MA429 – Algorithmic Techniques in Machine Learning

Formative Group Project Description

Introduction

This document is long but important. Please skim it before starting your project and refer to it as needed, and for periodic reminders of helpful hints.

Much detail follows, but a guiding principle is to make your project as interesting as possible. How would you summarise the work to a classmate, friend, or prospective employer? (I suggest actually doing so; it is a helpful and revealing exercise.) Did the project lead to any conclusions about its subject? Were some methods much more successful than others? Did you encounter any problems that might be worth further attention in the future? Whatever else you say, you are going to have to explain the data and what it represents, the question you're asking, and the techniques you're using. Doing so will require technical explanations and probably some notation, but these should be in service of something interesting, not technicalities for their own sake.

Overview

This formative group project will form a basis for the group summative project, and you will work in the same team for the two projects. Here you must work with the dataset below; for the summative project you will be able to choose a dataset (with approval from me), and you will be expected to go into greater depth. Both projects will be marked out of 100.

The main body of your report can only be **8 pages**; the summative will have a significantly longer page limit. I'm deliberately keeping this quite short; this is after all formative, and the purpose is to get some useful feedback from me and to practice working together as a group. See more details under "Organisation and presentation" below.

The project itself is a team work. The team will submit just one report, and normally (though not always) all students in the team will get the same mark for that. Each member of the team will also separately submit a contribution statement. For the summative project (but not the formative) there will be a short Q&A session with each team after the submission deadline, to discuss the project. This will not require preparation, but is intended for me to ask questions I may have based on my initial assessment, and if necessary, get clarity on each team member's contribution.

Marking

Marks will be awarded following this breakdown:

30 Context, modelling and originality: How well is the context of the problem, and the results, explained? As seen in the executive summary, introduction, literature review and conclusion.

Appropriate exploratory data analysis and understanding of data, intelligent modelling choices, and appropriate evaluation metrics.

The extent to which the project goes beyond standard existing approaches, or takes a novel perspective.

40 Analysis: How well is the main work of analysing the data done? This includes pre-processing, choice of learning methods, correct use of methods, correct methodology (e.g., train/validate/test), parameter tuning, interpretation of results.

Credit for handling challenges (e.g., using more difficult methods).

20 Organisation and presentation: In the large: logical and structured discussion, clarity of expression, appropriate level (avoid "mini-lectures").

In the small: layout, typesetting, table of contents, bibliography, clarity of figures, absence of typos – overall polish.

10 Code quality: Well-structured and easy-to-understand code; should be well-commented. It should be possible for me to reproduce your results from your code without too much difficulty.

Individual adjustment: This will not apply to the formative, but I may make some remarks if I see clear issues. For the summative, this will take into account information contained in the contribution statements, and on information I learn from the Q&A. In the hopefully most common case of (roughly) equal contribution, each team member will receive (roughly) the same mark.

Submission & deadline

The submission deadline is **Thursday 6th March 2025, 18:00**, to the Moodle submission link that will be provided.

Please include all group members' registration numbers and names within the submission, to be sure. Do **not** use your candidate exam numbers anywhere in these submissions!

Each **group** should submit the following, just *one copy per group*:

- A softcopy of the report, in pdf format. Please name the submission with one group member's registration number, e.g., `202054321 report.pdf`.
- An electronic appendix including your code, computational results, and any additional data you deem important, in a single zip file. Please name this by the same group member's registration number, e.g., `202054321.zip`.

Each **individual** should submit:

- A completed contribution statement via the Moodle form. This should be a detailed account of what parts of the project you were responsible for, and what parts your teammates were responsible for.

  The deadline for this will be slightly later – midnight Thursday 6th March.

Dataset and task

The dataset below is a classification problem from the UCI Machine learning repository, based on US census data.

https://archive.ics.uci.edu/dataset/2/adult

Your primary task is to predict whether income exceeds $50K based on the given features. Time and space permitting, some interesting exploration beyond this is encouraged. E.g., by making use of your classifier, can you glean interesting insights about factors influencing income, or about inequalities?

Data conversion

The datasets may require some work to open in R/Python. Look at the different files in the data folder to understand the data structure. You may be able to use standard packages to read the data, but if the input format is broken or deviates from the standards you may need to use some other editor first. Watch out for issues such as capitalisation of words, or number formats.

This step might be a bit tedious, but is doable with basic computer skills, not requiring specialist knowledge. It is an inherent part of working with real-world data and will recur in the summative project and in life after graduation.

Modelling & preliminary analysis

Before running experiments with the various classifiers, familiarise yourself with the data. This is a well-studied dataset with literature available online, see citations on UCI. You are not expected to become experts on census-taking, but you need to make an effort to *understand the dataset*. This is particularly important, and even more so in real life than in these projects.

