BMCS2114 Machine Learning

Assignment

Aims	1. Enable students to analyze and employ appropriate Machine Learning techniques to solve problems. 2. Enable students to use relevant tools and technology, such as python programming, to visualize and evaluate the performance of different machine learning algorithms.
Learning Outcomes Assessed	1. Justify the ML methods and strategies for the data set based on the performance evaluation. (A3, PLO7) 2. Produce ML programs using programming language or other available technologies to solve a real problem. (P4, PLO3)
Outline of Problem	For your assignment, you should critically evaluate current technologies, and then propose a project for the selected ML topic, implement an ML solution to solve the problem in the proposed project by using Python.
Details	The Assignment consists of **TWO (2)** parts, please refer to **Project Details Part A** and **Project Details Part B.** Both parts are related. Group yourself into a team of **THREE OR FOUR (3-4)** members. A list of suggested websites and data sets are posted below. You are encouraged to use one of the suggested websites to obtain data sets, because they have been successfully used for machine learning in the past. Please note that the following are the basic requirements of the assignment. **Fulfilling the requirements may lead you to an Average or Good grade.** Achieving Excellent grade requires evident **extra efforts**, such as learning new skills, introducing new idea, complex ML algorithms, demonstrating the ability to process big data, and/or producing excellent reports with working prototype. Your group is expected to produce idea that is originated from the members, but not to take the work or an idea of someone else (including the Web) and pass it off as your own. **NO GROUP IS ALLOWED** to share the same idea, i.e., each group must propose a unique title or solution.
Machine Learning Project Checklist	This checklist can guide you through your Machine Learning projects. There are eight main steps: 1. Frame the problem and look at the big picture. 2. Get the data. 3. Explore the data to gain insights. 4. Prepare the data to better expose the underlying data patterns to Machine Learning algorithms. 5. Explore many different models and **short-list the best ones.** 6. **Fine-tune your models** and combine them into a great solution. 7. Present your solution. 8. Launch, monitor, and maintain your system.

Project Details Part A – Shortlist promising models	1. Train many quick and dirty models from different categories (e.g., linear, naive Bayes, SVM, Random Forests, neural net, etc.) using standard parameters.
	2. Measure and compare their performance.
	   • For each model, use N-fold cross-validation and compute the mean and standard deviation of the performance measure on the N folds.
	3. Analyze the most significant variables for each algorithm.
	4. Analyze the types of errors the models make.
	   • What data would a human have used to avoid these errors?
	5. Have a quick round of feature selection and engineering.
	6. Have one or two more quick iterations of the five previous steps.
	7. Short-list the top three to five most promising models, preferring models that make different types of errors.
	
	Note: try to automate these steps as much as possible.

Project Details Part B – Fine-tune the system	1. Fine-tune the hyperparameters using cross-validation.
	   • Treat your data transformation choices as hyperparameters, especially when you are not sure about them (e.g., should I replace missing values with zero or with the median value? Or just drop the rows?).
	   • Unless there are very few hyperparameter values to explore, prefer random search over grid search. If training is very long, you may prefer a Bayesian optimization approach (e.g., using Gaussian process priors)
	2. Try Ensemble methods. Combining your best models will often perform better than running them individually.
	3. Once you are confident about your final model, measure its performance on the test set to estimate the generalization error.
	
	Note: try to automate these steps as much as possible. You may also make use of deep learning model to solve the problem.

Datasets	1) **Kaggle dataset (**https://www.kaggle.com**)**
	Example: UFC-Fight historical data from 1993 to 2019 (https://www.kaggle.com/rajeevw/ufcdata)
	   • Project Idea 1: Outlier detection on the players; find out who are the outstanding players.
	   • Project Idea 2: Predict the game outcome.
	
	2) **Precipitation data**
	This dataset has includes 45 years of daily precipitation data from the Northwest of the US (http://research.jisao.washington.edu/data_sets/widmann/)
	   • Project Idea 1: Weather prediction: Learn a probabilistic model to predict rain levels.
	   • Project Idea 2: Sensor selection: Where should you place sensor to best predict rain.
	
	3) **IMDB Movie database and user rating data (**https://www.imdb.com/interfaces/**)**
	   • Project Idea 1: User rating prediction based on past ratings
	   • Project Idea 2: discover clusters of similar movies or users
	
	4) **Computer Vision Datasets (**https://www.cs.toronto.edu/~kriz/cifar.html**)**
	   • Project Idea 1: Semantic segmentation
	   • Project Idea 2: Fine-grain object classification

Submission Deadlines	**Final Report: 5th Apr 2021 (Week 12)** **Presentation: Starts on Week 12 Practical class** Please submit to your tutor. Late submission will be penalized.
Contribution	This assignment contributes **100%** to the **coursework** component.
Academic Integrity and Plagiarism	There must be originality in your work, i.e. do not copy or refer to other group(s). You may only work with your team member(s) to produce the solution of this assignment. You must not share with nor refer to any part of the assignment (including the code) of anyone else except your team member(s) and your tutor. Before submitting your assignment, please make sure that you have complied with **TARUC plagiarism Policy**. Any cheating, attempt to cheat, plagiarism, collusion and any other attempts to gain an unfair advantage in assessment will cause the students concerned to be penalized. **IMPORTANT**: Students found to be dishonest are liable to disciplinary action.
Late Submission	Late submission without valid reason will NOT be tolerated. For late submission, there will be a reduction of total marks: • Late 1 to 3 days after deadline of submission: minus 10 marks • Late 4 to 7 days after deadline of submission: minus 20 marks • Late more than 7 days after deadline of submission: 0 marks In certain circumstances, a student may be allowed to submit the assignment late with valid reason. S/he must contact the tutor at least one week before the assignment is due. The tutor will evaluate whether the circumstance warrants submitting the assignment late, but no guarantee that the students will not be penalized. Failed to submit the reports & code will lead to failure of the coursework.

Project Details: Final Report Writing

Introduction	Your task for this part of assignment is to **identify a problem** for the selected ML topic, evaluate different **ML algorithms** using Python or other relevant softwares that help in solving the problem in the proposed project. Finally compose the work in an article.
What to hand in?	Students are required to come out with the final report for the selected ML topic. Project assessment (demo, question and answer) will be carried out during the practical class in week 12 or 13. The final report should contain the following: • Title of the proposed project. • Brief description of your propose ML solution to solve the problem. (Problem Statement) • Describes how your solution works on a level that does not lose interesting details. Make some sensible breakdown into headings (Methodology). • Includes and explain the evaluation outcome in the article (Result). • Use clear, well written English language. • Task allocations for each of the respective group member and a project plan. ** Each student is expected to produce **individual** work regarding the specified items. • The report shall be compiled and submitted in group.
Format for Deliverable	For your report, you are required to adhere to the followings: 1. Length : **Between 10 to 20 pages** 2. Paper size : A4 3. Font : Times New Roman / Arial, 11 points 4. Cover page: Please use Appendix 1 5. Last page: Please attach Appendix 2 **Note:** Submit all your materials (final report and source code) to Github Classroom assignment submission page.