Faculty of Engineering, Environment and Computing

7135CEM – Modelling and Optimisation under Uncertainty

Coursework Assignment Brief

Module Title **Modelling and Optimisation under Uncertainty**	**Individual**	Cohort: Resit 2020	Module Code **7135CEM**
Coursework Title **Coursework Assignment**			Hand out date: **24th September 2020**
Lecturer **Dr Ali Daneshkhah and Prof. Vasile Palade**			Due date and time: **11/12/2020, 18:00**
Estimated Time (hrs):	Coursework type: **Written Assignment**		% of Module Mark **100%: each task of this CW worth 50%.**

Submission arrangement online via AULA/CUMoodle:
Submit before 1800, late work will receive a mark of zero.
File types and method of recording: **Submit Two Word or pdf documents (or similar). One for each of the Two Tasks (see below).**
Mark and Feedback date: **21/12/2020**
Mark and Feedback method: **given on each script.**

Notes:

1. Please notify your registry course support team and module leader for disability support.
2. Any student requiring an extension or deferral should follow the university process.
3. The University cannot take responsibility for any coursework lost or corrupted on disks, laptops or personal computer. Students should therefore regularly back-up any work and are advised to save it on the University system.
4. If there are technical or performance issues that prevent students submitting coursework through the online coursework submission system on the day of a coursework deadline, an appropriate extension to the coursework submission deadline will be agreed. This extension will normally be 24 hours or the next working day if the deadline falls on a Friday or over the weekend period. This will be communicated via email and as a CU Moodle announcement.

On completion of this module the student should be able to:

1. Apply supervised and unsupervised learning applications using Gaussian process emulators.
2. Apply Dirichlet processes for unsupervised learning applications
3. Develop the knowledge and skills necessary to design, implement and apply the Graphical models to solve real world applications.
4. Evaluate the applications of fuzzy systems and their usage in hybrid intelligent systems, in combination with evolutionary computing and other machine learning methods.
5. Apply evolutionary computing methods to develop solutions for the real world optimisation problems and and appraise their advantages and limitations.

Task and Mark distribution:

This coursework consists of two tasks and you should attempt both and submit one Word or pdf file (or similar) for each task. Each task is worth 50 marks and the marks breakdown for each task is provided with each task. This coursework contributes 100% to your overall module mark.

Task 1: Advanced Machine learning algorithms for solving real-world problems in Regression, Classification, modelling data, and text mining

Individual Research Paper: 50% of the module mark

Context

During this module, you learned about different advanced machine learning techniques, associated concepts and applications. We explored the Gaussian process model, which is computationally efficient method for Regression, Classification, optimization, etc. We have also covered the Bayesian networks as promising tools for modelling the data with complex dependency structure. Finally, you have learned how to use Dirichlet Latent processes for unsupervised learning applications, particularly text mining.

In this assignment, you will have to select an application related to a regression, classification, modelling un-structured data, or text mining problem, and explore how best to apply the machine learning algorithms to solve it. The selected application for each of the methods mentioned above should have the following features:

1. **Gaussian Process regression and Classification**: The application selected for any of these two methods must consist of **at least four input variables** and a **single output variable**. You must also implement Gaussian process classification by appropriately define a threshold on the output variable to create a binary or multiple classes first, and then apply the Gaussian process classification on the categorized output.
2. **Bayesian network**: If you are choosing an application for this method, this application must consist of at least **eight random variables**. The random variables could be all discrete or continuous or hybrid.
3. **There is no restriction** on selecting the application to apply the Latent Dirichlet allocation model for topic modelling.

There are some potential projects listed below, which could be studied to get some ideas. However, please absolutely feel free to come up with your own. I strongly recommend you to come up with your own idea by reviewing some relevant articles.

1. This dataset from the UCI repository is quite interesting. The task is to predict the depth in the body (effectively, the depth along the spine) given the properties of a two-dimensional "slice" of the body. The hard part about this problem is that it is actually the output causing the input rather than the other way around. I have not had luck designing a good regression method for this data. Can you do this?
2. Find a Bayesian interpretation of elastic net regularization, and compare this method for regression against "standard" Bayesian regression (with a Gaussian prior) on a dataset of your choosing.
3. Probabilistic PCA using Gaussian Process is a Bayesian interpretation of the classical PCA algorithm for dimensionality reduction. Implement Gaussian Process based PPCA in

Python, R or Matlab, and compare its performance with other methods (such as "standard" PCA) on a dataset of your choosing.

