National University of Singapore
School of Computing
CS3244 Machine Learning

Term Project: ML Singapore!

Issue: February 17, 2022 Due: Apr 15, 2022

Important Instructions

1. *Your project report must be TYPE-WRITTEN using the latest AAAI Press Word template or LaTeX macro located in the author kit*[1]. *It should NOT be more than* 6 *pages (inclusive of diagrams, references, and team details described in bullet point* 5*).*

2. *Submit your project report in PDF format to the Files folder named "Project Report Submission" in IVLE by* **Apr 15, 2022, 11:59pm**. *Late submissions will NOT be accepted.*

3. *You should submit ONE project report per team.*

4. *Each team should have 5 members.*

5. *Indicate clearly your team number and all names and student numbers of the members of the team on the first page of the project report.*

6. *In a separate section, describe clearly the role(s) of each team member in this project.*

7. *Note that we may call any one member of a team to explain the submitted report during the marking process. You are advised NOT to free-ride.*

All members of the team will be given the same grade/mark[2]. The grading criteria is as follows:

[1] http://www.aaai.org/Publications/Templates/AuthorKit22.zip

[2] Free riders are given zero.

8%	**Are you hungry enough for success?** Suppose that you are envisioning a start-up based on a novel *machine learning* (ML) app that you have been aiming to design and develop for some time and you believe will be a potential game changer. This can now be achieved thanks to the availability of your project team-mates (at no cost :). Describe CLEARLY and concisely a REAL-WORLD APPLICATION scenario/use case in the urban city of Singapore motivating the use of your app and detailing why there is a need for using the ML models/algorithms of your choice. In particular, you may wish to consider application domains like healthcare (e.g., medical conditions like diabetes, high blood pressure, high cholesterol, and open issues like how to prevent misdiagnosis, personalized multi-drug combination therapy), cybersecurity (e.g., low-volume DDoS attacks, anomaly detection, learning "fingerprints" of attacks), fake media, epidemic outbreak, climate change, food self-reliance/resilience[3], trusted data sharing and data valuation[4], among others that you think will *improve the quality of life of the people living in Singapore*. Doing so will allow you to demonstrate the practical significance of your project work. For some examples of real-world applications, refer to the ML and AI news maintained by the Association for the Advancement of Artificial Intelligence (AAAI)[5], Deep AI[6], and my research group's webpage[7]. You may also wish to consider building an app that is capable of automatically mapping the *local R&D landscape in AI and ML*, with the goal of being able to answer questions such as "what expertise do AI/ML researchers in Singapore have?", "what applications are local companies using AI/ML technologies for?", "where can I find an AI/ML researcher working on X?", among others. This consists of several steps or subprojects, including (a) finding information about relevant entities (e.g., researchers, research centers, companies) on the Internet, (b) extracting key information about them (e.g., from research papers or press releases) and structuring it, and finally (c) establishing links and relationships between these entities (such information is sometimes represented in a knowledge graph). (To be continued on the next page)

[3] https://www.sfa.gov.sg/food-farming/sgfoodstory

[4] https://www.imda.gov.sg/news-and-events/Media-Room/Media-Releases/2019/Enabling-Data-Driven-Innovation-Through-Trusted-Data-Sharing-In-A-Digital-Economy, https://www.imda.gov.sg/-/media/Imda/Files/Programme/Data-Collaborative-Programme/Guide-to-Data-Valuation-for-Data-Sharing.pdf, https://www.comp.nus.edu.sg/~lowkh/research.html

[5] https://aitopics.org/class/AI-Alerts

[6] https://deepai.org/

[7] https://www.comp.nus.edu.sg/~lowkh/research.html

(Continuing from the previous page)

In particular, your description should address the following concerns, among others:

- How does your proposed application exploit the desirable properties of the ML models/algorithms/techniques of your choice?
- In the context of your proposed application, what qualitative advantages (and/or limitations) do your chosen ML models/algorithms/techniques provide, as compared to other available ML models/algorithms/techniques that can be used in your proposed application?
- Does your proposed application have any important requirements (e.g., real time, streaming data, big data, small data, high or variable input dimensions) that can be satisfied by your chosen ML models/algorithms/techniques or their more effective or scalable variants (e.g., Bayesian, parallel/distributed/decentralized, online, multi-output/multi-task variants)?
- How can the human users/experts in your proposed application understand, visualize, and interpret the outputs of your chosen ML models/algorithms/techniques and interact with them (if necessary)? More importantly, how do the outputs of your chosen ML models/algorithms/techniques help these human users/experts to plan and make decisions in your proposed application?

