Cranfield UNIVERSITY

Artificial Intelligence Assignment

2023-2024

Dr Jun Li <jun.li@cranfield.ac.uk>

Note: You are free to choose one of the three previous online competitions below for your coursework. You are advised to download and backup the dataset and project descriptions. Be aware of the data size. While it is great to use the whole amount of data for model training, it is fine to sample a subset with a justified sampling approach.

I anticipate that all three options require significant computational power. Though meeting the competition requirements is desirable, a proof of concept would suffice the coursework requirement (an innovative working prototype). You may explore the possibility of the university HPC or a free Cloud environment for your model implementation. The work can be potentially continued as your thesis project, but you have to discuss this with your supervisor.

Cranfield UNIVERSITY

Artificial Intelligence Assignment: Option-1
Dementia screening classification

This was a real-time competition: https://biomag2020.org/awards/data-analysis-competitions/. Please send an email to me for the dataset. Here is some essential information:

Description: Dementia is a chronic and progressive syndrome caused by brain diseases with a few effective pharmacological treatments. It is important to diagnose it at the early stage of the syndrome, to provide effective interventions and slow the progression of the disease. Number of published studies have showed that magnetoencephalography (MEG) is sensitive to the subtle changes of brain activity related to dementia. MEG is a patient-friendly clinical examination tool, because it is totally non-invasive and the preparation/scanning time is short.

In this competition, your goal is to classify each of the MEG data sets ("test data", N = 42) into one of three classes: healthy volunteer, mild cognitive impartment (MCI), and dementia. For coding your classification algorithms, another set of MEG data ("training data") recorded from healthy volunteers (N = 100), patients with MCI (N = 15) and dementia (N = 29) are provided. All MEG data were recorded with the same paradigm (5-min resting-state recording with eyes-closed), but at two different sites (A or B, both sites have identical MEG systems). The sampling rate differed between sites: 1000Hz in site A, and 2000Hz in site B. All MEG data were recorded using 160-channel gradiometer system (Yokogawa at the site A and RICOH at the site B) and provided in SPM-12 format.

Beside the MEG data, age and gender information is available for all samples. If available, MMSE (Mini-Mental State Exam) score of the data-provider (healthy volunteer or patient) are also provided for the "training-data". All participants and patients gave written informed consent to re-use their data for advances in medicine. Submitted answers will be evaluated from the viewpoint of clinical medicine, and submitter(s) who provided the most reasonable classifications will be selected as the winner(s) of the competition.

Limitation: You are not allowed to use/distribute the data for any purposes irrelevant to the competition, without organiser's permission.

Cranfield UNIVERSITY

Artificial Intelligence Assignment: Option-2

VinBigData Chest X-ray Abnormalities Detection

-- Automatically localize and classify thoracic abnormalities from chest radiographs

This was a real-time competition on Kaggle. You may go to here (https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection/overview) for more details such as the datasets and their descriptions. You have to sign-up for these information. Here is some essential information,

- You'll automatically localize and classify 14 types of thoracic abnormalities from chest radiographs (X-ray image datasets in dicom format).
- You'll work with a dataset consisting of 18,000 scans that have been annotated by experienced radiologists. You can train, validate and test your model with 15,000 independently-labeled images.
- Details on the datasets can be found in paper "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations" and on Kaggle link above.
- As the dataset is large (**about 200G**), you are free to use all or a subset for result demonstration. But the discussion of your model performance as part of the model evaluation has to be included in your report.
- You may consult your lecturer for a disk-to-disk copy if the downloading is time consuming for you.
- You may find lots of information such as the dicom image manipulation on the "Notebooks" tab of the Kaggle link above.

Cranfield UNIVERSITY

Artificial Intelligence Assignment: Option-3

Indoor Location & Navigation: Identify the position of a smartphone in a shopping mall

https://www.kaggle.com/c/indoor-location-navigation/overview

This was a real-time Research Prediction Competition on Kaggle. You may find more detail via the link above and more help in the references below. Here is a brief.

Your smartphone can detect and use your location to provide contextual information such as driving directions or finding a store based on GPS, which requires outdoor exposure for the best accuracy. Yet, there are many times when you're inside large structures, such as a shopping mall or event centre. Current positioning indoor solutions based on public sensors have poor accuracy, particularly in multi-level buildings, or generalize poorly to small datasets.

In this competition, your task is to **predict the indoor position of smartphones (to see the evaluation of prediction in next page)** based on real-time sensor data. You'll locate devices using "active" localization data. You'll work with a dataset of nearly 30,000 traces from over 200 buildings.

