Q1-R

August 15, 2024

1 Q1 - R section

In this section, I use R to perform differential expression analysis for feature selection. In my opinion, differential expression (DE) analysis is the most efficient feature selection method in this case because it not only considers the variance or mutual information of features but also how they are expressed across different classes. Therefore, I believe it will help us identify the best features (DEGs) for classification.

[23]:
```r
# Import needed libraries
library(limma)
library(edgeR)
```

[24]:
```r
# Load Data
normal_counts <- read.csv("train_normal_counts.csv")
meta_data <- read.csv("train_meta_data.csv")
```

[25]:
```r
head(normal_counts)
```

A data.frame: 6 × 134

	DLDR_0036 <dbl>	DLDR_0081 <dbl>	DLDR_0191 <dbl>	DLDR_0188 <dbl>	DLDR_0130 <dbl>	DLDR_0 <dbl>
1	5.820135	6.5462994	6.6040504	6.480745	6.550016	6.569252
2	-1.060061	0.5821648	-0.8650363	-1.083676	-1.222374	0.767254
3	4.388400	3.7520898	4.3514891	4.361634	4.534941	4.150470
4	4.080172	4.6451746	4.0721368	4.313540	4.370763	4.166038
5	2.564430	3.8408991	3.1431376	3.120196	3.512952	3.757010
6	3.552685	3.2010747	4.0374758	1.941859	2.517867	3.253653

[26]:
```r
dim(normal_counts)
```

1. 17396 2. 134

[27]:
```r
head(meta_data)
```

A data.frame: 6 × 1

	Simplified_class <chr>
1	Normal
2	Advanced_fibrosis
3	Normal
4	Normal
5	Non_advanced_Fibrosis
6	Normal

[28]:

```
dim(meta_data)
```

1. 134 2. 1

[29]:

```
labels <- factor(meta_data$Simplified_class)
```

[30]:

```
print(labels)
```

```
 [1] Normal                Advanced_fibrosis     Normal
 [4] Normal                Non_advanced_Fibrosis Normal
 [7] Advanced_fibrosis     Non_advanced_Fibrosis Non_advanced_Fibrosis
[10] Normal                Advanced_fibrosis     Advanced_fibrosis
[13] Advanced_fibrosis     Non_advanced_Fibrosis Advanced_fibrosis
[16] Advanced_fibrosis     Normal                Normal
[19] Normal                Normal                Advanced_fibrosis
[22] Advanced_fibrosis     Non_advanced_Fibrosis Non_advanced_Fibrosis
[25] Advanced_fibrosis     Non_advanced_Fibrosis Normal
[28] Non_advanced_Fibrosis Normal                Advanced_fibrosis
[31] Advanced_fibrosis     Advanced_fibrosis     Advanced_fibrosis
[34] Normal                Normal                Non_advanced_Fibrosis
[37] Non_advanced_Fibrosis Advanced_fibrosis     Advanced_fibrosis
[40] Non_advanced_Fibrosis Advanced_fibrosis     Advanced_fibrosis
[43] Advanced_fibrosis     Advanced_fibrosis     Non_advanced_Fibrosis
[46] Non_advanced_Fibrosis Normal                Advanced_fibrosis
[49] Advanced_fibrosis     Advanced_fibrosis     Non_advanced_Fibrosis
[52] Advanced_fibrosis     Normal                Non_advanced_Fibrosis
[55] Normal                Advanced_fibrosis     Advanced_fibrosis
[58] Non_advanced_Fibrosis Normal                Normal
[61] Normal                Non_advanced_Fibrosis Advanced_fibrosis
[64] Advanced_fibrosis     Normal                Normal
[67] Advanced_fibrosis     Normal                Advanced_fibrosis
[70] Advanced_fibrosis     Advanced_fibrosis     Advanced_fibrosis
[73] Normal                Advanced_fibrosis     Non_advanced_Fibrosis
[76] Normal                Advanced_fibrosis     Advanced_fibrosis
[79] Non_advanced_Fibrosis Non_advanced_Fibrosis Advanced_fibrosis
[82] Normal                Advanced_fibrosis     Non_advanced_Fibrosis
[85] Non_advanced_Fibrosis Advanced_fibrosis     Normal
[88] Non_advanced_Fibrosis Advanced_fibrosis     Non_advanced_Fibrosis
[91] Advanced_fibrosis     Normal                Normal
[94] Advanced_fibrosis     Non_advanced_Fibrosis Non_advanced_Fibrosis
```

```
 [97] Non_advanced_Fibrosis Advanced_fibrosis     Advanced_fibrosis
[100] Advanced_fibrosis     Normal                Normal
[103] Non_advanced_Fibrosis Non_advanced_Fibrosis Normal
[106] Normal                Advanced_fibrosis     Non_advanced_Fibrosis
[109] Non_advanced_Fibrosis Normal                Advanced_fibrosis
[112] Non_advanced_Fibrosis Normal                Normal
[115] Advanced_fibrosis     Normal                Normal
[118] Non_advanced_Fibrosis Advanced_fibrosis     Non_advanced_Fibrosis
[121] Normal                Advanced_fibrosis     Non_advanced_Fibrosis
[124] Non_advanced_Fibrosis Non_advanced_Fibrosis Normal
[127] Advanced_fibrosis     Non_advanced_Fibrosis Advanced_fibrosis
[130] Normal                Normal                Normal
[133] Normal                Normal
Levels: Advanced_fibrosis Non_advanced_Fibrosis Normal
```

