Group Project (Assignment 2)

CAB420, Machine Learning, Semester 1, 2021

This document sets out the instructions for the CAB420 group assignment (also referred to as Assignment 2). The assignment is worth **30%** of the overall subject grade, and is to be completed in groups of 3 to 4. Each student's contribution should be indicated by a clear explanation of the contributions, and the percentage of the whole work. Please note that your mark may be moderated depending on your individual percentage contribution specified in your report.

Instructions

Students are to select a machine learning problem/topic to investigate in groups of $3-4$. You may select your own project idea, or choose from a small list of provided topics. The problem/topic should ideally:

- Have one or more established datasets that can be used.
- Be capable of being addressed in multiple, diverse ways (for example, classification can be done using various formulations of DCNN, an SVM, a random forest, or several other techniques).

Groups are to implement multiple different methods to address their chosen problem/topic, and prepare a report that details these approaches and compares their performance on the chosen data. You are encouraged to discuss possible project ideas with teaching staff during practical sessions or after the consultation/support session.

Students are to self allocate into groups, and are encouraged to use their practical class and/or the slack channel to find a group. Upon forming a group students should:

- Send an email to `cab420query@qut.edu.au`, cc'ing all group members, advising that the group has been formed. You will be advised of your group number by return email.

Submission Details

The project comprises three separate pieces: a project proposal, a final report, and a short presentation. Note that the first (project proposal) is optional and not assessed, though it is strongly recommended you submit this to receive feedback on your proposed project. Details of these three components are outlined below. Note that only one group members needs to be submit the project proposal and final report.

Project Proposal, 1 - 2 pages, Due Date: 2nd May 2021, 11:59pm

A brief project proposal should be submitted at the end of Week 8. This should be one to two pages in length and should include the following information:

- Team members
- Project title
- Project motivation and objective
- Dataset(s) to be used and brief details on the proposed evaluation protocol

Note that no mark will be recorded for the project proposal, but you will receive feedback on the suitability of your proposed project and approach.

The Report, Due Date: 6th June, 2021, 11:59pm

The report should be structured as follows:

- Title page, containing project title, team number, and names of team members
- Main body of the report, which should be **9–12** pages long and include the following:
  - Introduction/Motivation: clearly motivate your project, and describe the research question, and how it relates to previous works that have been done in this area.
  - Related Work: briefly describe a small number of relevant existing approaches, and the objective of your work.
  - Data: clearly describe the data set, any pre-processing, and the data split into training/validation/test.
  - Methodology: Clearly explain the three algorithms that you used with citations to the literature. Please note that your project ideally should extend the existing approaches. You don't need to propose a novel algorithm, but you might be looking into approaches that have not previously investigated on your dataset. Note also that your three approaches should be different. For example, rather than simply using three feed-forward neural networks for a classification task, you could perhaps use (depending on the task) one feed-forward network, one GAN, and a non-deep learning method such as a random forest.

  - Evaluation and Discussion: Present the results of all your approaches clearly, and compare them with existing approaches and each other. Discuss why your methods are working better/worse than the existing approaches and each other.
  - Conclusions and Future Works: Clearly explain if the experiments match the objectives, the advantages/shortcomings of the proposed approached, and if any changes are required/ plans you have for the future investigations.

- An appendix that details the contribution of each group member towards the project should be included. This should list what each group member contributed, as well as an approximate overall percentage that each student contributed. This statement should be signed by each group member.

You may optionally wish to include further appendices to including items such as code or additional (non-critical) results. Though please be aware that all critical content should be included within the main body of the report, and content within the appendices will be considered supplementary.

Presentation, Due Date: 7th – 9th June, 2021 (Week 14)

All projects will have a 5 min presentation during scheduled during week 14, before exam block begins. The presentation will be made to the teaching team. At least one project member should be present for the presentation. The presentation should:

- Describe the dataset used;
- Briefly present the algorithms investigated;
- Provide a brief discussion of the results achieved.

Depending on your project, a small demonstration of the method and/or displaying sample outputs may also be appropriate.

Following your presentation, the teaching team will ask questions about your approach or results which you will be expected to be able to answer.

A schedule for the presentations will be posted towards the end of semester, with opportunities to reschedule presentations as needed, or present remotely if no other option is available.

Submission

Materials should be submitted via the provided TurnItIn links. Only one group member per group should make a submission. If multiple group members submit, we kindly request that you contact `cab420query@qut.edu.au` to advise the teaching and marking team that this has occurred.

Topic Suggestions

For this assignment, you can propose your own project ideas. If you wish to do this, please contact `cab420query@qut.edu.au`, or discuss the idea with the teaching team during practicals, or after lectures to see whether the proposed project is suitable as a project for CAB420.

You may also wish to look at Kaggle competitions. Kaggle is an online community of data scientists and machine learners, owned by Google. One of the more interesting aspects of Kaggle is that it provides datasets and organises many competitions for machine learning tasks. It also manages a leaderboard where the participants can publish their results. You are free to browse the competitions there to see if there are any challenges you are interested in, or if there are any datasets that you think may be of use for your project.

A small selection of project ideas are listed below that you may also choose from.

Enron E-mail Classification

The Enron E-mail data set contains about 500,000 e-mails from about 150 users. The data set is available here: `http://www.cs.cmu.edu/~enron/`. Can you classify the text of an e-mail message to decide who sent it?

DNN ⇒ Pre trained w/ vgg or sim

Clustering

Non-DNN

Object Recognition

The Caltech 256 dataset contains images of 256 object categories taken at varying orientations, varying lighting conditions, and with different backgrounds, and is available at `http://www.vision.caltech.edu/Image_Datasets/Caltech256/`. You can try to create an object recognition system which can identify which object category is the best match for a given test image, or possibly apply clustering to learn object categories without supervision.

WebKB Prediction

This dataset, `http://www-2.cs.cmu.edu/~webkb/`, contains webpages from 4 universities, labeled with whether they are a professor, student, project, or other pages. Learn classifiers to predict the type of webpage from the text.

Speaker Recognition

Speaker recognition is the task of recognising someone by the way that they speak. This is a classical problem in biometrics and machine learning. The Common Voice (`https://voice.mozilla.org/en/datasets`) dataset contains a large number of speakers and associated meta-data, and spans multiple languages. You could simply try to recognise a speaker, or investigate the impact of training and evaluating a model on different language, or explore how meta-data could be used to improve performance.

Crowd Counting

Crowd counting is the task of counting the number of people in a scene. The output of crowd counting can be simply a single number representing the total number of people in a scene, or a density map that indicates how people are distributed in a scene (or both). A large number of datasets have been released for crowd counting (see `https://github.com/gjy3035/Awesome-Crowd-Counting#datasets`). Using one or more of these you could investigate crowd counting methods, or explore how different types of methods generalise to different conditions.

Semantic Segmentation of Aerial Data

Semantic segmentation is the task of labelling each pixel of an image with a label indicating what is at the pixel. This is commonly used within scene understanding pipelines to identify objects and regions of interest in a scene. DroneDeploy have released a segmentation dataset and benchmark suite (see `https://github.com/dronedeploy/dd-ml-segmentation-benchmark`) for semantic segmentation from drone data. Along side colour imagery, elevation data has also been captured. How could you combine elevation and RGB data to improve segmentation performance?

Other Sources of Project Ideas

There are many other places you may turn to for project ideas. A few useful links include:

- `http://cs229.stanford.edu/projects2015.html`
- `http://cs229.stanford.edu/projects2016.html`
- `https://github.com/NirantK/awesome-project-ideas`