Q3-R

August 15, 2024

1 Q3 - R section

In this section, we use R to perform differential expression analysis for feature selection. We are going to do this analysis for both machines

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.

1.1 First Machine

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

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

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

A data.frame: 6 × 134

	DLDR_0036 <dbl>	DLDR_0081 <dbl>	DLDR_0191 <dbl>	DLDR_0188 <dbl>	DLDR_0130 <dbl>	DLDR_ <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]:
```
dim(normal_counts)
```

1. 17396 2. 134

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

A data.frame: 6 × 2

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

[28]:
```
labels <- factor(meta_data$class)
```

[29]:
```
print(labels)
```

```
  [1] Normal   Fibrosis Normal   Normal   Fibrosis Normal   Fibrosis Fibrosis
  [9] Fibrosis Normal   Fibrosis Fibrosis Fibrosis Fibrosis Fibrosis Fibrosis
 [17] Normal   Normal   Normal   Normal   Fibrosis Fibrosis Fibrosis Fibrosis
 [25] Fibrosis Fibrosis Normal   Fibrosis Normal   Fibrosis Fibrosis Fibrosis
 [33] Fibrosis Normal   Normal   Fibrosis Fibrosis Fibrosis Fibrosis Fibrosis
 [41] Fibrosis Fibrosis Fibrosis Fibrosis Fibrosis Fibrosis Normal   Fibrosis
 [49] Fibrosis Fibrosis Fibrosis Fibrosis Normal   Fibrosis Normal   Fibrosis
 [57] Fibrosis Fibrosis Normal   Normal   Normal   Fibrosis Fibrosis Fibrosis
 [65] Normal   Normal   Fibrosis Normal   Fibrosis Fibrosis Fibrosis Fibrosis
 [73] Normal   Fibrosis Fibrosis Normal   Fibrosis Fibrosis Fibrosis Fibrosis
 [81] Fibrosis Normal   Fibrosis Fibrosis Fibrosis Fibrosis Normal   Fibrosis
 [89] Fibrosis Fibrosis Fibrosis Normal   Normal   Fibrosis Fibrosis Fibrosis
 [97] Fibrosis Fibrosis Fibrosis Fibrosis Normal   Normal   Fibrosis Fibrosis
[105] Normal   Normal   Fibrosis Fibrosis Fibrosis Normal   Fibrosis Fibrosis
[113] Normal   Normal   Fibrosis Normal   Normal   Fibrosis Fibrosis Fibrosis
[121] Normal   Fibrosis Fibrosis Fibrosis Fibrosis Normal   Fibrosis Fibrosis
[129] Fibrosis Normal   Normal   Normal   Normal   Normal
Levels: Fibrosis Normal
```

Let's perform DE analysis

[30]:
```
# Create a design matrix
design <- model.matrix(~0 + labels)
colnames(design) <- levels(labels)
```

[31]:
```
fit <- lmFit(normal_counts, design)
```

[32]:
```
contrast.matrix <- makeContrasts(
    Fibrosis_vs_Normal = `Fibrosis` - Normal,
    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 Fibrosis vs Normal pair and save them

[33]:
```
# 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)

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

A data.frame: 6 × 6

	logFC <dbl>	AveExpr <dbl>	t <dbl>	P.Value <dbl>	adj.P.Val <dbl>	B <dbl>
6969	0.4767457	4.207436	11.50608	9.952564e-22	1.731348e-17	38.77566
17075	0.5206691	7.265926	11.08760	1.140214e-20	7.320483e-17	36.39218
4419	0.8110163	7.074491	11.04250	1.482760e-20	7.320483e-17	36.13539
10970	0.5251446	5.361546	11.01330	1.757697e-20	7.320483e-17	35.96911
7725	0.6071783	4.957512	10.97299	2.222858e-20	7.320483e-17	35.73958
1776	0.7526904	4.613934	10.92701	2.905256e-20	7.320483e-17	35.47785

[34]:
```
write.csv(top_genes_fib_vs_norm, "DEGs_Fibrosis_from_Normal.csv")
```

We have filtered the top 300 DEGs for each pair. The choice of =300 appears to be optimized based on our greedy search, which has not been included in this notebook.