Let's perform DE analysis

[31]:

```r
# Create a design matrix
design <- model.matrix(~0 + labels)
colnames(design) <- levels(labels)
```

[32]:

```r
fit <- lmFit(normal_counts, design)
```

[33]:

```r
contrast.matrix <- makeContrasts(
    AdvancedFibrosis_vs_Normal = `Advanced_fibrosis` - Normal,
    Fibrosis_vs_Normal = Non_advanced_Fibrosis - Normal,
    AdvancedFibrosis_vs_Fibrosis = `Advanced_fibrosis` - Non_advanced_Fibrosis,
    levels = design
)

# Apply contrasts to the fit
fit2 <- contrasts.fit(fit, contrast.matrix)

# Empirical Bayes moderation to get p-values
fit2 <- eBayes(fit2)
```

Now, we are going to extract the DEGs for each pair of classes and save them

[34]:

```r
# Get the top DEGs for the Advanced Fibrosis vs Normal comparison
top_genes_adv_vs_norm <- topTable(fit2, coef = "AdvancedFibrosis_vs_Normal",␣
 ↪adjust.method = "BH", number = Inf)

# Get the top DEGs for the Fibrosis vs Normal comparison
top_genes_fib_vs_norm <- topTable(fit2, coef = "Fibrosis_vs_Normal", adjust.
 ↪method = "BH", number = Inf)

# Get the top DEGs for the Advanced Fibrosis vs Fibrosis comparison
```

```
top_genes_adv_vs_fib <- topTable(fit2, coef = "AdvancedFibrosis_vs_Fibrosis",␣
 ↪adjust.method = "BH", number = Inf)

# View the top DEGs
head(top_genes_adv_vs_norm)
head(top_genes_fib_vs_norm)
head(top_genes_adv_vs_fib)
```

A data.frame: 6 × 6

	logFC	AveExpr	t	P.Value	adj.P.Val	B
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
10728	-1.3278914	1.902497	-10.235955	1.733536e-18	3.015659e-14	31.41974
13385	1.0088994	6.051398	9.886439	1.299982e-17	9.020981e-14	29.46392
10694	-1.3132140	3.432308	-9.855214	1.555699e-17	9.020981e-14	29.28958
16113	-3.4233202	-1.392599	-9.734863	3.105946e-17	1.350776e-13	28.61834
16278	-2.8530739	-0.172656	-9.680406	4.245065e-17	1.476943e-13	28.31501
6969	0.4379282	4.207436	9.600224	6.721385e-17	1.948753e-13	27.86885

A data.frame: 6 × 6

	logFC	AveExpr	t	P.Value	adj.P.Val	B
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
13623	0.7946468	5.255822	11.73980	2.831866e-22	4.926315e-18	40.01880
10970	0.6321228	5.361546	11.47202	1.339959e-21	1.021819e-17	38.49831
5442	-1.4931333	2.180779	-11.38733	2.190883e-21	1.021819e-17	38.01728
17075	0.6202635	7.265926	11.37529	2.349550e-21	1.021819e-17	37.94887
4461	0.6782287	6.569435	11.29235	3.802820e-21	1.323077e-17	37.47774
6563	0.6004911	5.545244	11.13147	9.676618e-21	2.805574e-17	36.56387

A data.frame: 6 × 6

	logFC	AveExpr	t	P.Value	adj.P.Val	B
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
16863	1.223013	2.9208664	9.097265	1.182040e-15	2.056276e-11	24.90970
3296	1.594761	1.9955120	8.122841	2.755851e-13	1.021684e-09	19.70546
673	1.485822	1.1568774	8.121636	2.774199e-13	1.021684e-09	19.69912
14913	-1.153029	2.6115748	-8.115173	2.874679e-13	1.021684e-09	19.66515
12060	1.675825	0.1951384	8.111304	2.936548e-13	1.021684e-09	19.64482
3227	1.279852	4.3441375	8.016339	4.947792e-13	1.434530e-09	19.14676

[35]:
```
write.csv(top_genes_adv_vs_norm, "DEGs_AdvancedFibrosis_vs_Normal.csv")
write.csv(top_genes_fib_vs_norm, "DEGs_Fibrosis_vs_Normal.csv")
write.csv(top_genes_adv_vs_fib, "DEGs_AdvancedFibrosis_vs_Fibrosis.csv")
```

We have filtered the top 200 DEGs for each pair. The choice of =200 appears to be optimized based on our greedy search, which has not been included in this notebook.

