INT305 Machine Learning 2020
Coursework 3 Task Sheet

Overview

One of the learning outcomes of INT305 Machine Learning is to prepare you to apply machine learning algorithms for specific problems. Equipped with the machine learning techniques, we want you to be able to solve practical tasks on real-world dataset. The Coursework 3 is designed to start you in this direction.

Coursework 3 is 70% of your final grade, and it is an individual research project.

Conduct an original practical research by applying machine learning algorithms and techniques to a public dataset. Pick an application that interests you, and explore how best to apply machine learning algorithms and techniques to solve it. It could be something that you are passionate about; or if you are already working in an industry, a work-related project that machine learning might apply to.

There are suggested dataset sources in the end of this task sheet. You are allowed to use Python data processing, data visualization, and machine learning libraries in your research project. However, please be honorable and not use any part of other people's research result as part of your research without proper citation.

Timeline (Beijing Time/CST/GMT+8) and Deliverables

Week 4 : Lecture 2	Coursework 3 Task Sheet released
Week 9 : Friday, 13 November 2020, 23:59	**Project Proposal** due
Week 13 : Friday, 11 December 2020, 23:59	**Project Interim Report** due
Exam Week : Monday, 11 January 2021, 00:00	**Project Video** due
Exam Week : Monday, 18 January 2021, 00:00	**Project Final Report** due

In addition, during/after Exam Week in January, we may ask you to give a short presentation/Q&A to show authenticity of your project.

Outline

The rest of the task sheet will describe all the project stages, the project grading scheme, and the dataset sources in detail.

Coursework 3 – Project Proposal

Project proposal is not graded, but it is *mandatory* to submit. You will *fail* this project if you do not submit it (0 total marks). Its purpose is to make sure you start early and we can give you feedback on your idea.

Project Proposal Submission

The project proposal is a PDF document, and need not be long. Write a few paragraphs describing what you have decided to do according to the subsection below.

Talk and discuss briefly with one of the TAs or the instructor during the second half of **Lab 5 (Fri, 30 Oct)** or **Lab 7 (Fri, 13 Nov)**. After getting approval, please submit the project proposal through the Learning Mall due **Friday, 13 November 2020, 23:59**.

Structure and Content

Your project proposal should include the following information:

- Background information: Explain what the application domain or field of research is. Why is the problem difficult or important? What specific questions will you try to answer?
- Dataset: Explain about what dataset(s) you plan to use. How many samples and features? If it's visual data, include an illustration if possible.
- Method: Explain what machine learning algorithms and techniques you are planning to use and why.
- Prior research: Explain specific and relevant literatures on the application. What has other people done?
- Initial exploration: If you actually have done some preliminary works, explain what you have already done (e.g., downloaded and played with data, tried k-nearest neighbors).
- Target contribution: What will be the core of the work for your project? How do you expect to spend most of your time? Do you expect to be judged primarily on your writing, your programming, your exhaustive exploration of machine learning algorithms and data, something else, or some combination thereof?
  (Since you are allowed to use Python data processing, data visualization, and machine learning libraries, you must be clear on what your substantial new contribution will be.)

Note that if your proposed project will be done jointly with another class' project/FYP, you must obtain approval from the other instructor ***and*** approval from me.

Coursework 3 – Project Interim Report

Project interim report is not graded, but it is *mandatory* to submit. You will *fail* this project if you do not submit it (0 total marks). Its purpose is to make sure you're on track and we can give you feedback on what you plan to do next.

Project Interim Report Submission

The project interim report is a PDF document, and need not be long. Briefly write a few paragraphs describing what you have done and accomplished so far, as well as what your next goals are according to the subsection below.

Talk briefly and convince one of the TAs or the instructor during the second half of **Lab 10 (Fri, 4 Des)** or **Lab 11 (Fri, 11 Des)** that you are making reasonable progress. After getting approval, please submit the project interim report through the Learning Mall, due **Friday, 11 December 2020, 23:59**.

Structure and Content

Write your project interim report as if it is an early draft version of the first few pages of your final project report. You can then re-use most of the interim report text in your final report.

Your project interim report should include the following information:

- Problem description: What problem(s) are you end up attacking? Are there any changes in the target contribution written in your proposal?
- Method: What machine learning algorithms and techniques have you tried and why? Which one seems to work well and why?
- Preliminary experiments: Describe the experiments that you have run, the outcomes, and any error analysis that you have done. You should have tried at least one baseline.
- Next steps: Given your preliminary outcomes, what are the next steps that you are considering?

Coursework 3 – Project Video

The project video will contribute to 50% of the total marks.

