Data Lab 2: Breast Cancer Subtyping

ACMS 40950/60876 (Spring 2025)

Due Friday, February 21, 2025 5:00 PM via GitHub

Background and Goals

In recent years, there have been many notable scientific discoveries that have been driven by unsupervised learning. One example is the discovery of breast cancer subtypes, which were identified using hierarchical clustering applied to miRNA array data. Without going into too many biological details, breast cancer is known to be a very heterogeneous disease, and these subtypes are a useful way of defining more homogeneous groups of patients, whose cancers behave similarly. Importantly, by identifying and understanding these breast cancer subtypes, clinicians and scientists have been able to develop targeted and more effective treatment plans for their patients based upon their breast cancer subtype.

In this lab, you will conduct your own unsupervised learning analysis in order to rediscover these breast cancer subtypes. Through this process, you will gain experience with (1) applying various dimension reduction and clustering methods on real-world data and (2) developing intuition for the many different data preprocessing and modeling choices that can substantially impact your findings.

Instructions

You will be engaging with RNASeq gene expression data from breast cancer patients in The Cancer Genome Atlas. (Optional: If you are curious and would like to learn more about RNASeq, here is an introductory video). In this dataset, each row corresponds to a different breast cancer patient, each column corresponds to a different gene, and each (i, j) entry in the data matrix corresponds to the expression of gene j for patient i. Generally, you can think of high gene expression as a gene being turned "on" and producing lots of proteins downstream while low gene expression means that the gene has been turned "off" and does not produce many proteins. In what follows, you will be guided through steps to conduct an unsupervised learning analysis to rediscover breast cancer subtypes.

Note: your RNASeq gene expression dataset is different from the miRNA data used to identify breast cancer subtypes in the seminal research. Consequently, do not be under the impression that you have to obtain the same results as what has already been discovered about breast cancer. When you conduct your analysis, proceed as if you were a clinician/scientist and want to obtain the most *reliable* findings that are best supported by the *given* data.

As in Lab 1, please create a reproducible report with your analysis and critical discussion. Submission details can be found at the end of this document.

Exploratory Data Analysis

1. Please download the gene expression data for this lab from Canvas in the `lab2/` folder. This dataset has 1043 rows/samples and 5000 columns/genes. Conduct a brief exploratory data analysis for this dataset. At a minimum, please (a) plot the distribution of all gene expression values in a histogram or density plot and (b) produce one other EDA plot. What are your main observations about this distribution of gene expression values?

Dimension Reduction

2. Pick at least two different dimension reduction methods that you believe are appropriate for this data, and jot down why you made these choices. Next, apply these dimension reduction methods to the raw gene expression dataset, and plot the results of each dimension reduction method. Are the dimension reduction results similar between the different methods? Justify your answer (e.g., via plots, statistical summaries, other quantitative or qualitative evidence).

3. In the field of genomics, it is common to transform raw gene expression data, denoted by X, using a log-transformation by computing $\tilde{X} = log(X + 1)$. Let us refer to $\tilde{X}$ as the *log-transformed* gene expression data. Using the same dimension reduction methods chosen in problem (2), re-apply them on the log-transformed gene expression data $\tilde{X}$, and plot the results of each dimension reduction method. Are the dimension reduction results similar between the different methods? Are the dimension reduction results applied to X similar to those applied to $\tilde{X}$? Justify your answers (e.g., via plots, statistical summaries, other quantitative or qualitative evidence).

4. Based upon your observations of the dimension reduction results from problem (2) and problem (3) above, what is the "issue" with blindly applying dimension reduction methods to the raw gene expression data X, and how does the log transformation $log(X + 1)$ help to mitigate this issue?

5. *For ACMS 60876 students only*: Are there other data preprocessing choices or steps that could have been taken that might affect our analysis/results?

Clustering

6. Choose at least two different clustering methods that you believe are appropriate for this data, and jot down why you made these choices. For each clustering method chosen here and for each dimension reduction method chosen above in problem (3), apply the clustering method using the dimension-reduced data from problem (3) as the input. Plot the results for $k = 2, 3, 4, 5$, where k denotes the number of clusters.