[36]:
```
filtered_genes_adv_vs_norm <- top_genes_adv_vs_norm[1:200,]
filtered_genes_fib_vs_norm <- top_genes_fib_vs_norm[1:200,]
filtered_genes_adv_vs_fib <- top_genes_adv_vs_fib[1:200,]

# View filtered DEGs
head(filtered_genes_adv_vs_norm)
head(filtered_genes_fib_vs_norm)
head(filtered_genes_adv_vs_fib)
```

A data.frame: 6 × 6

	logFC	AveExpr	t	P.Value	adj.P.Val	B
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
10728	-1.3278914	1.902497	-10.235955	1.733536e-18	3.015659e-14	31.41974
13385	1.0088994	6.051398	9.886439	1.299982e-17	9.020981e-14	29.46392
10694	-1.3132140	3.432308	-9.855214	1.555699e-17	9.020981e-14	29.28958
16113	-3.4233202	-1.392599	-9.734863	3.105946e-17	1.350776e-13	28.61834
16278	-2.8530739	-0.172656	-9.680406	4.245065e-17	1.476943e-13	28.31501
6969	0.4379282	4.207436	9.600224	6.721385e-17	1.948753e-13	27.86885

A data.frame: 6 × 6

	logFC	AveExpr	t	P.Value	adj.P.Val	B
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
13623	0.7946468	5.255822	11.73980	2.831866e-22	4.926315e-18	40.01880
10970	0.6321228	5.361546	11.47202	1.339959e-21	1.021819e-17	38.49831
5442	-1.4931333	2.180779	-11.38733	2.190883e-21	1.021819e-17	38.01728
17075	0.6202635	7.265926	11.37529	2.349550e-21	1.021819e-17	37.94887
4461	0.6782287	6.569435	11.29235	3.802820e-21	1.323077e-17	37.47774
6563	0.6004911	5.545244	11.13147	9.676618e-21	2.805574e-17	36.56387

A data.frame: 6 × 6

	logFC	AveExpr	t	P.Value	adj.P.Val	B
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
16863	1.223013	2.9208664	9.097265	1.182040e-15	2.056276e-11	24.90970
3296	1.594761	1.9955120	8.122841	2.755851e-13	1.021684e-09	19.70546
673	1.485822	1.1568774	8.121636	2.774199e-13	1.021684e-09	19.69912
14913	-1.153029	2.6115748	-8.115173	2.874679e-13	1.021684e-09	19.66515
12060	1.675825	0.1951384	8.111304	2.936548e-13	1.021684e-09	19.64482
3227	1.279852	4.3441375	8.016339	4.947792e-13	1.434530e-09	19.14676

[37]:

```
dim(filtered_genes_adv_vs_norm)
dim(filtered_genes_fib_vs_norm)
dim(filtered_genes_adv_vs_fib)
```

1. 200 2. 6

1. 200 2. 6

1. 200 2. 6

These top 200 DEGs are biologically meaningful in addition to their role in computationally classifying the data. They are likely genes whose expression changes significantly when transitioning from one class to another. These genes are probably among the most correlated with the class labels, though they are not necessarily causal genes. The change in class labels may have a substantial impact on their expression, potentially affecting their associated pathways or other related biological processes.

[38]:

```
genes_adv_vs_norm_names <- rownames(filtered_genes_adv_vs_norm)
genes_fib_vs_norm_names <- rownames(filtered_genes_fib_vs_norm)
genes_adv_vs_fib_names <- rownames(filtered_genes_adv_vs_fib)
```

then we combined the filtered DEGs to create a new feature space

[39]:
```
combined_gene_names <- unique(c(genes_adv_vs_norm_names,
                                genes_fib_vs_norm_names,
                                genes_adv_vs_fib_names))
```

[40]:
```
length(combined_gene_names)
```

527

[41]:
```
common_genes <- intersect(rownames(normal_counts), combined_gene_names)
selected_normal_counts <- normal_counts[common_genes, ]
head(selected_normal_counts)
```

A data.frame: 6 × 134

	DLDR_0036	DLDR_0081	DLDR_0191	DLDR_0188	DLDR_0130	DLDR
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
10	4.5895546	5.4821690	5.01315395	5.0660709	4.6558168	4.2990
57	-0.7190239	0.5821648	-1.15454295	-1.0836762	-0.2894877	-0.0064
265	0.5876374	-2.2251901	-0.07654044	0.1964317	0.2723912	2.5231
275	1.6624052	2.5296974	1.50842206	1.9916119	1.4679420	1.9341
278	5.5299023	5.8462723	5.55358296	5.3592673	5.7745493	4.9626
297	7.5494873	7.7592284	7.61253797	8.0700567	7.4394046	7.6716

[42]:
```
dim(selected_normal_counts)
```

1. 527 2. 134

We extracted a subset from the data based on selected features. Let's save it and continue the analysis in Python Jupyter Notebook

[43]:
```
write.csv(selected_normal_counts, "subset_data.csv")
```