Written Report – 6.419x Module 2

Name: NAI CHUN

- **Problem 2 Larger unlabeled subset**

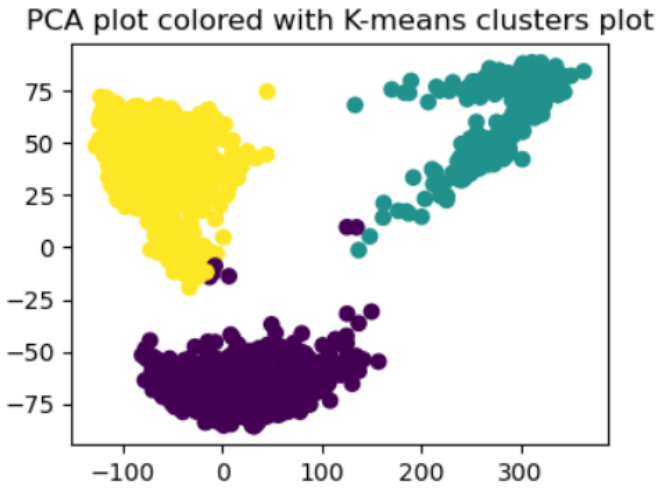
Part 1: Visualization

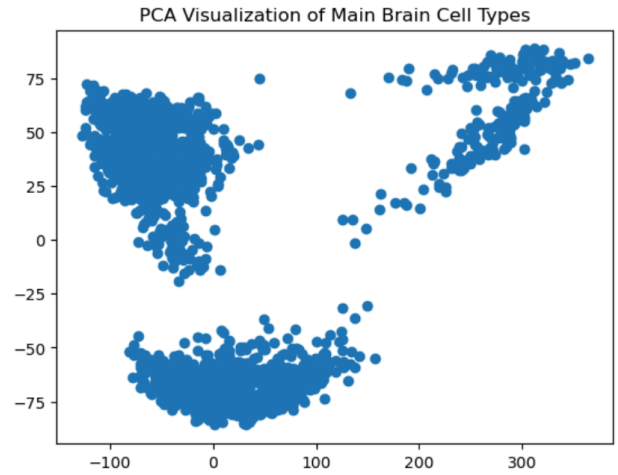
A scientist tells you that cells in the brain are either excitatory neurons, inhibitory neurons, or non-neuronal cells. Cells from each of these three groups serve different functions within the brain. Within each of these three types, there are numerous distinct sub-types that a cell can be, and sub-types of the same larger class can serve similar functions. Your goal is to produce visualizations which show how the scientist's knowledge reflects in the data.

As in Problem 1, we recommend using PCA before running T-SNE or clustering algorithms, for quality and computational reasons.

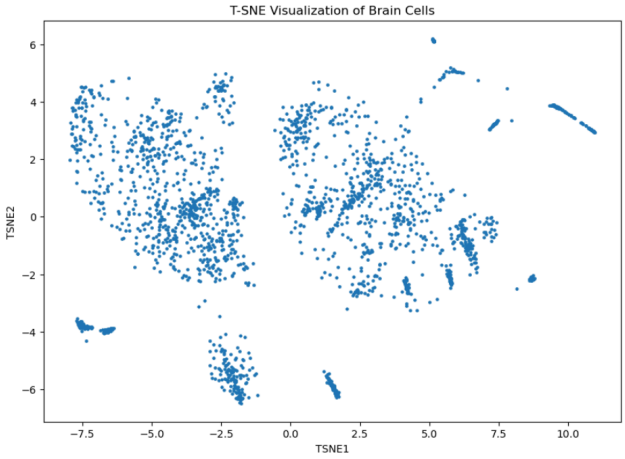
1. (3 points) *Provide at least one visualization which clearly shows the existence of three main brain cell types as described by the scientist, and explain how it shows this. Your visualization should support the idea that cells from different groups can differ greatly.*