The dataset for this competition consists of dense indoor signatures of WiFi, geomagnetic field, iBeacons etc., as well as ground truth (waypoint) (locations). The data found in path trace files (*.txt) corresponds to an indoor path between position p_1 and p_2 walked by a site-surveyor. The datasets are as shown:

- train - training path files, organized by site and floor; each path files contains the data of a single path on a single floor
- test - test path files, organized by site and floor; each path files contains the data of a single path on a single floor, *but without the waypoint (x, y) data*; the task of this competition is, for a given site-path file, predict the floor and waypoint locations at the timestamps given in the sample_submission.csv file
- metadata - floor metadata folder, organized by site and floor, which includes the following for each floor:
    - floor_image.png
    - floor_info.json
    - geojson_map.json

Prediction evaluation

Cranfield UNIVERSITY

Submissions are evaluated on the `mean position error` as defined as:

$$\text{mean position error} = \frac{1}{N}\sum_{i=1}^{N}\left(\sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2} + p \cdot |\hat{f}_i - f_i|\right)$$

where:

- N is the number of rows in the test set
- $\hat{x}_i, \hat{y}_i$ are the predicted locations for a given test row
- x_i, y_i are the ground truth locations for a given test row
- p is the floor penalty, set at 15
- $\hat{f}_i, f_i$ are the predicted and ground truth integer floor level for a given test row

IMPORTANT: The integer `floor` used in the submission must be mapped from the char/int floors used in the dataset. The mapping is as follows:

- F1, 1F $\rightarrow$ 0
- F2, 2F $\rightarrow$ 1
- etc.
- B1, 1B $\rightarrow$ -1
- B2, 2B $\rightarrow$ -2

To know more detail, please refer to here:

1. https://www.kaggle.com/c/indoor-location-navigation
2. https://www.kaggle.com/c/indoor-location-navigation/overview/evaluation
3. https://github.com/location-competition/indoor-location-competition-20
4. https://www.youtube.com/watch?v=xt3OzMC-XMU
5. https://developer.android.com/guide/topics/sensors
6. https://developer.android.com/reference/android/net/wifi/ScanResult.html
7. https://developer.android.com/reference/android/bluetooth/le/ScanRecord
8. https://location20.xyz10.com/

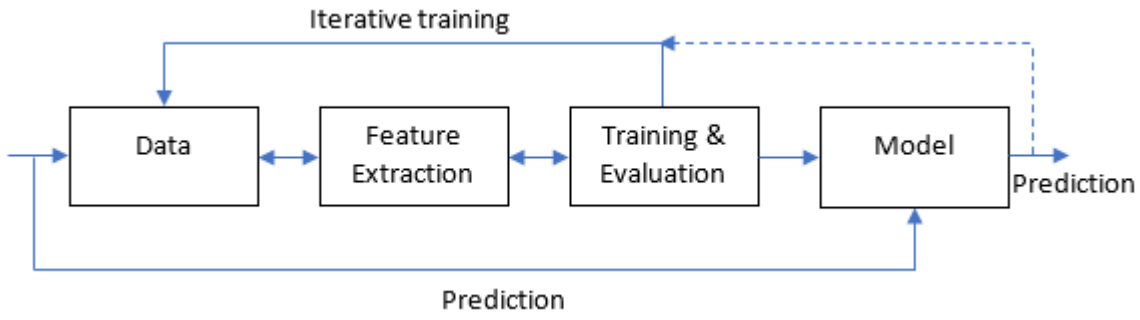
Note: The challenges lie in seeking the appropriate algorithms developed with right feature engineering for highly sparse datasets, and large computation with a data size about more than 50 GB, which may take you few days to fully download. For your project, you may use a subset only for a proof of idea with a implementable prototype model, which can be easily scaled to a full running model, given a powerful server or cluster e.g., HPC without significantly compromising the prediction precision of the original model.

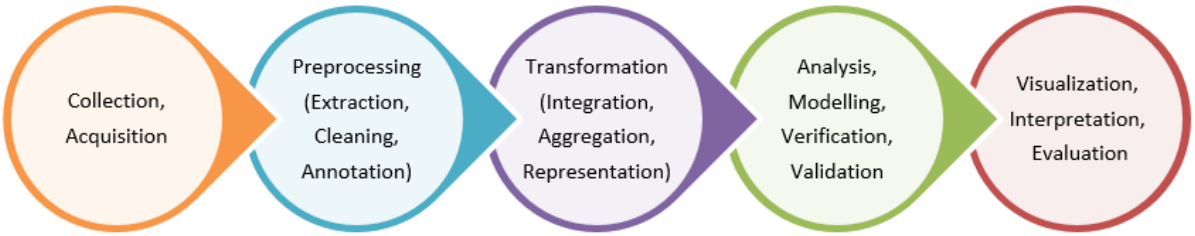
In the coursework, only the training dataset is used. You may split the training dataset into training, validation and testing datasets for model training and performance testing.