Pay careful attention to the meaning of the fnlwgt value, which is admittedly poorly documented. It describes the **number of people** represented by this instance. So rather than repeating an identical instance (say) 20 times, there would just be a single instance with fnlwgt = 20. *Think very carefully about how you will make use of this value.*

Can you observe any interesting patterns? What could be the impact of different attributes, which ones seem more likely to influence success, etc. Think of a few interesting questions or hypotheses about the data. These should address the *domain* (and not the machine learning techniques or parameter values). What are potential applications of the data, and what useful/interesting conclusions could you hope to draw from it?

Input data & preprocessing

Before experimenting with various machine learning methods, you should do some exploratory analysis and pre-processing. Here are some general guidelines; not everything is necessarily applicable, and further methods may also be useful.

- Is the full dataset relevant and necessary? Remove instances that seem erroneous or redundant.
- If you are working on a classification problem, and the representation of the different classes is highly imbalanced, consider whether this should impact your ideas about how to evaluate the performance of a method, or your methodology going forward.
- Are there missing values? (How they are shown may vary from dataset to dataset.) Decide how you will interpret and handle missing values.
- Is the format of each attribute (nominal or numeric) appropriate to its contextual meaning? Is the range of the numeric attributes appropriate? Carry out conversions and normalizations as appropriate.
- Are there any obvious outliers? Instances with values very different from the typical range might represent input errors; is there evidence for that in the data?
- Are there any apparently irrelevant attributes that should be removed?
- Does it make sense to work with a smaller set of attributes?

In the report, describe your dataset – what types of attributes it has, perhaps the value ranges, interesting characteristics, relevant features, etc. Give your reader a general understanding of the data, and details that you think are important or illustrative. You do not necessarily need to describe every single input parameter. Your report should describe all preprocessing steps you performed, with justification of the decisions you made. You can carry out the preprocessing steps in R/Python or otherwise.

Keep the discussion to what is relevant or interesting: avoid unnecessary descriptions of 50 different predictors, their precise definitions, the distribution of each, all the pairwise correlations, and so on. If you are logarithmically transforming a set of predictors because they have a wide range and seem to give a better fit when transformed, say so and put lengthy details in an appendix. (This means your work can be reproduced and checked, while the typical reader can skip the boring stuff.)

Performance metrics

Explain how you evaluate the success of your methods, e.g., misclassification rate, confusion matrix, etc. Keep in mind the limitations of misclassification rate that we have discussed; it may or may not be a good choice, depending on the context and goals. Make sure that in your evaluations no data that is used for testing has been previously accessed for learning, model selection or hyperparameter tuning. I will be harsh on poor methodology.

Experiments with machine learning methods

You should try a variety of methods on the dataset, then focus on the most promising ones. Ideally, having identified which methods seem to work best, you will refine your approach further; however, there may not be time for much of that in the formative.

- You can use arbitrary methods, not limited to those covered in the lectures.

- For methods not covered in the course material, give references and a brief description, including the method's input, output, and aims (perhaps what error function it minimises, or attempts to minimise). Give enough information so that a reader with basic data mining knowledge (MA429 level) can understand roughly what the method is doing and interpret its results, and to demonstrate that you understand the method and why you are using it.
- For methods we have or have not covered, do *not* give a lecture about the method in the report.
- Where your method relies on parameter choices, state the values you used and justify them (briefly or in more detail as appropriate).
- Describe the resulting model(s), especially the more promising ones. If a model is relatively concise, try to summarise it in words, focusing on important or interesting patterns. Reflect on how the model answers the objectives of the experiment.
- State the performance of your model in terms of your chosen performance metrics. Say how long it took to fit the model.
- Justify the choice of methods you experimented with. You are encouraged to experiment with multiple methods, but you cannot do everything. *A good analysis with fewer methods is preferable to a superficial analysis with more.* This holds especially for groups with only 2 students instead of 3.
- The main task is classification. Consider other perspectives, such as clustering, only if you wish to and feel that there may be something to be gained. Especially in this short formative project, there is neither time nor space to try too many things.
- You may run into problems with running time on large data sets. If so, first check that you are not doing something silly with parameter settings or whatnot. If the issue is real, take smaller samples of different sizes and see how the running time scales: when you take twice as much data, is the running time 2 times as long, 4 times as long, or does vary in some other way? This will give you a sense of how big a training set you can accommodate within your project scope.

Summary of results

- For whatever problem(s) you considered, compare the performance of the various models you used, with one another and what you found in the literature. Discuss their comparative advantages and disadvantages. How could you take things further in subsequent work?
- Elaborate on the initial questions and hypotheses you may have formulated about the data. Did the models elucidate them? Did you get acceptable answers, justifications or refutations? What were your most surprising or interesting discoveries?
- What is the possible impact of your findings? Are there any potential practical consequences?
- What are the ethical implications, if any, in aspects ranging from the data collection to the potential impact of your findings?

- What is the overall conclusion of the project? Discuss possible strengths and weaknesses.
- Remember that your *results do not have to be good.* In academic work, especially a very limited project like this, it is perfectly reasonable to report that you tried some things and they did not work well. Of course, good results are always nicer than poor ones, but you will be marked on having tried sensible things in a sensible way, and described the work clearly, more than on the success of your work.