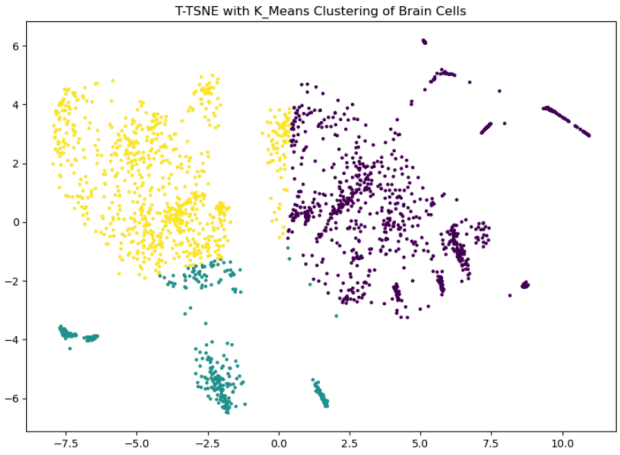
Solution:





The log transformed data produced plus PCA could clearly define three clusters. These three different groups could be seen as either excitatory neurons, inhibitory neurons, or non-neuronal cells.



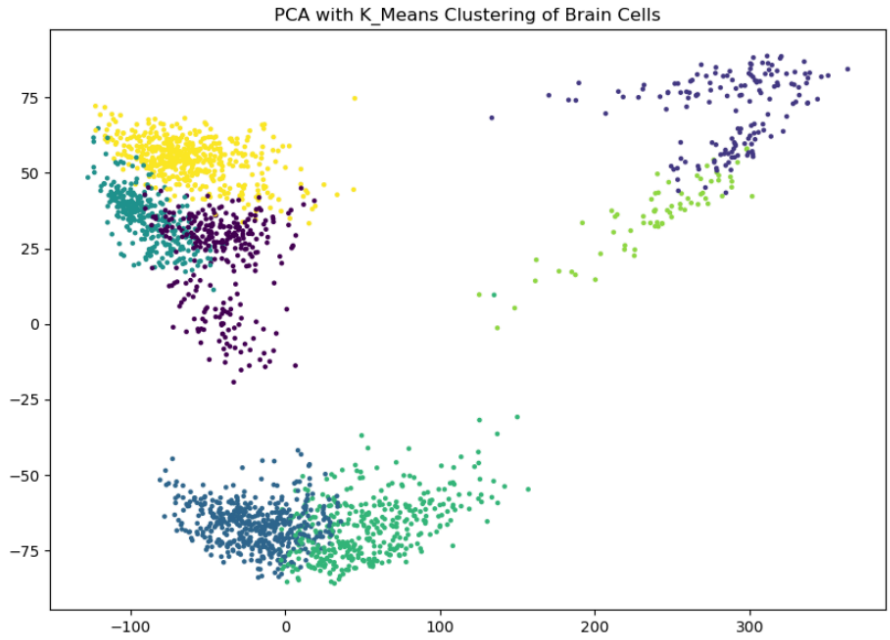


The diagram conducted by T-SNE without PCA also shows three clusters, which clearly correspond to the three main cell types: excitatory neurons, inhibitory neurons, and non-neuronal cells.

2. (4 points) Provide at least one visualization which supports the claim that within each of the three types, there are numerous possible sub-types for a cell. In your visualization, highlight which of the three main types these sub-types belong to. Again, explain how your visualization supports the claim.

Solution:

Use PCA for initial dimensionality reduction, K-Means clustering to identify the main types and sub-types of brain cells.



Subtypes: There are 2-3 potential sub-types represented by K-means clusters in each three visual clusters. Each color in the chart represents a different cell type, indicating significant variations even within the same broad category.

Part 2: Unsupervised Feature Selection

Now we attempt to find informative genes which can help us differentiate between cells, using only unlabeled data. A genomics researcher would use specialized, domain-specific tools to select these genes. We will instead take a general approach using logistic regression in conjunction with clustering. Briefly speaking, we will use the p2_unsupervised dataset to cluster the data. Treating those cluster labels as ground truth, we will fit a logistic regression model and use its coefficients to select features. Finally, to evaluate the quality of these features, we will fit another logistic regression model on the training set in p2_evaluation, and run it on the test set in the same folder.

