Unsupervised Learning and Dimensionality Reduction Report
CS7641: Machine Learning
Summer 2025

1 Assignment Weight

The assignment is worth 15% of the total points.

Read everything below carefully as this assignment has changed term-over-term.

2 Objective

Now it is time to explore unsupervised learning algorithms. This part of the assignment asks you to use some of the clustering and dimensionality reduction algorithms we've looked at in class and to revisit earlier assignments. The goal is for you to think about how these algorithms are the same as, different from, and interact with your earlier work.

The same ground rules apply for programming languages and libraries. You may program in any language that you wish insofar as you feel the need to program. As always, it is your responsibility to make sure that we can actually recreate your narrative, if necessary.

3 Procedure

3.1 The Problems Given to You

You are to implement five algorithms.

The first two are clustering algorithms. You can choose your own measures of distance/similarity. Justify your choices.

- Expectation Maximization
- K-Means Clustering

The last three are linear dimensionality reduction algorithms:

- Randomized Projections
- Principle Component Analysis (PCA)
- Independent Component Analysis (ICA)

You are to run several experiments with the goal of disseminating how dimensionality reduction affects your data. You will use the same two datasets from the previous assignments. You will develop hypotheses based on your datasets and the following exploration. These hypotheses should be well-posed and grounded in theory from the lectures and readings.

Extra Credit Opportunity:

There is an opportunity to add 5 points of extra credit. In addition to the above algorithms, you will also implement a **Non-linear Manifold Learning Algorithm of Your Choice** as both a comparison and a visualization of your datasets. You will need to justify your choice and you should briefly describe whatever it is that you do use. This is not mandatory and may require more time than allotted.

3.2 Exploration

The following should comprise your exploration.

1. Apply the clustering algorithms on the datasets. You will report on each of the clustering algorithms for each dataset, resulting in 4 demonstrations.
2. Apply the dimensionality reduction algorithms on the datasets. You will report on each of the dimensionality reduction algorithms, resulting in 6 demonstrations.
3. Re-apply the clustering algorithms on the set of dimensionality reduction datasets. This will result in 12 combinations of results of datasets, dimensionality reduction, and clustering methods. You should look at the full scope of the results and note how they might pertain to your hypotheses. Tables are helpful with this section. In particular, focus on more interesting findings. You will be reporting what combination performs the best. Justification will be especially important as space is limited in the report.
4. Choose one of the datasets. Re-run your neural network learner from the SL Report with each of the dimensionality reduction algorithms applied. You will need to give observations and discussion for the different linear methods from Step 3 (RP, PCA, or ICA) and what performed the best for each dataset with reason. If you are doing the extra credit, you will also need to use the non-linear manifold learning features as well to receive full extra credit.

Analysis writeup is limited to 8 pages. The page limit does include your citations. Anything past 8 pages will not be read. Please keep your analysis as concise while still covering the requirements of the assignment. As a final check during your submission process, download the submission to double check everything looks correct on Canvas. Try not wait until the last minute to submit as you will only be tempting Murphy's Law.

In addition, your report must be written in LaTeX on Overleaf. You can create an account with your Georgia Tech email (e.g. gburdell3@gatech.edu). When submitting your report, you are required to include a 'READ ONLY' link to the Overleaf Project. If a link is not provided in the report or Canvas submission comment, 5 points will be deduced from your score. Do not share the project directly with the Instructor or TAs via email. For a starting template, please use the IEEE Conference template.

Add in github description.

Update for Summer 2025

The following datasets are required for the Summer 2025 cohort. Each semester these datasets will change. This is due to a variety of reasons concerning simplicity and overuse of common ML datasets. These datasets are mid-sized and provide many angles of analysis due to the complexity of features and domain knowledge. Each dataset can be found on Canvas if access to the original download is limited. If these datasets are not used, you will receive a zero for the assignment.

- **Global Cancer Patients**: Kaggle Repository: *Global Cancer Patients*
- **Company Bankruptcy**: Kaggle Repository: *Company Bankruptcy*

3.3 Acceptable Libraries

Here are a few **examples** of acceptable libraries. You can use other libraries as long as they fulfill the conditions mentioned above.

Machine learning libraries:

- scikit-learn (python)
- Weka (java)
- e1071 (R)
- ML toolbox (matlab)

Plotting:

- matplotlib (python)
- seaborn (python)
- yellowbrick (python)
- ggplot2 (R)

4 Submission Details

All scored assignments are due by the time and date indicated. Here "time and date" means Eastern Time (ET). Canvas does not currently support Anywhere On Earth, so this is the best alternative we can offer being at Georgia Tech. Please double check your settings and assignments for the exact due dates to mark your calendars appropriately. As a good check, you should go to settings on Canvas and set your time zone.