Different from the above, there is also an opportunity to choose to work on a research-oriented problem instead of an application-oriented one. It is highly exploratory and hence involves a non-trivial amount of risk and effort; the reward/achievement can be potentially significant. As you would have observed from the "Neural Networks" lecture slides, a perceptron unit or a multi-layer network of perceptron units cannot be trained using gradient descent due to its discontinuous Heaviside step activation function which makes it undifferentiable. Can it scale to a sufficiently large network and still be trained to reach a competitive predictive performance like that of the gradient-based counterparts (e.g., multi-layer neural network using sigmoid or ReLU activation function)? What are its pros and cons relative to the gradient-based counterparts? Are there specific application domains or target functions where it can be trained to give better predictive performance? One potential suggestion is to consider the use of gradient/derivative-free/black-box optimization methods like Bayesian optimization.

6%	**Are you resourceful enough to succeed?** Bryan is unfortunately not able to teach you everything you need to know about ML in CS3244; we probably need an undergrad/postgrad degree program to do so. Can you identify, understand, and exploit the state-of-the-art ML tools/techniques (often beyond what will be taught in our class), resources, and data that you need to develop your app? Fortunately, you are not alone in this – you are blessed with fellow team-mates to learn, research, discuss, and understand the necessary technical background and tools together and make them work for your proposed application. Many heads are better than one! Specifically, your description should address the following concerns, among others: • Investigate an advanced ML topic(s) or the use of state-of-the-art ML model(s)/algorithm(s)/technique(s) for developing your proposed app. Describe CLEARLY and concisely the technical details of your chosen ML model(s)/algorithm(s)/technique(s). Doing so will allow you to demonstrate the theoretical rigor and understanding of your project work. I have listed below some recommended ML topics and state-of-the-art ML models/algorithms/techniques with links to online resources, workshops, and video-recorded tutorials at the flagship ML conferences discussing the latest advances: – Convolutional neural networks (CNNs), residual neural network (ResNet), and recurrent neural nets (RNNs), variational autoencoders (VAEs), generative adversarial networks (GANs)[8] – Automated ML[9]: Hyperparameter optimization, Bayesian optimization (BO)[10], neural architecture search[11] – Data-efficient ML (Learning with less data): Active learning, BO[8], meta-learning[12], transfer learning – Privacy-preserving ML[13], federated learning[14] – Incentives in collaborative ML/federated learning[15] – Kernel methods like Gaussian processes (GPs) and deep GPs[16] – Bayesian deep learning[17] – Interpretable ML[18] (e.g., local interpretable model-agnostic explanations (LIME)[19]) – Algorithmic bias, fairness, transparency, and ethics[20] – Adversarial ML[21] – ML on mobile and IoT devices[22] (To be continued on the next page)

[8] https://keras.io, https://github.com/yunjey/pytorch-tutorial,

https://github.com/Hvass-Labs/TensorFlow-Tutorials, https://www.tensorflow.org/tutorials, https://r2rt.com/recurrent-neural-networks-in-tensorflow-i.html, https://r2rt.com/recurrent-neural-networks-in-tensorflow-ii.html, https://r2rt.com/recurrent-neural-networks-in-tensorflow-iii-variable-length-sequences.html, https://www.facebook.com/nipsfoundation/videos/1552060484885185/

[9] https://sites.google.com/site/automl2017icml/

[10] http://bayesopt.com, http://www.comp.nus.edu.sg/~lowkh/research.html, https://sheffieldml.github.io/GPyOpt/

[11] https://arxiv.org/abs/1808.05377

[12] http://metalearning.ml/2019/

[13] https://vimeo.com/248492174, https://sites.google.com/view/psml

[14] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

[15] https://gradanovic.github.io/incentives_in_ML_icml2020_ws/index.html, https://www.comp.nus.edu.sg/~lowkh/pubs/icml2020r.pdf