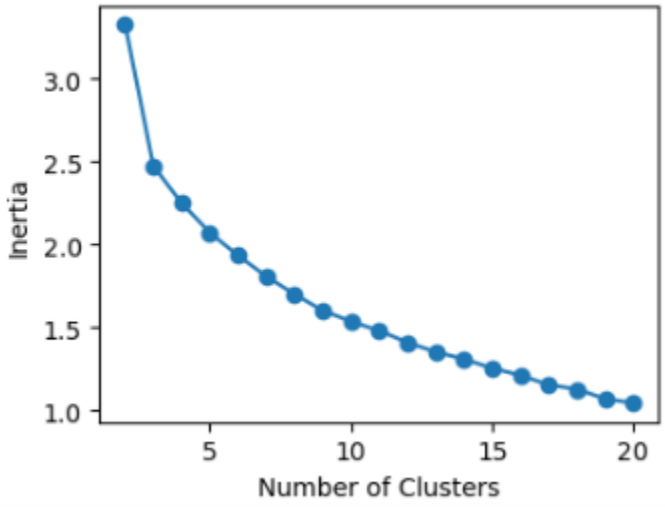
The following steps summarized by instructions from staff.

- ***Data Clustering:*** *Implement K-Means clustering on the* ***p2_unsupervised*** *dataset to categorize the data into clusters.*
- ***Logistic Regression Analysis:*** *Treat the resulting cluster labels as if they were actual ground truth and apply a logistic regression model to pinpoint key features, which in this case are genes.*
- ***Feature Selection:*** *Choose the most significant features by examining the coefficients from the logistic regression model, with a focus on genes that have larger coefficients.*
- ***Quality Assessment:*** *Test the effectiveness of the chosen genes by training a new logistic regression model on the* ***p2_evaluation*** *training dataset and then evaluating its performance on the test dataset within the same directory.*

(4 points) Using your clustering method(s) of choice, find a suitable clustering for the cells. Briefly explain how you chose the number of clusters by appropriate visualizations and/or numerical findings. (to cluster cells into the subcategories instead of categories)

Solution:

Elbow Method is the method I applied to select a number of clusters.
In the plot of the k-Means clustering criterion WGSS versus the number of clusters k, we could find the number of cluster 4 or 5 clusters where the elbow is most prominent.



(6 points) We will now treat your cluster assignments as labels for supervised learning. Fit a logistic regression model to the original data (not principal components), with your clustering as the target labels. Since the data is high-dimensional, make sure to regularize your model using your choice of , or elastic net, and separate the data into training and validation or use cross-validation to select your model. Report your choice of regularization parameter and validation performance.

Multi-class logistic regression: When the underlying data has more than two classes involved, we can adapt Logistic Regression which is usually used for binary classification by one-versus-rest approach. In particular, if we have classes, we train separate binary classification models using logistic regression. Each classifier is trained to determine the probability of a data point belonging to the class . To predict the class for a new point , we run all classifiers on and choose the class with the highest probability, i.e.,

Solution:

Step1: Select the optimal number of clusters as 4 using the Elbow Method, and then Applied K-Means clustering to generate pseudo labels.
Step2: Select top 100 dominant coefficients, and train a logistic regression model on the training set with these selected features.
Step3 : Employed the original high-dimensional dataset and applied a logistic regression model with 5-fold cross-validation and elastic net regularization to it.

For computation issues, in this report, the data applied to these steps was /p2_unsupervised_reduced/X.npy, and got validation accuracy results as: 0.65

(9 points) Select the features with the top 100 corresponding coefficient values (since this is a multi-class model, you can rank the coefficients using the maximum absolute value over classes, or the sum of absolute values). Take the evaluation training data in p2_evaluation and use a subset of the genes consisting of the features you selected. Train a logistic regression classifier on this training data, and evaluate its performance on the evaluation test data. Report your score. (Don't forget to take the log transform before training and testing.)

