Assignment 3

Thomas Li

2025-05-14

Marking Rubric

On a high level:

- C grade (50%-65%): You have attempted the assignment, but in a limited way that might not show demonstrate full understanding of the concepts involved.
- B grade: (65%-79%) You have done what you were asked to do, and you did it correctly
- A grade (80%-100%): In addition to satisfying the requirements for a B grade, you have done something beyond, like:
    - exploring an additional avenue of research,
    - including work beyond the lecture content,
    - discussing current debates in the subject area,
    - demonstrating extraordinary investigative research or
    - writing an exceptional report,
    - …

You are usually not expected to write any long essays, so keep it brief, legible, and understandable. Support your arguments with figures, mathematical arguments, or references, where suitable.

In slightly more detail: Your assignment will be marked according to the following criteria. The relative weightings are only an approximate guide and will vary between questions.

Skills	Expectation for a C grade	(additional) Expectation for a B grade	(additional) Expectation for an A grade	approximate relative weighting
Communicating clearly via text	Text is in understandable English. For most questions, a few short sentences will be totally adequate, unless stated otherwise.	Uses adequate technical jargon, and concise language.	“Generic garbage” as sometimes produced by GenAI will lead to detractions here.	**necessary prerequisite.** Anything else cannot be marked if it is not communicated clear enough

Skills	Expectation for a C grade	(additional) Expectation for a B grade	(additional) Expectation for an A grade	approximate relative weighting
R Markdown and format of submission	Submission is a single file in either .html or .pdf format. File is generated by R Markdown. All code is shown in the html document.	—	—	**necessary prerequisite.**
R programming skills	Syntax correct, code runs without errors.	Code does what it is intended to do. Code is readable and uses code comments and suitable naming conventions to clarify what is being done. `dplyr`, `ggplot`, and statistical learning modules are employed in a suitable way.	Computations are reasonably efficient.	15%
Knowing and applying statistical learning algorithms	Basic statistical learning task is addressed.	Correct choice (within constraints of admissible modules) and application of algorithm.	—	15%
Interpreting and evaluating algorithm results and output	Some interpretation of the findings is provided, even if not correct.	Algorithm results are (when required) put into context, correctly explained, and implications as relevant for the assignment are correctly identified.	—	25%
Communicating clearly via figures	An attempt at visualisation (if required) is made.	Axes are labelled, figure is suitably scaled, intention is clearly communicated. Interpretation of the figure is given in either an expressive caption or in the accompanying full text.	—	35%

Skills	Expectation for a C grade	(additional) Expectation for a B grade	(additional) Expectation for an A grade	approximate relative weighting
Extra Effort	—	A good report will use the models and statistical methods to the extent developed in class.	A very good assignment (e.g., A- or better) will extend to something not covered, and/or show reflection beyond the narrow research question. Sometimes, question parts marked with “extra effort” give an indication for possible avenues to extend your work.	10%

Question A: Classifying volcanic rock

In this question, you are allowed to use any method available to you in R, either programmed yourself, or from packages of your choosing. The important thing here is that you communicate your results very well, in terms of language, illustration, and precision.

In this question you will analyse a dataset containing volcanic rocks and their chemical composition, collected from various sites in New Zealand. You will predict whether the `Source` of a given rock sample is “Okataina” (a geographical location), or whether the sample comes from somewhere else.

Load the annotated dataset `rocks.csv` . and the unlabelled dataset `rocks_unlabelled.csv` . You will find that rows in the latter set do not contain an entry for `Source` , i.e., they are of unknown origin. Your goal will be to fill these blanks, using the techniques you have learned in this course so far. The following guidelines might be useful to consider:

- Use a classification-tree-based method (can also be random forests or boosting), and at least one additional different data mining algorithms (two tree variants are not considered “different”). Compare their performance in a suitable way.
- It will not be enough to just predict some `Source` labels for the missing data, but you will need to quantify and justify the accuracy of your predictions. It is OK if you think that your method is not perfect, but be sure to point out its limitations in this case. You will not be judged on accuracy, but on correctness of procedure, and clarity of communication.
- How exactly you prepare and work with the dataset is your choice, but follow good practice models as demonstrated in the course (for example, consider “data hygiene”, i.e., don’t test on training data etc.).
- Be brief, exact, correct, and communicate well.

Question B: Clustering seeds

In this question, you are not allowed to use any pre-implemented modules for performing k-means, expectation maximisation, or other clustering algorithms. You will need to provide your own implementation.

We will work with a dataset `seeds.csv` containing measurements of a large number of seeds. Each seed is measured in terms of “Area”, “Perimeter”, “MajorAxisLength”, “ConvexArea”, and other optical characteristics of this seed. Your goal in this question will be to find clusters in this dataset, i.e. try and find groups of “different types of seed” (the seeds come from different plants, but this information has been lost, so all we can do is try and estimate how many different types of seeds we have).

1. Perform data preprocessing, if appropriate.
2. Investigate the question “how many types of seeds are there”, and present your results briefly, but coherently.
3. *Opportunities for showing extra effort:* The additional dataset `seeds_annotated.csv` contains (few) datapoints which have class labels (A,B,…,G), annotated by an expert of seeds. Use this, in combination with your results from your clustering analysis, to train a seed type prediction algorithm that takes the measurements of a seed and predicts its type. Test your algorithm in a suitable way and communicate your findings (you need to be careful not to check performance purely on the annotated training set). [You do not need to implement your algorithms yourself in this part of the question, feel free to use whatever code works best for you]