Project Video Submission

1. Upload the video file through the course Learning Mall, along with the source materials used inside, for example the PowerPoint slides, by **Monday, 11 January 2021, 00:00** (= Sunday, 10 January 2021, 24:00).
2. You must also upload the video on YouTube, YouKu, or Bilibili; and provide us with the link through the course Learning Mall.
   You may choose to keep the video private, where only those with the link can view it, in which case only the instructors will be able to do so. You can make the video public if you want to.

Structure and Content

1. Clear and understandable: the video should describe everything you think is important about your project (e.g., motivation, algorithms and techniques, findings and results).
2. Self-contained: any INT305 student should be able to understand the content of the videos without needing to consult any other materials.
3. Content is what matters: you can make the video as simple as slides with a voice overlay, or as fancy as you want.
   - As long as it is clear and understandable, you will not be graded on the fanciness of the video. We will grade solely based on the content.
   - However, if you have some spare time, do make it fun to watch!
4. Length: the video can be at most 3 minutes long.
   This is a very strict requirement: a video of length 3 minutes and 1 second will *not* be accepted. The length is whatever the uploaded file on Learning Mall and YouTube, YouKu, or Bilibili say it is.

Inspiration

Use the following videos as an inspiration for designing your video presentation:
https://www.youtube.com/results?search_query=3+minute+thesis
https://www.youtube.com/watch?v=BxZGyPg3L0E

Coursework 3 – Project Final Report

The project final report will contribute to 50% of the total marks.

Final Report Submission

The final report of the project is a Jupyter Notebook (.ipynb format). Write a concise report according to the outline below, and then submit it through the course Learning Mall by **Monday, 18 January 2021, 00:00** (= Sunday, 17 January 2021, 24:00).

The submitted report should contain all and only the Python code used in the project to produce the result. Specifically, the notebook should be arranged so that the reader can replicate your workflow by running the cells in the notebook in order.

In addition, you can either include a link to a GitHub repository or submit a zip file with the code for your project. You do *not* have to include the data or additional libraries.

Report Outline

Follow the required outline of your final report:

1. **Introduction**
   Explain the problem. Explain why it is important and your motivation for solving that problem. Give some background if necessary.
   Give a concise description of your project. State the input, the algorithm, and the output explicitly. For example, "The input to our predictor is a medical image. We then use logistic regression to output a predicted tumor growth rate".
   Enable the reader to get a rough picture of your report before continue reading it.

2. **Related Work**
   Find 2-5 existing works/papers, group them into categories based on their approaches, and discuss their strengths and weaknesses, as well as how they are similar to and differ from your work.
   Explain which approaches were clever/good, what the state-of-the-art is.
   Include previous attempts by others at your problem, previous technical methods, or previous learning algorithms.
   You may want to use Google Scholar to help you with that.
   https://scholar.google.com/

3. **Problem Formulation**

   Formulate the application as a machine learning problem.
   Start by describing your dataset: what the data points are, what features and labels characterize the data points, how many training/validation/test examples, and how you split them.
   Include a citation on where you obtained your dataset from.
   Include an example of your data, e.g. show an image, a waveform.

   Explain the preprocessing you did, normalization, data augmentation, data discretization.
   Describe feature engineering and extraction methods you used, for example using Fourier transforms, word2vec, HOG, PCA, ICA.

4. **Methods**

   Describe your machine learning algorithms and techniques.
   Describe what loss function is used to measure the quality of the predictor.
   Describe the metric that serves as the measure of quality of a machine learning model on your problem, e.g. the mean-squared-error.
   Make sure to include relevant notation. For example, you can briefly include an optimization objective formula from the lecture.

   For each algorithm, give a very-short description of how it works, in your own words. We are looking for your understanding of how these machine learning algorithms work. Although the teaching staff probably know the algorithms, future readers may not. Additionally, if you are using an algorithm not covered in the lecture, you may want to give a longer description.

   Describe how you applied the machine learning algorithms to solve the problem.
   Explain how you learned the predictor, which Python library you used.

5. **Experiments and Results**

   Describe the hyperparameters that you chose to optimize the model, and how you chose them.
   Describe the validation/cross-validation that you used, how many folds.

   Explain what your primary metrics are, e.g. accuracy, precision, AUC.
   Provide equations for the metrics if necessary.

Discuss the results obtained from the methods.
Explain what the training/validation error is, and how the results depends on the hyperparameters of the models.
Explain estimated performance of the final model on new data with respect to the chosen performance metric.