Compare the obtained score with two baselines: random features (take a random selection of 100 genes), and high-variance features (take the 100 genes with highest variance). Finally, compare the variances of the features you selected with the highest variance features by plotting a histogram of the variances of features selected by both methods.

***Note:** The histogram should show the distribution of the variances of features selected by both methods. You could show the comparison by overlaying both histograms in the same plot.*

Solution:

The accuracy score got from the p2_evaluation with 100 corresponding coefficient values is 0.82
Seems like the top 100 corresponding coefficient values with log_trasformation could well represent the data, even if the data is unsupervised.

```
Indices of top features selected:
[15386 19725 13671  9652 13953 14651  8851 13964 18067  6854  8773 16116
  5008 11124 14832 18016 19085 15455 14099  8544 19341 15458 14199 16834
  9360 14496 14709  7898  1855 19033 13913  9204 14403  5879 14163  5062
 18508 18856  8685 14970 11217  2220 18245  3073 12999 16143 14852 10357
 15676 10113 18253  8186  4451 17564 13078  5651 15428 18493 16022  5625
 16541 18235 19833 18084 15818 14889  8408  6916 14753  9912 18317 15617
  5278 11379 17298 13943  8440  6206 12518 16014  3330  2724  8804 14675
 13283  6349 14278  9659 12557 14702 18853  2000  1942  9076  1662  3504
 11011  3453 18572  9291]
```

The subset of the genes consisting of the features could well-represented the whole image of data.

Accuracy: 0.8194945848375451
Classification Report:

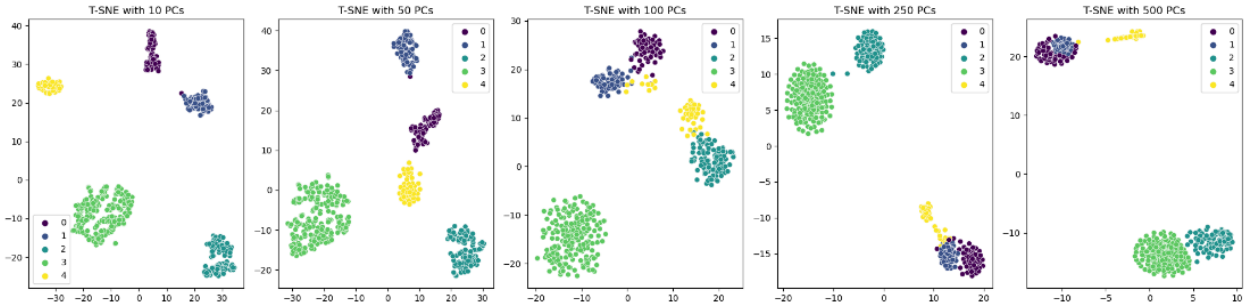
	precision	recall	f1-score	support
0	0.93	0.82	0.87	17
1	0.85	0.89	0.87	19
2	0.92	0.96	0.94	73
3	0.95	1.00	0.97	19
4	0.94	1.00	0.97	17
5	1.00	0.77	0.87	13
6	0.91	0.71	0.80	14
7	0.53	0.62	0.57	37
8	0.85	0.81	0.83	54
9	0.73	0.55	0.63	20
10	1.00	0.87	0.93	30
11	0.78	0.88	0.82	16
12	0.81	0.98	0.89	45
13	0.88	0.44	0.58	16
14	1.00	1.00	1.00	27
15	0.96	1.00	0.98	23
16	0.62	0.67	0.64	27
17	0.84	0.82	0.83	169
18	0.77	0.72	0.74	32
19	0.43	0.56	0.49	18
20	0.86	0.91	0.88	55
21	0.61	0.85	0.71	13
22	0.81	0.93	0.87	14
23	0.94	0.91	0.93	56
24	0.68	0.83	0.75	18
25	0.50	0.17	0.25	12
26	0.70	0.90	0.79	31
27	0.96	0.99	0.97	70
28	0.82	0.60	0.69	15
29	0.38	0.75	0.50	12
30	0.77	0.71	0.74	14
31	0.89	0.62	0.73	13
32	0.93	0.59	0.72	22
33	0.67	0.70	0.68	20
34	0.40	0.20	0.27	10
35	0.91	0.83	0.87	47
accuracy			0.82	1108
macro avg	0.79	0.77	0.77	1108
weighted avg	0.83	0.82	0.82	1108