All assignments will be due at 11:59:00 PM ET on the the final Sunday of the unit. However, since we will not be looking at the assignments until morning, you will have officially until 7:59:00 AM ET until the assignment is marked late. I understand that there are many circumstances that you may need an additional hour or two to complete the assignment. I will be asleep through the night and see no issue in giving the extra time.

However, I need to heed a stern warning. You should use the 11:59PM timestamp as your internal deadline rather than the 7:59AM official cutoff. Staying up all night is a detriment to your mental health and may not be as conducive to constructive writing. I know there is a colloquialism where nothing would get done unless for the last minute, however I do hope you all manage your time wisely. Please note the exact time for the submission as many situations may incur Murphy's Law. Allow a couple of minutes for the submission upload and check as it does take a few seconds on average to upload an assignment in Canvas.

Late Due Date [20 point penalty per day]: Indicated as "Until" on Canvas. The late penalty is not on a racked scale, but rather wholistic day-to-day. Meaning, if you do utilize the late penalty, you have the full 24 hours before another 20 point penalty incurs.

You will submit **two PDFs**:

1. You must submit a PDF containing your SL Report. Your document must be written in LaTeX using Overleaf.
2. Additionally, you will submit a second PDF titled `DOCSTRING-GTUsername`, where `GTUsername` is the first part of your Georgia Tech email address (e.g., `gburdell3@gatech.edu` → `gburdell3`). This document must include two links and code instructions:
   (a) A **READ ONLY** link to your Overleaf project.
   (b) A **GitHub commit hash** from the final push of your report.
   (c) Instructions to run your code.

   - When submitting your answers, you are required to include a READ ONLY link to the Overleaf Project. **Please do not send any email invitations to join the project.**
   - You are required to use the GT Enterprise GitHub for all course-related code. While personal GitHub accounts are common, using the GT Enterprise GitHub helps mitigate potential plagiarism and violations of the student code of conduct. This must be the actual hash, not a general link.
   - You need to include instructions for running your code. Typically, this will be the content you create for your README.md on Github. We need to be able to get to your code and your data. Providing entire libraries isn't necessary when a URL would suffice; however, you should at least provide any files you found necessary to change and enough support and explanation so we can reproduce your results on a standard Linux machine.

For a starting template, we recommend using the IEEE Conference template[1].

Only your **latest submission** will be graded. Please double-check that **both PDFs** are submitted.

[1] `https://www.ieee.org/conferences/publishing/templates.html`

Your report should contain:

- Brief description of the datasets, and hypotheses you want to highlight in your report.
- Explanations of methods. This is your opportunity to demonstrate nuances needed to support your hypotheses.
- Grounded descriptions of resulting clusters. Support descriptions with data-driven evidence.
- Analyses of your results. Why did you get the clusters you did? Do they make "sense"? If you used data that already had labels (for example data from a classification problem from assignment #1) did the clusters line up with the labels? Do they otherwise line up naturally? Why or why not? Compare and contrast the different algorithms. What sort of changes might you make to each of those algorithms to improve performance? How much performance was due to the problems you chose? Be creative and think of as many questions you can, and as many answers as you can. Take care to justify your analysis with data explicitly.
- Can you describe how the data looks in the new spaces you created with the various dimensionality reduction algorithms? For PCA, what is the distribution of eigenvalues? For ICA, how kurtotic are the distributions? Do the projection axes for ICA seem to capture anything "meaningful"? Assuming you only generate k projections (i.e., preforming dimensionality reduction), how well is the data reconstructed by the randomized projections? How much variation did you get when you re-ran your random projections several times? How does noise affect each algorithm? What is the rank of your data? Can you describe how colinear your data is both qualitatively and quantitatively? How might specific properties of your data influence outputs of various algorithms?
- When you reproduced your clustering experiments on the datasets projected onto the new spaces created by ICA, PCA, and RP, did you get the same clusters as before? Different clusters? Why or why not? Remember to justify why one output might be more interesting when choosing your demonstrations.
- When you re-ran your neural network algorithms were there any differences in performance? Speed? Consider how you might judge differences in performances and include these notes in your discussion.

It might be difficult to generate the same kinds of graphs for this part of the assignment as you did in previous assignments; however, you should come up with some way to describe the kinds of clusters you produce. If you can achieve this visually all the better. However, a note of caution. Figures should remain legible as we are asking for several demonstrations in many sections. Do not try to squish figures together in specific sections where axis labels become 8pt font or less. We are looking for clear and concise demonstration of knowledge and synthesis of results in your demonstrations. Any paper that solely has figures without formal writing will not be graded. Be methodical with your space.