Include a mixture of tables and plots for results.
You should have both quantitative and qualitative results.
If you are solving a classification problem, you should include a confusion matrix or AUC/AUPRC curves. Include performance metrics such as precision, recall, and accuracy.
For regression problems, for example, state the average error.

Include visualizations of results, heatmaps, examples of where your algorithm failed, and a discussion of why your algorithms failed or succeeded.
In addition, explain whether you think you have overfitted to your training set, and what you did to mitigate that.

6. **Conclusion and Future Works**

Summarize your main findings during the project work, and reiterate key points.
Describe which algorithms were the highest-performing, and explain why some algorithms worked better than others.

For future work, if you had more time, or more computational resources, explain what would you explore, or what would you do differently to overcome some methodological shortcomings in your project.
Describe whether the results suggesting that the problem is solved satisfactorily, or there might be room for improvement.

7. **Bibliography**

Include citations for: (1) Any papers mentioned in the related work section. (2) Papers describing algorithms that you used which were not covered in class. (3) Code or libraries you downloaded and used. This includes Python libraries such as scikit-learn and TensorFlow.
Use this order: author(s), title, conference/journal, publisher, year.

Report Writing and Experiment Tips

- Writing a good report may take a surprisingly long time. The final report including the accompanying code and the video are the only insights we get into your project, so do put some effort into the report if you are aiming for a good grade.

- In contrast, the performance of your model is less important for grading, provided that your methodology is sound and that your model is a reasonable choice for the chosen problem. Spend enough time, but not too much time squeezing the final increases in performance on your task. Instead, use that effort to write a great project report.

- Balance your time in formulate the problem as a machine learning problem, engineering features, selecting and training the model, and tuning the hyperparameters. In particular, use model validation and parameter tuning tools available in scikit-learn model_selection module with good documentation.

- Try to use the notation used on the course if you use mathematical formulas or symbols. In the case that you want to use different notation, use good scientific writing principles and clearly define the meaning of your symbols.

- Please comment your code. The commenting doesn't have to be extremely comprehensive, but it should give some indication of what is happening in different sections of your code.

- Your plots should include legends, axis labels, and have font sizes that are legible when printed.

Coursework 3 – Project Grading Scheme

The same criteria will be used to grade both the project video (50 points) and the project final report (50 points). Each criterion will be worth max 10 points.

Evaluation

The video and the final report will be evaluated based on 5 criteria:

1. Relevance
   - Is the presentation or the report related to machine learning?
   - Is the problem formulated correctly as a machine learning problem?
   - Is the algorithms and techniques used related to topics taught in INT305?
   - Is the analysis, discussion, and/or conclusion framed as ones in machine learning formulation?

2. Significance and usefulness
   - Does the student choose a real-word problem to work on, or only a small toy problem?
   - Is this work likely to be useful and/or have impact?
   - Does the work answer the questions should be clearly stated in the proposal?
   - Does the problem sufficiently motivated or written well in the introduction?

3. Soundness
   - Is the presentation or the report technically correct? Reasonable arguments? Any methodological flaws?
   - Does the chosen dataset have enough examples to get statistically significant results?
   - Does the student conduct sound numerical experiments, such as splitting the data into training/validation/test sets; making comparative result tables using validation or cross-validation; using the test set only for final assessment?
   - Do graphs and other means of visualization added to support an argument, a selection, or a result?
   - Is the metric that serves as the measure of predictor quality defined, used, and computed correctly?

4. Clarity and presentation
    - Are the presentation and report organized well? Logical placement of code cells and images? Excellent timing? Good bibliography?
    - Are the methods and process used in solving the machine learning problem described in an effective flow of presentation? Sufficiently detailed?
    - Are the results of the experiments presented clearly?
    - Are there enough graphs and visual support?
    - Concise presentation/report and not lengthy? Pleasure to watch/read?

5. Novelty and originality
    - Is this project applying a common technique to a well-studied problem, or is the problem or the algorithm relatively unexplored?
    - Is there enough literature review, where the most important papers or significant results in the area are mentioned and give proper credit?
    - Does the student convey novel insight about the problem and/or algorithms after comparing to those original problems or existing approaches?

Coursework 3 – Data Sources

There are a number of public data sources and machine learning tasks, including:

1. Kaggle
   https://www.kaggle.com/datasets
   https://www.kaggle.com/competitions
2. CodaLab
   https://competitions.codalab.org/my/
3. AIcrowd
   https://www.aicrowd.com/challenges
4. OpenML
   https://openml.org/
5. Machine Learning Contests
   https://mlcontests.com/

You may choose other sources.

This is the end of the INT305 2020 Coursework 3 Task Sheet.