[35]:
```
filtered_genes_fib_vs_norm <- top_genes_fib_vs_norm[1:300,]

# View filtered DEGs
head(filtered_genes_fib_vs_norm)
```

A data.frame: 6 × 6

	logFC <dbl>	AveExpr <dbl>	t <dbl>	P.Value <dbl>	adj.P.Val <dbl>	B <dbl>
6969	0.4767457	4.207436	11.50608	9.952564e-22	1.731348e-17	38.77566
17075	0.5206691	7.265926	11.08760	1.140214e-20	7.320483e-17	36.39218
4419	0.8110163	7.074491	11.04250	1.482760e-20	7.320483e-17	36.13539
10970	0.5251446	5.361546	11.01330	1.757697e-20	7.320483e-17	35.96911
7725	0.6071783	4.957512	10.97299	2.222858e-20	7.320483e-17	35.73958
1776	0.7526904	4.613934	10.92701	2.905256e-20	7.320483e-17	35.47785

These top 300 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 Normal to Fibrosis class. These genes are probably among the most correlated with the class labels, though they are not necessarily causal genes. This change in class label may have a substantial impact on their expression, potentially affecting their associated pathways or other related biological processes.

[36]:
```
genes_fib_vs_norm_names <- rownames(filtered_genes_fib_vs_norm)
```

[37]:
```
common_genes <- intersect(rownames(normal_counts), genes_fib_vs_norm_names)
selected_normal_counts <- normal_counts[common_genes, ]
head(selected_normal_counts)
```

A data.frame: 6 × 134

	DLDR_0036 <dbl>	DLDR_0081 <dbl>	DLDR_0191 <dbl>	DLDR_0188 <dbl>	DLDR_0130 <dbl>	DLDR <dbl>
10	4.5895546	5.48216905	5.013154	5.0660709	4.655817	4.2990
17	5.4498461	5.62655895	5.668470	5.8249308	5.645748	5.1724
67	5.5138387	6.02747535	6.073207	5.8621856	5.755947	4.9978
278	5.5299023	5.84627227	5.553583	5.3592673	5.774549	4.9626
301	5.7134081	6.17141469	5.919896	6.0819486	5.670238	5.1750
310	0.3184508	0.09673801	1.033084	0.3568963	0.362589	1.1890

[38]:
```
dim(selected_normal_counts)
```

1. 300 2. 134

We extracted a subset from the data based on selected features. Let's save it and continue with the second machine.

[39]:
```
write.csv(selected_normal_counts, "subset_data1.csv")
write.csv(meta_data, "meta_data1.csv")
```

1.2 Second Machine

[40]:
```
meta_data <- subset(meta_data, Simplified_class != 'Normal')
```

[41]:
```
head(meta_data)
```

A data.frame: 6 × 2

	Simplified_class <chr>	class <chr>
2	Advanced_fibrosis	Fibrosis
5	Non_advanced_Fibrosis	Fibrosis
7	Advanced_fibrosis	Fibrosis
8	Non_advanced_Fibrosis	Fibrosis
9	Non_advanced_Fibrosis	Fibrosis
11	Advanced_fibrosis	Fibrosis

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

1. 90 2. 2

[43]:
```
rownames(meta_data)
```

1. '2' 2. '5' 3. '7' 4. '8' 5. '9' 6. '11' 7. '12' 8. '13' 9. '14' 10. '15' 11. '16' 12. '21' 13. '22' 14. '23' 15. '24' 16. '25' 17. '26' 18. '28' 19. '30' 20. '31' 21. '32' 22. '33' 23. '36' 24. '37' 25. '38' 26. '39' 27. '40' 28. '41' 29. '42' 30. '43' 31. '44' 32. '45' 33. '46' 34. '48' 35. '49' 36. '50' 37. '51' 38. '52' 39. '54' 40. '56' 41. '57' 42. '58' 43. '62' 44. '63' 45. '64' 46. '67' 47. '69' 48. '70' 49. '71' 50. '72' 51. '74' 52. '75' 53. '77' 54. '78' 55. '79' 56. '80' 57. '81' 58. '83' 59. '84' 60. '85' 61. '86' 62. '88' 63. '89' 64. '90' 65. '91' 66. '94' 67. '95' 68. '96' 69. '97' 70. '98' 71. '99' 72. '100' 73. '103' 74. '104'