You may submit the assignment as many times as you wish up to the due date, but, we will only consider your last submission for grading purposes.

Note: we need to be able to get to your code and your data. Providing entire libraries isn't necessary when a URL would suffice; however, you should at least provide any files you found necessary to change and enough support and explanation so we can reproduce your results on a standard linux machine.

5 Feedback Requests

When your assignment is scored, you will receive feedback explaining your errors and successes in some level of detail. This feedback is for your benefit, both on this assignment and for future assignments. It is considered a part of your learning goal to internalize this feedback. We strive to give meaningful feedback with a human interaction at scale. We have a multitude of mechanisms behind the scenes to ensure grading consistency with meaningful feedback. This can be difficult, however sometimes feedback isn't always as clear as you need. If you are confused by a piece of feedback, please start a private thread on Ed and we will jump in to help clarify.

Change for Summer 2025. Reviewer Response. In an effort to learn and grow assignment-to-assignment, we will provide a mechanism to edit and respond to your feedback. We will call this the *Reviewer Response.* You will have one week from the assignment grade being posted to edit and provide a two-page maximum response with both edits made and reviewer feedback. You will need to reasonably respond and edit your initial paper submission to improve your paper in good faith. Both the initial submission, revised submission, and two-page response will be needed for a proper Reviewer Response. If satisfied, you will receive half of the missed points

back for the assignment. For example, if the initial grade was a 70/100, if everything is satisfied for the Reviewer Response, there will be 15 points added resulting in an 85/100. Further examples will be provided when the assignment grades are posted.

6 Plagiarism and Proper Citation

The easiest way to fail this class is to plagiarize. **Using the analysis, code or graphs of others in this class is considered plagiarism**. The assignments are designed to force you to immerse yourself in the empirical and engineering side of ML that one must master to be a viable practitioner and researcher. It is important that you understand why your algorithms work and how they are affected by your choices in data and hyperparameters. The phrase "as long as you participate in this journey of exploring, tuning, and analyzing" is key. We take this very seriously and you should too.

What is plagiarism?

If you copy any amount of text from other students, websites, or any other source without proper attribution, that is plagiarism. The most common form of plagiarism is copying definitions or explanations from wikipedia or similar websites. We use an anti-cheat tool to find out which parts of the assignments are your own and there is a near 100 percent chance we will find out if you copy or paraphrase text or plots from online articles, assignments of other students (even across sections and previous courses), or website repositories.

What does it mean to be original?

In this course, we care very much about your analysis. It must be original. Original here means two things: 1) the text of the written report must be your own and 2) the exploration that leads to your analysis must be your own. Plagiarism typically refers to the former explicitly, but in this case it also refers to the latter explicitly.

It is well known that for this course we do not care about code. We are not interested in your working out the edge cases in k-nn, or proving your skills with python. While there is some value in implementing algorithms yourselves in general, here we are interested in your grokking the practice of ML itself. That practice is about the interaction of algorithms with data. As such, the vast majority of what you're going to learn in order to master the empirical practice of ML flows from doing your own analysis of the data, hyper parameters, and so on; hence, you are allowed to steal ML code from libraries but are not allowed to steal code written explicitly for this course, particularly those parts of code that automate exploration. You will be tempted to just run said code that has already been overfit to the specific datasets used by that code and will therefore learn very little.

How to cite:

If you are referring to information you got from a third-party source or paraphrasing another author, you need to cite them right where you do so and provide a reference at the end of the document [Col]. Furthermore, "if you use an author's specific word or words, you must place those words within quotation marks and you must credit the source." [Wis]. It is good style to use quotations sparingly. Obviously, you cannot quote other people's assignment and assume that is acceptable. Speaking of acceptable, citing is not a get-out-of-jail-free card. You cannot copy text willy nilly, but cite it all and then claim it's not plagiarism just because you cited it. Too many quotes of more than, say, two sentences will be considered plagiarism and a terminal lack of academic originality.

Your README file will include pointers to any code and libraries you used.

If we catch you...

We report all suspected cases of plagiarism to the Office of Student Integrity. Students who are under investigation are not allowed to drop from the course in question, and the consequences can be severe, ranging from a lowered grade to expulsion from the program.

7 Version Control

- v1.0 - 05/16/2025 - TJL finalized SL Report for Summer 2025 term.

References

[Col] Williams College. *Citing Your Sources: Citing Basics*. URL: `https://libguides.williams.edu/citing`.

[Wis] University of Wisconsin - Madison. *Quoting and Paraphrasing*. URL: https://writing.wisc.edu/handbook/assignments/quotingsources.