7. For each clustering method chosen in problem (6), apply the clustering method using the full log-transformed gene expression data ($p = 5000$ features) as the input. Plot the results for $k = 2, 3, 4, 5$, where k denotes the number of clusters.

8. Compare and contrast the clustering results across the different choices of clustering

methods, choices of dimension reduction methods (or lack thereof), and choices of k. Include any visuals, statistical summaries, and other quantitative or qualitative evidence that support your discussion.

9. Repeat problems (6) and (7) using different subsamples of the data (e.g., take subsamples of size 0.8). Comment on the stability of clusters across the different subsamples. Are there particular choices (methods and/or k) that lead to more stable or more unstable clusters?

10. How many clusters k (i.e., breast cancer subtypes) would you say are in the data? Carefully justify your choice of k.

11. *For ACMS 60876 students only*: Discuss the limitations of your clustering analysis. If you were able to conduct a follow-up breast cancer subtyping study, what would you do differently when designing or conducting the study?

12. To conclude this analysis, decide on your best guess of breast cancer subtype labels for two choices of k (i.e., the k you chose in problem (10) and your next best guess for k). Your subtype label guesses could be the clusters from a particular dimension reduction + clustering method, a combination of many methods, or whatever creative scheme that you deem appropriate. The only requirements are that (1) you are able to detail how these breast cancer subtype labels are computed (be sure to include sufficient detail so that a general reader can exactly reproduce your analysis), (2) you can justify your analysis decisions, and (3) your code is reproducible. Please save your best guess of breast cancer subtype labels to `results/best_subtype_labels.csv`. This `.csv` file should be a 1043×2 matrix of cluster labels (or integers) in the same order as the original gene expression data. Rows with the same cluster label (or integer) should correspond to samples that belong in the same cluster. An example of what this `.csv` should look like is provided at https://github.com/tiffanymtang/dsip-s25/blob/main/lab2/results/best_subtype_labels.csv.

    - Bonus points will be awarded to cluster label submissions that most closely align with the true breast cancer subtype labels for the true k, which will be released at the conclusion of this data lab.

Throughout this lab, you will be generating lots of plots. Please use tools such as `patchwork` (in R) or `subplots` (in Python) to combine several plots into a single image whenever appropriate. This will make it easier for you to compare/contrast results across plots.

Submission Details

Please push a folder named `lab2/` to your `dsip` GitHub repository **by 5:00 PM on Friday, February 21, 2025**. I will run an *automated* script that pulls from each of your GitHub repositories promptly at 5:01 PM and attempts to reproduce your report.

The structure of your `lab2/` folder should follow the project structure discussed in class:

```
lab2/
```

```
├── data/
├── R/ (for R users)
├── python/ (for python users)
├── scripts/
├── notebooks/
│   ├── lab2.qmd (or .Rmd or .ipynb)
│   └── lab2.html (or .pdf or other rendered output)
├── renv/ (for R users)
├── renv.lock (for R users)
├── environment.yml (for python users)
└── conda-lock.yml (for python users)
```

The `R/` and/or `python/` folders should contain all functions (and only functions, no scripts) necessary to reproduce your report. The `data/` folder should contain the raw breast cancer data file, but **do not** push the `data/` folder to GitHub. In general, it is not good practice to store data on GitHub due to their restrictions on maximum file size (max: 100MB).

(Optional) You may include other files such as `README.md`, `LICENSE`, and `.gitignore`, which are considered good practice and very common in project workflows. You may also add other folders or rename the folders as you wish, but **do not** rename the `notebooks/` and `lab2.*` files since my automated script will be specifically looking for these file/folder names.

I will attempt to reproduce your report by clicking 'render' in RStudio or 'run all' in Python, so please be sure to include all necessary code in your repository. Keep in mind that my `data/` folder will only contain the raw data file that was initially provided to you.

A Note on Grading + Rubric

You will be graded on both the quality and reproducibility of your analysis.

A detailed rubric can be found on Canvas.

Please recall the course policy regarding collaboration: Collaboration *of ideas* with the instructor and with classmates is encouraged throughout this course, with the following caveats:

- You must write up the final code and text by yourself.
- If you collaborate or use any resources other than course texts, you must explicitly acknowledge your collaborators (e.g., in writing at the end of your report) and cite the resources you used.