75. '107' 76. '108' 77. '109' 78. '111' 79. '112' 80. '115' 81. '118' 82. '119' 83. '120' 84. '122' 85. '123' 86. '124' 87. '125' 88. '127' 89. '128' 90. '129'

[44]:
```
fibrosis_normal_counts <- normal_counts[as.integer(rownames(meta_data))]
```

[45]:
```
head(fibrosis_normal_counts)
```

A data.frame: 6 × 90

	DLDR_0081 <dbl>	DLDR_0130 <dbl>	DLDR_0079 <dbl>	DLDR_0131 <dbl>	DLDR_0135 <dbl>	DLDR_00 <dbl>
1	6.5462994	6.550016	5.9652146	6.436726	6.3378589	6.432976
2	0.5821648	-1.222374	-0.2056643	1.225274	0.3287041	-1.235636
3	3.7520898	4.534941	3.1093774	4.104779	4.6099902	4.061103
4	4.6451746	4.370763	4.8092861	4.411547	4.3253607	4.620430
5	3.8408991	3.512952	3.9527651	3.099743	3.3085263	3.682309
6	3.2010747	2.517867	2.7857226	3.357725	1.9136666	1.506867

[46]:
```
dim(fibrosis_normal_counts)
```

1. 17396 2. 90

[47]:
```
labels <- factor(meta_data$Simplified_class)
```

[48]:
```
print(labels)
```

```
 [1] Advanced_fibrosis     Non_advanced_Fibrosis Advanced_fibrosis
 [4] Non_advanced_Fibrosis Non_advanced_Fibrosis Advanced_fibrosis
 [7] Advanced_fibrosis     Advanced_fibrosis     Non_advanced_Fibrosis
[10] Advanced_fibrosis     Advanced_fibrosis     Advanced_fibrosis
[13] Advanced_fibrosis     Non_advanced_Fibrosis Non_advanced_Fibrosis
[16] Advanced_fibrosis     Non_advanced_Fibrosis Non_advanced_Fibrosis
[19] Advanced_fibrosis     Advanced_fibrosis     Advanced_fibrosis
[22] Advanced_fibrosis     Non_advanced_Fibrosis Non_advanced_Fibrosis
[25] Advanced_fibrosis     Advanced_fibrosis     Non_advanced_Fibrosis
[28] Advanced_fibrosis     Advanced_fibrosis     Advanced_fibrosis
[31] Advanced_fibrosis     Non_advanced_Fibrosis Non_advanced_Fibrosis
[34] Advanced_fibrosis     Advanced_fibrosis     Advanced_fibrosis
[37] Non_advanced_Fibrosis Advanced_fibrosis     Non_advanced_Fibrosis
[40] Advanced_fibrosis     Advanced_fibrosis     Non_advanced_Fibrosis
[43] Non_advanced_Fibrosis Advanced_fibrosis     Advanced_fibrosis
[46] Advanced_fibrosis     Advanced_fibrosis     Advanced_fibrosis
[49] Advanced_fibrosis     Advanced_fibrosis     Advanced_fibrosis
[52] Non_advanced_Fibrosis Advanced_fibrosis     Advanced_fibrosis
[55] Non_advanced_Fibrosis Non_advanced_Fibrosis Advanced_fibrosis
[58] Advanced_fibrosis     Non_advanced_Fibrosis Non_advanced_Fibrosis
[61] Advanced_fibrosis     Non_advanced_Fibrosis Advanced_fibrosis
[64] Non_advanced_Fibrosis Advanced_fibrosis     Advanced_fibrosis
[67] Non_advanced_Fibrosis Non_advanced_Fibrosis Non_advanced_Fibrosis
[70] Advanced_fibrosis     Advanced_fibrosis     Advanced_fibrosis
[73] Non_advanced_Fibrosis Non_advanced_Fibrosis Advanced_fibrosis
```

```
[76] Non_advanced_Fibrosis Non_advanced_Fibrosis Advanced_fibrosis
[79] Non_advanced_Fibrosis Advanced_fibrosis     Non_advanced_Fibrosis
[82] Advanced_fibrosis     Non_advanced_Fibrosis Advanced_fibrosis
[85] Non_advanced_Fibrosis Non_advanced_Fibrosis Non_advanced_Fibrosis
[88] Advanced_fibrosis     Non_advanced_Fibrosis Advanced_fibrosis
Levels: Advanced_fibrosis Non_advanced_Fibrosis
```