4. Bayesian optimization is very important issue with a wide range of applications. However, this was not fully studied during lectures, but it can be easily implemented using Gaussian Process. The Python codes and some examples can be found here!
5. The squared exponential covariance is widely used for Gaussian process regression. It is probably used in 90+% of all GP publications. That said, it is widely believed to be "too smooth" for many real-world regression tasks. Compare the squared exponential covariance versus the Matéern covariance on several datasets via Bayesian model selection. How often is the squared exponential the right choice?
6. Latent Dirichlet allocation (LDA) is a Bayesian method for creating "topic models" of text documents. There are plenty of interesting text datasets available (e.g., DBpedia could be a good resource!). One idea would be to compare the behavior of LDA with other techniques, such as latent semantic analysis.

You may be able to get relevant dataset and ideas by visiting the following sites:

- This compentition site consists of some relevant data, and the relevant ideas could be developed by analyzing this data. Check also dataset in Kaggle competitions.
- This website has a fantastic compilation of 100 interesting, relevant datasets from all sorts of application areas.
- The creators of libSVM have also compiled a great list of datasets, all in a standardized format. The libSVM codebase also includes *libsvmread* for reading these in MATLAB.
- The UCI Machine Learning Repository is a mainstay in machine-learning research. There is a wide range of datasets there from many different application areas and with many different properties (large, small, high-dimensional, low-dimensional, classification, regression, etc.).
- DBpedia is an amazing resource that automatically extracts structured data from Wikipedia. They have all sorts of data available for download in convenient formats. This tool can be used to extract labeled graphs from DBpedia, but there is so much more you could do.

The purpose of this coursework is to

- Examine the fundamental concepts of machine learning, their implementation and application.
- Perform appropriate preparation of a dataset and evaluate the performance of different learning algorithms on this dataset.
- Gain practical experience in selecting machine learning algorithms for solving a real-life Regression, classification, modelling data with complex dependency structure, or text-mining problems.
- Demonstrate effectiveness in project teamwork and leadership.

You will be required to:

- Work individually, and submit progress on your work regularly to get formative feedback and improve the final submission;

Your final submission on TASK 1 will include a scientific paper of up to 8 A4 pages (written individually), based on the experience and results gained during the project work. You are encouraged to target a certain conference or journal and submit the proposed paper to it. Submission guidelines can be found on the conference or journal web page you choose to submit.
List of reputed conferences and journals:

1- IJCNN Conference
2- NIPS Conference
3- International Conference of Machine Learning
4- Machine Learning Journal
5- Neural Networks Journal
6- Others (please let us know)

The paper should broadly include the following sections:

- Abstract
- Introduction (where you introduce the problem along a short literature review of related work; if the literature review is longer, it is recommended to be a section on its own)
- Problem and Data set(s) description (where you describe in detail the problem you want to solve and its significance)
- Methods (where you shortly describe the machine learning methods and/or other methods employed to solve the problem)
- Experimental setup (including data pre-processing, feature selection and extraction)
- Results
- Discussion and Conclusions
- References

These are generic section titles, which you may adapt appropriately to the application/problem that is being investigated. You may include sections describing modifications of algorithms or developments that are novel and specific to your work. You may include figures, tables, pseudo-code, and appendices with the actual code that has been developed.

More information of how to write a paper is available at the following link: "Crafting Papers on Machine Learning", by Pat Langley.
You will need to follow the formatting guidelines of the IEEE Manuscript Template for Conference Proceedings (A4)

The project general guidelines and milestones:
Please note, the following guidelines are good practice and should lead to better result, but you have the freedom to pick whatever is suitable for your style:

- You have to select a real-world regression, classification, modelling data with complex dependency structure, or text-mining problems and one or more appropriate dataset(s) as suggested above. You may also use the following links, which have numerous problems and datasets:
  - UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/;
  - ICML 2019 accepted papers: https://icml.cc/Conferences/2019/Schedule?type=Poster;
  - Kaggle competitions: http://www.kaggle.com/competitions;

- Stanford machine learning projects:
  http://cs229.stanford.edu/projects2013.html,
  http://cs229.stanford.edu/projects2012.html,
  http://cs229.stanford.edu/projects2011.html,
  http://cs229.stanford.edu/projects2016.html .

