Stat 437 Project 1

Ling Jin (ID: 011880184)

Introduction

In this project, the goal was to explore clustering and classification methods using high-dimensional gene expression data to distinguish different types of cancer. The dataset, obtained from TCGA, includes expression profiles from 801 subjects across five cancer types: BRCA, KIRC, COAD, LUAD, and PRAD. Each subject has expression data for over 20,000 genes.

Working with such a large and complex dataset presents several challenges. First, not all genes are equally informative, some might not be expressed at all in most samples. Second, the sheer number of features compared to the number of samples can introduce noise and make models prone to overfitting. To tackle this, I applied both unsupervised learning, clustering, and supervised learning, classification methods. Specifically, I used K-means clustering, hierarchical clustering, quadratic discriminant analysis, and K-NN. The idea was to see how well these methods could group samples by cancer type and how their performance changes depending on the number of genes used.

Methods and Results

Task A: Clustering Analysis

Step A1: Data Preprocessing

The data was first cleaned by removing genes that had zero expression in at least 300 samples. This left us with a more manageable and meaningful dataset called gexp2. From this, I randomly selected 1000 genes and 30 samples to create a smaller dataset for clustering (gexpProj1), which was then standardized.

Step A2.1: K-Means on 50 Genes

Using 50 randomly selected genes, I applied the gap statistic to estimate the best number of clusters. Interestingly, the gap statistic suggested that the optimal number of clusters was 1, meaning there wasn't a clear clustering structure. Still, I went ahead and used k = 5 (the number of actual cancer types) for K-means clustering. The results showed some grouping, KIRC and PRAD were somewhat separated, but other types like BRCA and LUAD were mixed together. The confusion matrix confirmed that misclassification was common. The cluster visualization also showed lots of overlap.

Step A2.2: Elbow Method

I plotted the total within-cluster sum of squares for k = 1 to 10. The "elbow" wasn't very obvious, but there was a noticeable bend around k = 4 or 5, which matches the number of cancer types. Still, it didn't align perfectly with the gap statistic, and the clustering was far from ideal.

Step A2.3: K-Means on 250 Genes

I repeated the analysis using 250 genes instead of 50. Surprisingly, the gap statistic again suggested k = 1, and the elbow method gave a similar unclear pattern. This showed that just increasing the number of genes doesn't necessarily improve clustering, more features might just mean more noise.

Step A3: Hierarchical Clustering

I tried three linkage methods, which are average, single, and complete on another set of 250 genes. I then cut the average linkage dendrogram to create five clusters, matching the number of cancer types. Some clusters (like KIRC) matched the true labels reasonably well, but others were very mixed. BRCA and LUAD, in particular, were spread across multiple clusters, indicating weak separation.

Task B: Classification Analysis

Step B1: Preprocessing

For the classification task, I focused only on BRCA and LUAD samples. After filtering and randomly picking 1000 genes, I created a standardized dataset called stdgexp2.

Step B2: QDA with 3 Genes

Using only 3 randomly selected genes, I trained a QDA model. I made sure to remove any highly correlated features before fitting the model to avoid issues with the covariance matrix. The classification results were pretty good, most BRCA and LUAD samples were correctly identified, although there were still some errors, especially with LUAD being misclassified as BRCA.

Step B3: QDA with 100 Genes

I expected the model to perform better with more genes, but that wasn't the case. The QDA model completely failed to classify LUAD correctly, all samples were predicted as BRCA. This is likely because estimating a covariance matrix in such a high-dimensional space (with relatively few samples) becomes unstable, leading to overfitting.

Step B4: K-NN with 100 Genes

I repeated the classification using the k-nearest neighbors method with k = 3. This time, the results were excellent. Almost all samples were classified correctly, with very few errors. This shows how non-parametric models like K-NN can handle high-dimensional data better than QDA when you don't have a lot of samples.

Discussion

This project taught me a lot about the challenges of working with high-dimensional gene expression data. One of the biggest takeaways was that using more genes does not always lead to better results. When I increased the number of genes from fifty to two hundred fifty, the clustering actually got worse, likely because the additional features added more noise than meaningful information. Clustering itself was difficult. Methods like K means and hierarchical

clustering struggled to clearly separate the different cancer types, although there were some exceptions like KIRC that showed more distinct patterns. In contrast, supervised learning methods performed much better. K nearest neighbors were especially effective, while quadratic discriminant analysis worked well only when the number of genes was small and became less reliable as the data grew more complex. Overall, this project made it clear that success in analyzing this kind of data depends not just on choosing a method but also on thoughtful data preparation, careful selection of features, and a good understanding of how different models behave with high dimensional data.

Appendix:

The full R code used for analysis is provided in the attached file: **stat437Proj1(Ling Jin).Rmd**