Cranfield UNIVERSITY

Machine Learning analysis and report

You should follow a general Machine Learning analysis procedure such as,





- Deep Learning though may be a good candidate for all options (you may still have to choose a specific type of deep neural network to fit the purpose), you are free to choose your own Machine Learning method(s) for the specific option.
- Employ a software package of your choice e.g. Python Notebooks, Google Colab or MATLAB.
- Evaluate the results with the accuracy, in terms of error (you should define how the error is computed) computational time (with respect to the computer specification hardware you are performing the computations on).
- Write a report to summarise your findings. The report should be no more than 3000 words, should include any number of figures and tables as long as they are referenced within the main text. The report should have the following sections
    - Introduction – literature review, fundamentals of AI&ML, scope and objectives.
    - Methodology and development – algorithm and structure description and mathematical representation, include diagrams and figures as well as model and software development, testing and validation.
    - Results and Discussion – write your findings and present them in the form of figures and tables. Use scientific rigour to justify your choices findings and argumentation, use relevant literature to support your discussion.
    - **In addition**, at the end of the discussion you need a paragraph to discuss the potential ethical issues and challenges related to Artificial Intelligence using scenarios such as the scenarios in this assignment.

- Conclusions – summarise your findings and state possible future directions.
- Again, please read the "Note" of the document. Let me know if you have any questions. If you are interested in a Kaggle competition, this may be a good start: https://www.analyticsvidhya.com/blog/2015/06/start-journey-kaggle/

Submission

- You must submit your report through Canvas.
- Full program source code (e.g. Python Notebooks) for your program(s) and model outputs must be submitted through Canvas separately in a .zip format. A Readme.txt or Readme.docx to illustrate your code and how to run your model should also be included.
- You may find the separate folders on Canvas to submit report and codes, where the marking proportion is an estimated one.

Submission Deadline (No late submissions):

- Full-time: XXX, 2024 (UK time)
- Part-time: XXX, 2024 (UK time)

Note: For the final deadline, please refer to the Assignments on Canvas.

Plagiarism

You must not plagiarise your work! You may use program source code from the provided course examples or other open libraries, but this usage must be acknowledged in the comments of your submitted file. Automated software tools will be used to detect cases of source code plagiarism in this practical exercise (taking into account the provided common source code examples from the course and the assignment itself).

You should have been made aware of the Cranfield University policy on plagiarism. Anyone unclear on this must consult the course lecturer prior to submission of this practical. University Plagiarism Guidance:

https://intranet.cranfield.ac.uk/ResearchLearnTeach/Documents/Plagiarism.pdf

Cranfield UNIVERSITY

Marking Criteria

Introduction, Literature Review, Conclusion	20
Methodology, Design	20
Development (coding)	30
Results and Discussion	20
Report structure and Writing	10

Marking criteria breakdown with Option-1 as an example

1. Introduction, Literature Review, Conclusion (20)
   - Introduction (e.g. the background and issues, objectives). 5 marks
   - Critical literature review (in width and depth). 10 marks
   - Conclusion (method used, findings, weakness) and future work. 5 marks
2. Methodology, Design (20)
   - Modelling (e.g. exploratory analysis, pre-processing such as Word2Vect, model design, training and testing), feasibility and running environment consideration (e.g. HPC, Cloud or Computer specification). 10 marks
   - Extra and innovative techs (e.g. sentence length consideration: bucketing and padding, rare words handling, new architecture with justification, attention or tracking context techs), and data if involves further data collection. 10 marks
3. Development (coding) (30)
   - Essential functions (consistent with the analysis in the report) with training and testing, data structures, and demonstration of successful runs. 10 marks
   - Innovative and independent functions (reflecting the architecture in design and performance in testing). 10 marks
   - Coding skills such as modulization, and documentation such as code comments and Readme.txt. 10 marks
4. Results and Discussion (20)
   - Model performance (incl. translation quality and improvement, rare words and sentence length handling, computational time and improvement through architecture and computing environment improvement). 10 marks
   - Results presentation (through training and testing), and model and results evaluation and discussion (e.g. the main finding, what and why significance of the results, implications of the findings, strength and limitation of the model and methods applied, potential improvement). 10 marks
5. Report structure and Writing
   - Visualization (such as data, model and results visualization using figures and tables). 5 marks
   - Structure (following a technical report structure and reflecting the modelling process and software development process) and English. 5 marks