Organisation and presentation

"Presentation" doesn't mean flashy graphics and such but a clear, accurate, easy-to-grasp presentation of your work. Your report should be well structured and *concisely written.*

- The title should be short but informative.
- Use an 11pt font for your body text.
- Start your report with an executive summary of around **half a page**. This should summarise the setting of the problem investigated and your main conclusions. It should focus on the problem domain and should not contain any technical detail.
- The report should include background information on the dataset, your preliminary analysis, preprocessing steps, the various data mining techniques you applied, and a summary of your results. **This should be kept brief.**
- Include only the most important and relevant graphs, tables, decision trees, etc. in the body of your report. Any that are not crucial for the analysis but still potentially interesting can be put into appendices (clearly labelled, and referenced within the body of the report).
- The main body of the report, everything excluding the cover page, executive summary, table of contents, bibliography and appendices, may not exceed **8 pages**. This is a (strict) upper bound, not a target. There are no page limits for the appendices, but you are strongly advised to keep them short, and include computer output only if it is referred to in the main text.
- There are many ways to write a good report; there is no one best format. Structure your report in the way that best fits the story you are telling.

Code

Use R or Python for your analysis, taking advantage of any packages you wish. Your analysis must be reproducible. That is, it should be fairly straightforward for me to take your submitted code along with the dataset from the UCI repository, and (after installing any needed packages) execute the code on my own system and obtain the results you quote in your report. You may wish to include versions of the dataset with your submission; e.g., if you created a new version of the dataset with some new synthetic features, and then trained a method on that, including this file may be the easiest way to make everything reproducible.

Code should ideally be in a well-documented and commented R markdown file, or a well-documented and commented Jupyter notebook. It's best if it's a single notebook, well-structured so that it's easy to map between the report and the code. Spending some time at the end polishing this is recommended.

Plain (well-commented) R or Python code is acceptable, but extra care will be needed to ensure that the code is readable, and that it's clear in what context a piece of code should be applied, especially if there are multiple files. If the code in file 1 applies some pre-processing to the dataset and produces a new pre-processed dataset, and file 2 needs this pre-processed dataset to exist first before it can be used, make sure that this is clear.

Literature & methods

The UCI websites contain publications related to the dataset. You are encouraged to read these: they can include useful background information on the datasets, and they describe some data mining techniques applied to them. You can refer to these in your report. However, if you choose to *apply* methods used by previous researchers (not just refer to them for background) you must perform these methods yourself. A critical reading of the papers is advised.

As previously noted, you are not restricted to using methods covered in the course. You are encouraged (more for the summative project than for this one) to explore further chapters of the course textbooks or other sources: there is a vast amount of good material on the Web. Be sure you understand any technique you use (at least well enough to use it properly), especially if it is significantly different from the techniques we have covered. All the techniques we covered have extensions, variants, and hybrids; I suggest first thinking of ideas you think might help, and then looking to see if they have been thought of before (and perhaps implemented in an R package).

References, citations, plagiarism, formatting

Cite, in proper bibliographic format, any works you rely upon. Even for standard techniques like SVM, it makes sense to say "see for example [5]" (where [5] is a proper citation to "ISL"). Please use a bibliography style that lists citations in alphabetical order by author. Citing sources and indicating direct quotations (with quotation marks as well as citations) is good form and shields you against plagiarism concerns.

If you are a LaTeX user, it (especially with BibTeX) can help make your citations and bibliography easy, flexible, and standard. You can also use `knitr` to embed R code and output into your TeX document.

Team work

You can divide the work however you like, as long as it is fair. Since the whole group is judged on the one project, it would be wise for everyone to look over everything, as it progresses and at the end. Budget time for this.

Historically, teams have mostly worked well. At the end, most groups reported a fairly equitable distribution of labour, and group members received equal or similar marks. It's also fine if group members say "A did more than B did more than C", especially if you all agree: A will get a few marks more, C a few marks less, and everyone should be happy. **In cases where a team member has made no serious contribution, this can be**

treated the same as the lack of a serious attempt at an exam, and the student will have to perform an alternative, individual piece of work at a later date.

Should team work be going poorly, first try to resolve the issue yourselves, but then alert me. Do so as early as possible, so that there is still time to remedy the situation; this is something to handle during the formative, so that things go smoothly in the summative. Our experience so far has been that most groups work well, and we expect it to continue so.

I strongly recommend meeting in person (or by Zoom etc) especially if there are any emotional issues: email is terrible for that, usually making things worse fast. It is generally recommended that any meeting is summarised in print right away, especially any "action items". I recommend email over other print communication, using your LSE email account and leaving a paper trail should it be needed for any reason.

Contribution statement

You will each individually (not as a group) submit a form on Moodle explaining the role and level of contribution of each team member (including yourself). Be specific, without going into unnecessary detail. E.g., not: 'we all wrote some code'; what specific aspects did you code? What parts of the report were you primarily responsible for? Etc. Experience suggests it is best if you organise this by group member: for each person, say what they contributed.

This will be submitted using an online form in Moodle. Write this in a text editor or Word, and copy this into the form.