Top 100 variances could captures most of the information, in contrast, if the variables are randomly selected then the accuracy drops.

- **Problem 3 Influence of Hyper-parameters**

The hyper-parameter choices used in data analysis techniques can have a large impact on the inferences made. As you may have encountered, finding the best choice of parameter such as perplexity in T-SNE or the number of clusters can be an ambiguous problem. We will now investigate the sensitivity of your results to changes in these hyper-parameters, with the goal of understanding how your conclusions may vary depending on these choices.
(3 points) When we created the T-SNE plot in Problem 1, we ran T-SNE on the top 50 PC's of the data. But we could have easily chosen a different number of PC's to represent the data. Run T-SNE using 10, 50, 100, 250, and 500 PC's, and plot the resulting visualization for each. What do you observe as you increase the number of PC's used?

Solution:



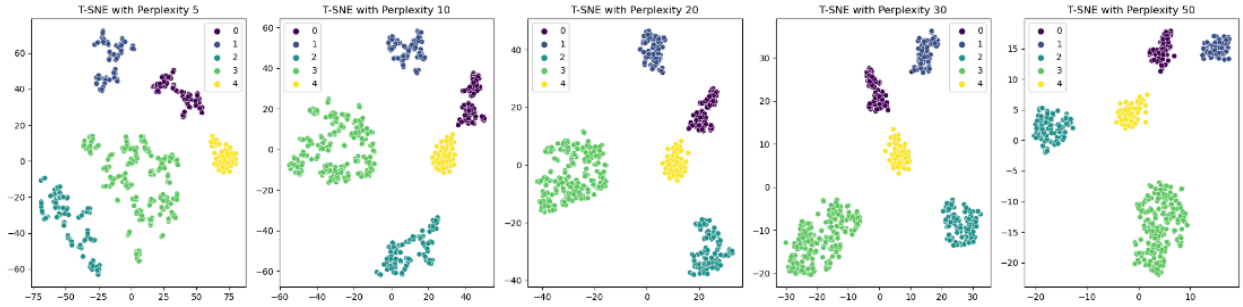
10 PCs, the clusters could lack clear distinction or significance.
50 PCs, capturing more data variance, and makes the groups more clearly separated.
100 PCs, detection of subtle patterns makes the sub-groups be captured.
250 PCs, over this 250, the significant variances are captured.
500 PCs, less improvement

Maybe 50 PCs has already well-represented the data.

(13 points) Pick three hyper-parameters below (the 3 is the total number that a report needs to analyze. It can take a) 2 from A, 1 from B, or b) 1 from A, 2 from B.) and analyze how changing the hyper-parameters affect the conclusions that can be drawn from the data. ***Please choose at least one hyper-parameter from each of the two categories (visualization and clustering/feature selection)****. At minimum, evaluate the hyper-parameters individually, but you may also evaluate* ***how joint changes*** *in the hyper-parameters affect the results. You may use any of the datasets we have given you in this project. For visualization hyper-parameters, you may find it productive to augment your analysis with experiments on synthetic data, though we request that you use real data in at least one demonstration.*