[16] https://www.facebook.com/nipsfoundation/videos/1552223308202236/, https://github.com/SheffieldML/, https://www.gpflow.org/ https://sheffieldml.github.io/GPy/, https://github.com/arikcj/pgpr, https://github.com/qminh93/RVGP, https://github.com/sdfond/onlineGP, https://github.com/markvdw/GParML, https://github.com/qminh93/dSGP_ICML16

[17] https://www.facebook.com/nipsfoundation/videos/1555493854541848/, http://bayesiandeeplearning.org

[18] http://people.csail.mit.edu/beenkim/icml_tutorial.html, https://sites.google.com/view/whi2017/home

[19] https://github.com/marcotcr/lime, https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

[20] https://www.facebook.com/nipsfoundation/videos/1553500344741199/, https://vimeo.com/248490141

[21] https://en.wikipedia.org/wiki/Adversarial_machine_learning, https://medium.com/cltc-bulletin/adversarial-machine-learning

[22] https://sites.google.com/view/nips-2017-on-device-ml/, https://sites.google.com/site/tinyml2017/, https://developer.apple.com/machine-learning/, https://caffe2.ai/, https://www.tensorflow.org/mobile/, https://opensource.googleblog.com/2017/06/mobilenets-open-source-models-for.html

(Continuing from the previous page)

- What technical insights have you gathered about the behavior of your chosen ML models/algorithms/techniques that you have found to be particularly interesting?
- Do your chosen ML models/algorithms/techniques fit the requirements of your proposed application exactly? Have they met your expectation when used for your proposed application? What have worked well and what have not? What is the most rewarding and/or painful experience with using your chosen ML models/algorithms/techniques?
- Can you propose some novel modifications/revisions/improvements (if any) of your chosen ML models/algorithms/techniques to improve their performance (e.g., predictive accuracy, time efficiency) or to make them fit your proposed application better?
- How have you effectively organized and exploited your manpower resources (i.e., the resourcefulness and strengths of each team-mate)? Can each of you reflect what you have learned that you have previously not known about ML through this project?

6%	**Have you run extensive tests and experiments to demonstrate feasibility and success?** Describe clearly and CONCISELY how you empirically demonstrate the usefulness of your chosen ML models/algorithms/techniques in your proposed real-world application and empirically evaluate their performance (e.g., trade-off between predictive accuracy and time efficiency). In particular, you should consider the following guidelines: • Describe your experimental setup clearly and carefully. • What (possibly surrogate) real-world dataset(s)[23] will you be using for training and testing? • Are there more than one type of ML models/algorithms/techniques that can be used for your proposed application? If so, have you empirically compared and analyzed their performance? What quantitative/empirical advantages (and/or limitations) do they provide? What observations/conclusions can you draw from your experiments? • List the online resources, tools, and code implementations of your chosen ML models/algorithms/techniques that you have used for your proposed application.
4%	**Will Singapore (or maybe just Bryan) invest in you?** Basically, when I read your report, I should go like "Wow! Your proposed application and research work are NOVEL, INTERESTING, and really cool and would significantly improve the quality of life of many people living in Singapore!" In a technically sound and pleasantly surprising way, of course. I have the budget to give *at least* 9 teams FULL MARKS in this category.
1%	**Following instructions**: The instructions are provided on page 1.

[23] Here are some links to real-world datasets:

- https://data.gov.sg, http://www.epa.gov/airdata/
- https://archive.ics.uci.edu/ml/datasets.html
- http://www.metoffice.gov.uk/hadobs/emslp/
- http://www.metoffice.gov.uk/hadobs/hadsst2/
- http://www.metoffice.gov.uk/hadobs/mohsst/
- http://ai.google/tools/datasets/, http://toolbox.google.com/datasetsearch
- http://www.esrl.noaa.gov/psd/data/gridded/, http://coastwatch.noaa.gov
- http://select.cs.cmu.edu/data/index.html
- https://sites.google.com/site/goovaertspierre/pierregoovaertswebsite/download, http://www2.hawaii.edu/~matt/680/brooms_barn.txt
- https://www.kaggle.com/brandao/diabetes
- https://www.kaggle.com/uciml/pima-indians-diabetes-database
- https://physionet.org/challenge/2019/