussian Process regression and Bayesian optimization. You should modify this code in any way you see appropriate.

(a) Choose any ONE of the following functions: **Problem05, Problem07, Problem09, Problem10, Problem14, Problem21** from http://infinity77.net/global_optimization/test_functions_1d.html and represent it as a Python function.

The function should take in a variable *x,* and return the value of the function evaluated at *x*. You should name your function similar to its name on the webpage, e.g. `problem05`

Then:

- Sample the function at five random points within its bounds.
- Use these data to train a Gaussian Process regression model (using the RBF kernel).
- Produce a figure which shows: the five points sampled, the true function, the Gaussian Process regression prediction trained on the five sampled points, and the standard deviation s(x) in the prediction.

[5 marks]

(b) You are now going to use the **Probability of Improvement** acquisition function to minimize your chosen function. Implement the probability of improvement acquisition function so that it uses an arbitrary target *T* as follows:

$$POI(x) = \Phi\left(\frac{T - \hat{y}(x)}{s(x)}\right)$$

In this form, POI(*x*) gives the probability that *y*(*x*) is less than some target value *T* for the objective function. Note that *T* must be less than y_{min} where y_{min} is the lowest function value sampled so far. Values of *T* close to y_{min} will favour exploitation, whilst values of *T* much lower than y_{min} will favour exploration (searching unexplored regions of design space, where uncertainty is high).

Experiment with your chosen function to find three values of *T* that illustrates how POI shifts from exploitation to exploration as *T* becomes more ambitious (lower), and demonstrate this by producing a suitable plot. **[6 marks]**

(c) Typically we wish to balance exploration and exploitation during an optimization search. For example, a common strategy is to focus more on exploration at the beginning of an optimization search, and then gradually become more exploitative as the iterations proceed.

Design a strategy to vary the value of *T* used in the POI acquisition function at each iteration of a Bayesian optimization search, in order to effectively balance exploration and exploitation. The search should begin with five initial random points, and then proceed iteratively using the POI acquisition function to select further points to evaluate, with the value of *T* as set by your strategy. In total, the POI should be used to select TEN points for evaluation. So at the end of a search, the function will have been evaluated 15 times in total (5 initial random points, plus 10 further points selected using POI).

Present the results of using your strategy on your chosen function and one further test function from the list of functions listed in (a), presenting your results in whatever way you best see fit. **[9 marks]**

Part 2 Supervised Learning (20%)

The dataset that we will use for this task will be provided and you can downland it from the Blackboard.

1. **[Pre-processing]** In this phase you will be focussing on preparing the dataset, i.e. loading, exploring dataset, displaying, removing noise and normalisation etc. Then, split the data for training and the remainder for test. Take appropriate measures to ensure that the test set is not biased in any way. Collect and record statistics on the resulting training and test sets.

   [5 marks]
2. **[Features Extraction**] Extract the features you will need for the remainder of the analysis. You may revisit this stage many times as you become more familiar with the dataset and the kinds of features that may be useful for the classification task. The features you choose may affect the performance of the final classifier. Choose something you think is

reasonable to start with and later you can experiment with alternatives on the test set.

[4 marks]

3. **[Supervised classification]** Select two classification models (such as KNN, SVM, decision tree or Naïve Bayes) for comparison. Train a supervised classification model on your features. Use a validation set or cross-validation to compare the performance (confusion matrix, accuracy, precision, recall, specificity, F1-score and AUC) of different models. Create plots to compare a subset of the models that you investigated during model selection. Retain the most effective model for evaluation.

[8 marks]

1. **[Summary and Discussion]** Summarise your findings, discuss the factors that could have impacts on your results and propose potential improvement.

[3 marks]

Part 3 Unsupervised Learning (20%)

Use the dataset provided and clustering techniques in Part 3.

1. **[Pre-processing]**: The first column of the data is the label information, and you shall exclude them in clustering. You shall pre-process the data, such as inspecting the data, visualising the data distribution, removing noise, and normalising the data.

[4 marks]

2. **[K-means clustering]**: apply the K-means clustering algorithm to analyse the data. You shall (1) tune the parameter setting, select the optimal cluster number, and (2) visualise the centroids and the clustering results.

[5 marks]

3. [**hierarchical clustering]:** apply the clustering algorithm to analyse the data. You shall (1) plot a hierarchical clustering dendrogram, (2) Compare characteristics of different linkage methods for the hierarchical clustering results, and (3) estimate the correct number of clusters using clustering validity indices, also display the results.

[6 marks]

4. **[Summary and Discussion]** Compare the performance of hierarchical and K-means clustering and discuss the results. Also examine the obtained partitions with the known labels. Are samples belonging to the same classes generally clustered together? Can outliers or exceptions be identified?

[5 marks]

Submission

1. The assignment is to be submitted electronically to Blackboard Learn on or before 06 January 2022 at noon.
2. The submission should contain and document all experimental design, process, results and discussion.
3. Only **one** copy of the work should be submitted, with **student name and student IDs clearly marked on all the submitted work**.
4. Note: When submitting an Assignment in Blackboard, particularly ones which include file attachments, you should allow for time to confirm your submission. There are two ways that you can tell your Assignment has been submitted successfully:

a) an on-screen confirmation message when you submit the assignment, or

b) in your My Grades area, an exclamation mark (!) will appear next to any assignment that has been successfully submitted.

You are recommended to take and keep a screen shot of your submission confirmation message. For full information go to Blackboard and see Blackboard Learn Student Orientation Course -> Submitting Assignments