- You will write a proposal (maximum of 1 A4 page), giving the title of the project, your name, the description of the problem and the plan of the work. You will need to submit this proposal to your tutor for formative feedback by the end of **October 2020**.
- You will need to investigate and read related work, and you have to submit an individually written short literature review of your findings in order to get formative feedback.
- You have to select, implement and apply appropriate machine learning algorithms to the selected problem, performing data pre-processing, if needed, and record the results from the experiments.
- You will receive regular feedback on your progress from the module leader or tutor in the labs when once you have submitted your progress report on each part.
- At the end, you have to write up your final paper, and submit it by 11/12/2020, 6.00pm.

Criterion	**Mark**
Technical quality	
1) Rigour and extent of the experiments.	5%
2) Correct application of the selected algorithms and suitability of the methods.	5%
3) Data preparation - technical quality.	5%
4) Extent of evidence of running the experiments provided in appendices.	5%
Evaluation	
5) Evaluation and discussion of the results. Why the results are important? How would the results be useful to other researchers or practitioners?	8%
6) Is this a "real" problem or a small "toy" problem? How does the paper advance the state of the art?	2%
7) **Social, ethical, legal and professional considerations** related to the problem in question.	2%

Clarity of the writing: 8) Is there sufficient information for the reader to reproduce the results? Is the language used in the paper good? 9) References and general presentation; Are results clearly presented, with appropriate visualisations?	 5% 3%
Originality: 11) Is there some original approach to the problem, original use of techniques? 12) Is there any (and how much) difference from previous contributions?	 5% 5%

Task 2: Evolutionary and Fuzzy Systems

Fuzzy Logic Optimized Controller for a Commercial Greenhouse

Design and Implement a Fuzzy Logic Controller (FLC) to be used to the control climate and/or irrigation for a commercial greenhouse. The environmental parameters to be controlled could be ambient temperature humidity, lighting, soil moisture content and PH. These could be controlled with actuators such as cooling fans, heaters, sprinklers water pumps and lamps.

https://news.lift.co/touring-aphrias-greenhouses-in-ontarios-banana-belt/

FLC Design

The FLC should be based on determining the input and output parameters of the system, depending on what control behaviour(s) the FLC will implement. ***Note*** *that depending on the control behaviours you wish to implement, you can select to use a subset of the input sensors so think about the behaviour(s) the FLC should control*.

Design choices should be made to consider the type and number of fuzzy sets for input parameters and or output parameters.

A set of suitable control rules should be defined which can be experimented with to achieve a good control performance of the chosen behaviour(s).

The FLC should therefore implement the following elements:

- Consideration of which Fuzzy Inference model to use: Mamdani or Sugeno (TSK) fuzzy models
- Mapping the crisp data input and output parameters into designed fuzzy sets.
- Map input fuzzy sets into output fuzzy sets *(for Mamdani model)* based on a set of designed rules that capture the desired control behaviour of the system.
- Employ appropriate inference operation (*rule implication*) that handles the way in which rules are activated and combined together (*composition and aggregated*).

- The outputs of the fuzzy inference engine will define a modified output fuzzy set (for Mamdani model) that specifies a possibility distribution of the control actions in relation to activated rules.

- Use an appropriate defuzzifier to convert the modified fuzzy outputs into nonfuzzy (crisp) control values that can then be used to set the output actuation parameters.

Working alone, complete the following tasks:

Part 1 – Design and Implementation of the FLC

(30 Marks)

Design and implement a demonstrable FLC, which can be a simulated system programmed in Matlab, FuzzyLite or Juzzy, see links below:

Matlab Fuzzy Logic Toolbox (http://uk.mathworks.com/videos/getting-started-with-fuzzy-logic-toolbox-part-1-68764.html,

http://www-rohan.sdsu.edu/doc/matlab/toolbox/fuzzy/fuzzyt10.html)

Fuzzylite (*http://www.fuzzylite.com*)

Juzzy (*http://juzzy.wagnerweb.net*)

Provide suitable evidence of your implementation in the form of diagrams and screenshots of the different components.

(18 marks)

Discuss and justify your design decisions for the choice fuzzy sets: membership functions, fuzzy rules, FLC inference mechanism selected and defuzzification method that was chosen. Back up your explanations with evidence in the form of appropriate diagrams and screenshots.

(6 marks)

Perform analysis of the output behaviour of the controller showing the rules activation, controller output and control surface plots demonstrating how the controller achieves the specified behaviours in relation to an operational scenario.

