MA429 Algorithmic Techniques for Data Mining, 2019-20

Gregory Sorkin, 2020.02.28 (v1)

Formative ("Mock") Group Project

Introduction

This document is relatively long but important. Please skim it before starting your project and refer back to it as needed.

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? (Actually doing so is a good 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, 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 just technicalities for their own sake.

Overview

This group mock 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 one of two datasets below; for the summative project you will have a wide choice of datasets, and will be expected to go into greater depth. Both projects will be marked out of 40.

The mock project itself is a team work. The team will submit just one report, and all students in the team will get the same mark for it. Each member of the team will also separately submit an Individual Reflection and Contribution form. This will result in individual marks.

Dataset options

The two datasets below are classification problems from the UCI Machine learning repository and another public source. You can find a detailed description of each set at the given links.

US census dataset

> https://archive.ics.uci.edu/ml/datasets/Adult
> A larger, more challenging version is also available:
> https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29
> You can use either of the two versions.

Heart disease dataset

> https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Data conversion

The datasets may require some work to open in R. Look at the different files in the data folder to understand the data structure. You may be able to use standard R 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.

Preliminary analysis

Before running experiments with the various classifiers, familiarise yourself with the data. These are well-studied datasets with literature available online, see citations on UCI. You are not expected to become experts on census-taking or cardiology, 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.

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 data mining 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 data mining techniques, you should do some exploratory analysis and pre-processing. Here are some general guidelines; not everything is necessarily applicable to all datasets, 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, you may want to resample the data to represent the classes more equally. (Or not. What is your goal in terms of, say, false positives and false negatives. How will rebalancing affect that? Use reasoning and experimentation.)
- 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, the value ranges, interesting characteristics, relevant features, etc. Some preprocessing steps have already been applied for some of the datasets. 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 or in otherwise. Keep the discussion to what is relevant or interesting: avoid unnecessary descriptions of 50 different predictors, their precise definitions, the distribution of each, 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 that for example misclassification rate has limited value if the class proportions are skewed or different error types have different costs. The problem domain may suggest a metric (maximise net profit). Make sure that in your evaluations no data that is used for testing has been previously accessed for learning.

Experiments with data mining methods

You should try a variety of methods on the dataset, then focus on the most promising ones.

- 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 for R to construct 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.

- In both datasets, 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 mock 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 work.

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.

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.
- Start your report with an executive summary of about a page, maybe less. 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.

- 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 report should not exceed 12 pages plus cover page, bibliography and appendices. This is an 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. We've provided several samples, but you don't have to follow any of them. Write that way that suits your personal style, and that fits the story you are telling.

Software usage

Please use R for your analysis, taking advantage of any packages you wish. Your analysis must be reproducible: make sure you describe all technical steps and parameter settings, so that someone working with the same dataset can obtain the same results. Include high-level aspects of this in the body of the report, and further details in appendices. *Submit a clearly written and well-documented R code.*

Literature & methods

The UCI websites contain publications related to the datasets. 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 [ISLR]" (but with "ISLR" replaced by a proper citation). 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 protects 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.

Division of 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 mark or two more, C a mark or two less, and everyone should be happy. In extreme 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 and/or Rebecca Batey. Do so as early as possible, so that there is still time to remedy the situation (especially, before the summative project). However, our experience so far has been good, and we expect it to continue so.

Individual reflection & contribution form

You will each individually (not as a group) submit an Individual reflection & contribution form. It has two purposes.

1. To explain the role and level of contribution of each team member (including yourself). Please try to keep this fairly short and comprehensible. Experience suggests it is best if you organise this by group member: for each person, say what they contributed.
2. To reflect on how the project went, particularly on your own performance: what you did that you are happy with, what you are unhappy with, and what you would do differently next time. (Or start / stop / continue, if you prefer. It's your reflection!)

Marking

Marks will be awarded following this breakdown:

Preprocessing and preliminary analysis	7
Experiments with different methods	8
Summary and interpretation of results	7
R implementation	6
Structure and presentation	6
Individual reflection & contribution	6
	40

Submission & deadline

The submission deadline is **Thursday 19th March 2020, 4pm.**

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 **individual** should submit:

- A completed Individual Reflection and Contribution form.

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

- A print copy of the report, to COL 4.01.
- A soft copy of the report, in pdf format, via the Moodle link. Please name the submission with one group member's registration number, e.g., 201912345.pdf.
- An electronic appendix including your R code, computational results, and any additional data you deem important, in a single zip file, also via the Moodle link. Again, please name this by the same group member's registration number, e.g., 201912345.zip.