Let's perform DE analysis

[49]:
```
# Create a design matrix
design <- model.matrix(~0 + labels)
colnames(design) <- levels(labels)
```

[50]:
```
fit <- lmFit(fibrosis_normal_counts, design)
```

[51]:
```
contrast.matrix <- makeContrasts(
    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 Advanced Fibrosis vs Non-Advanced Fibrosis pair and save them

[53]:
```
# 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_fib)
```

A data.frame: 6 × 6

	logFC <dbl>	AveExpr <dbl>	t <dbl>	P.Value <dbl>	adj.P.Val <dbl>	B <dbl>
16863	1.223013	2.9231954	8.604058	2.266513e-13	3.942826e-09	19.90756
4770	0.482585	4.4402232	7.784892	1.119579e-11	4.824356e-08	16.20400
12710	1.181496	-0.2294500	7.774583	1.175495e-11	4.824356e-08	16.15769
673	1.485822	1.1041118	7.739811	1.385410e-11	4.824356e-08	16.00156
3296	1.594761	2.1489644	7.739625	1.386628e-11	4.824356e-08	16.00073
12060	1.675825	-0.0107199	7.562489	3.195875e-11	9.265907e-08	15.20726

[54]:
```
write.csv(top_genes_adv_vs_fib, "DEGs_Advanced_from_NonAdvanced.csv")
```

We have filtered the top 300 DEGs for each pair. The choice of =300 appears to be optimized based on our greedy search, which has not been included in this notebook.

[55]:
```
filtered_genes_adv_vs_fib <- top_genes_adv_vs_fib[1:300,]

# View filtered DEGs
head(filtered_genes_adv_vs_fib)
```

A data.frame: 6 × 6

	logFC <dbl>	AveExpr <dbl>	t <dbl>	P.Value <dbl>	adj.P.Val <dbl>	B <dbl>
16863	1.223013	2.9231954	8.604058	2.266513e-13	3.942826e-09	19.90756
4770	0.482585	4.4402232	7.784892	1.119579e-11	4.824356e-08	16.20400
12710	1.181496	-0.2294500	7.774583	1.175495e-11	4.824356e-08	16.15769
673	1.485822	1.1041118	7.739811	1.385410e-11	4.824356e-08	16.00156
3296	1.594761	2.1489644	7.739625	1.386628e-11	4.824356e-08	16.00073
12060	1.675825	-0.0107199	7.562489	3.195875e-11	9.265907e-08	15.20726

These top 300 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 Non-Advanced Fibrosis to Advanced Fibrosis class. These genes are probably among the most correlated with the class labels, though they are not necessarily causal genes. This change in class label may have a substantial impact on their expression, potentially affecting their associated pathways or other related biological processes.

[56]:
```
genes_adv_vs_fib_names <- rownames(filtered_genes_adv_vs_fib)
```

[57]:
```
common_genes <- intersect(rownames(fibrosis_normal_counts),␣
 ↪genes_adv_vs_fib_names)
selected_normal_counts <- fibrosis_normal_counts[common_genes, ]
head(selected_normal_counts)
```

A data.frame: 6 × 90

	DLDR_0081 <dbl>	DLDR_0130 <dbl>	DLDR_0079 <dbl>	DLDR_0131 <dbl>	DLDR_0135 <dbl>	DLDR_ <dbl>
57	0.5821648	-0.2894877	-1.138550	-4.060128	-2.1042553	0.01812
209	7.0012221	7.2057963	6.750193	7.459017	7.3791575	7.11479
232	7.1692726	6.9989939	7.278919	7.887874	7.7238812	7.41458
275	2.5296974	1.4679420	2.031375	2.048397	0.5280129	1.13001
297	7.7592284	7.4394046	7.481170	7.556880	8.3222192	8.61647
390	8.6335680	8.2483778	8.367261	8.726754	7.2475674	7.84261

[58]:
```
dim(selected_normal_counts)
```

1. 300 2. 90

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

[59]:
```
write.csv(selected_normal_counts, "subset_data2.csv")
write.csv(meta_data, "meta_data2.csv")
```