(6 marks)

Part 2 – Optimize the FLC developed for Part 1

(10 Marks)

Consider the Fuzzy Logic Controller (FLC) for Controlling the smart home you have designed for the above part. The purpose of this part is to optimize the fuzzy controller you have previously developed. A data set of n examples, (x_i,y_i), i = 1,2 …, n, is available to evaluate the performance of your controller.

Keeping the same structure of the FLC as you have used in Part 1, design a genetic algorithm to adjust the membership functions of the input and output variables of the FLC in order to optimize the performance of your FLC. Give details of the genetic algorithms you have used, i.e., problem encoding, genetic operators, fitness function. Some of you may have designed the FLC as a Mamdani model, while others may have used Sugeno models. Clarify what is the length of

the chromosomes used in your solution. In case you have used a Mamdani model to implement your FLC, describe how the genetic algorithm solution would change if you were to use a Sugeno model for your FLC. Conversely, if you have used a Sugeno model in Part 1, describe how the genetic algorithm solution would change if you were to use a Mamdani model for your FLC.

Part 3 – Compare different optimization techniques on CEC'2005 functions

(10 Marks)

Choose two functions from the CEC'2005 suite of benchmark functions available here:
http://www.cmap.polytechnique.fr/~nikolaus.hansen/Tech-Report-May-30-05.pdf
More details about the special session at CEC'2005 can be found here:
https://www.ntu.edu.sg/home/EPNSugan/ (click "Benchmarks" then "CEC'05 Special Session / Competition")
https://www.al-roomi.org/benchmarks/cec-database/cec-2005
Some of the links in these pages are broken, but **you will be able to download the Matlab code for the functions if you click "Resources Database (Different Formats) [Download]"** on the last web page indicated above.

This part is to compare the performance of at least 2 different optimization techniques on the two functions you have chosen, for both D=2 and D=10, where D is the number of dimensions. If you want to challenge yourself, you may try higher dimensional spaces, for example D=100, but this is optional. As optimization techniques to compare in this part, you may choose Genetic Algorithms, Particle Swarm Optimization, Simulated Annealing or other optimization methods available in the Global Optimization Toolbox or the Optimization Toolboox in Matlab, or developed as standalone programs by yourself.

To make the comparison meaningful you would have to run each optimization algorithm 15 times and report the average performance (including the standard deviation of the obtained results), as well as the best and the worst performance among the 15 runs. You may try to compare your results with results reported in the literature on the same functions.

In your report, you should include the description of the functions you have selected, the Matlab code for those functions, the results obtained and the parameters of the optimization algorithms used to obtain the reported results, any other Matlab scripts or code used in your simulations, convergence graphs, etc.

Parts and Mark distribution:

Part 1	30
Part 2	10
Part 3	10

Notes:

1. You are expected to use the Coventry University Harvard Referencing Style. For support and advice on this students can contact Centre for Academic Writing (CAW).

2. Please notify your registry course support team and module leader for disability support.
3. Any student requiring an extension or deferral should follow the university process as outlined here.
4. The University cannot take responsibility for any coursework lost or corrupted on disks, laptops or personal computer. Students should therefore regularly back-up any work and are advised to save it on the University system.
5. If there are technical or performance issues that prevent students submitting coursework through the online coursework submission system on the day of a coursework deadline, an appropriate extension to the coursework submission deadline will be agreed. This extension will normally be 24 hours or the next working day if the deadline falls on a Friday or over the weekend period. This will be communicated via your Module Leader.
6. You are encouraged to check the originality of your work by using the draft Turnitin links on your Moodle Web.
7. Collusion between students (where sections of your work are similar to the work submitted by other students in this or previous module cohorts) is taken extremely seriously and will be reported to the academic conduct panel. This applies to both courseworks and exam answers.
8. A marked difference between your writing style, knowledge and skill level demonstrated in class discussion, any test conditions and that demonstrated in a coursework assignment may result in you having to undertake a Viva Voce in order to prove the coursework assignment is entirely your own work.
9. If you make use of the services of a proof reader in your work you must keep your original version and make it available as a demonstration of your written efforts.
10. You must not submit work for assessment that you have already submitted (partially or in full), either for your current course or for another qualification of this university, unless this is specifically provided for in your assignment brief or specific course or module information. Where earlier work by you is citable, ie. it has already been published/submitted, you must reference it clearly. Identical pieces of work submitted concurrently will also be considered to be self-plagiarism.