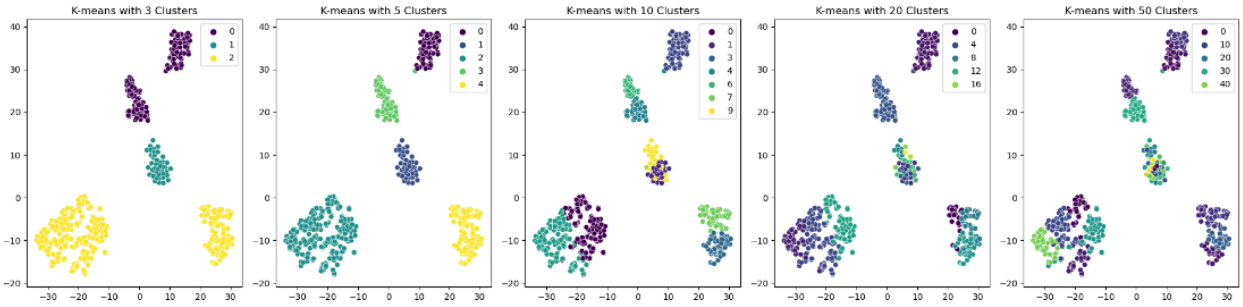
Solution:

1. **Perplexity:** a key factor to capture between global and local structure.**(Visualization)**



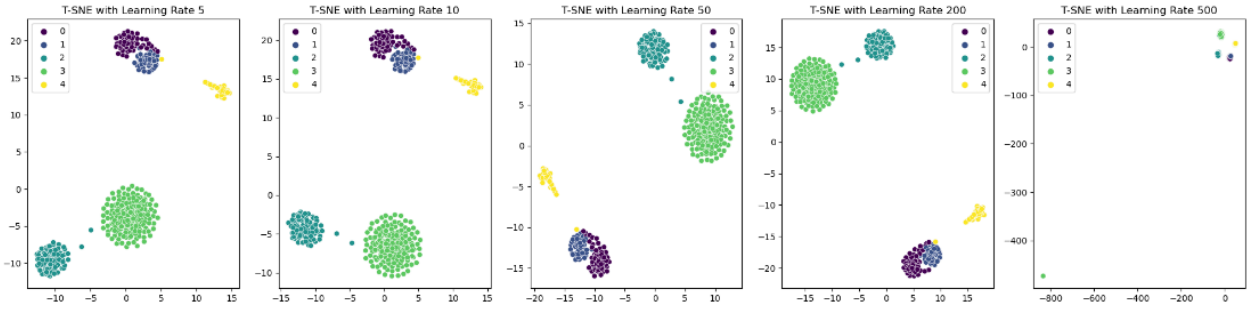
We could estimate the clusters of the points from different values of perplexity.
Perplexity equals to 5, capturing more local structure and information.
Perplexity increases, likes from 5 to 20, could capture a mixture with global and local structure.
When Perplexity increases to 50, it captures mostly global information.

2. **Number of Clusters:** a key factor of labeling data correctly **(Clustering/ Feature Selection)**



3 clusters cannot well represent the data.
5 clusters represent the data perfectly.
When the number of clusters increases, the more noise and local information captures.

3. **T-SNE Learning Rate: A important factors to converge (Visualization)**



Learning rate is highly correlated to the learning rate. If the learning rate is low or too high it will lead to poor convergence, in addition , the visualization is affected deeply by the learning rate.
In the p1 datasets, when the learning rate at both 5, and 500 cannot capture the data effectively.

Reference

[1] Torang, A., Gupta, P. & Klinke, D.J. An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and T helper cell subsets. BMC Bioinformatics 20, 433 (2019). https://doi.org/10.1186/s12859-019-2994-z

[2] B. Gustavii, How to write and illustrate a scientific paper, Cambridge University Press, 2017.

[3] Wikipedia, “Principal component analysis,” Accessed: Sep. 2021. [Online]. Available: https://en.wikipedia.org/wiki/Principal_component_analysis

[4] OpenAI. (2024). ChatGPT (June 26 version) [Large language model]. https://chat.openai.com/chat

[5] How to use t-SNE effectively [Online] Available: https://distill.pub/2016/misread-tsne/