Assignment I: Lung Cancer Data Analysis **Duddu Sai Meher Karthik**
High-Dimensional Low Sample-Size (HDLSS) Data Roll Number 130101084
EE622: Advanced Machine Learning August 31, 2016

Problem Statement

Build a classifier for cancer vs. no cancer by using HDLSS techniques (such as elastic net). Use the Gene Expression Omnibus (GEO) data series GSE4115, containing data from 192 human subjects, each with 22,283 profiled genes. Each subject can have one of three disease states: cancer, no cancer, or suspected cancer

Proposed Solution Structure and Properties

- Models tried:
  - Binomial Model with Lasso, Ridge, Elasticnet Penalty (logistic)
  - Gaussian Model with Lasso, Ridge, Elasticnet Penalty (linear with threshold)
- Dataset: Split the dataset into 75% training and 25% testing. Eliminated suspected cancer patients (total 5 in number in 192 samples). Eliminated features(genes) absent across all patients, and used median-imputation for partially-missing genes.
- Good machine learning protocol: Used 10-fold cross-validation for hyperparameter selection, and standardization of features before model training
- Tables and figures showing variation of hyperparameters (alpha, lambda regularization parameters) and their impact on accuracy, comparison of different models, which genes were the most important in each technique, which genes were consistently important across techniques have been presented.
- Insightful analysis: Grouping tendency of coefficients in ridge and elasticnet, variation of non-zero coefficients in Lasso penalty are analyzed and observed.
- Clean, legible, commented code

Model 1: Binomial Model (Logistic)

Experiment 1: Using Lasso Penalty

Accuracy Model training and error calculation calculated using misclassification error. Optimal value of regularizer weightage gave a testing error of 29.7%

Cross-validation and Optimal Lambda Cross-validation error shown in Fig.1. Value of Lambda minimizing misclassification error: 0.1673883

Lasso Non-Zero Coefficients Variation in number of non-zero coefficients with lambda is shown in Fig.2. For the optimal value of lambda, 7 non-zero coefficients were observed.

Important Genes This was calculated considering the absolute value of the feature weights in the trained model. Accordingly, CYR61.1, SLC5A1, BRF2, GSDMB.1, ARAP1.1, IL13RA1.3 were observed to be the most important genes, in that order.

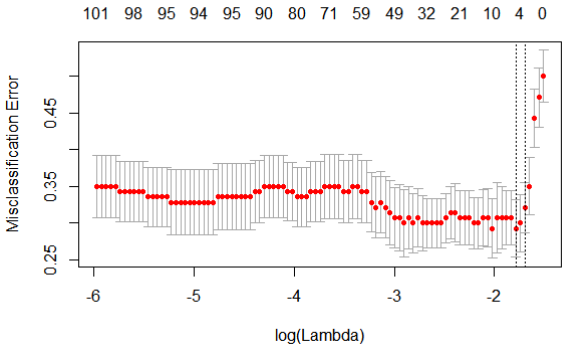

Figure 1: Variation in Validation Miscalculation Error with Lambda

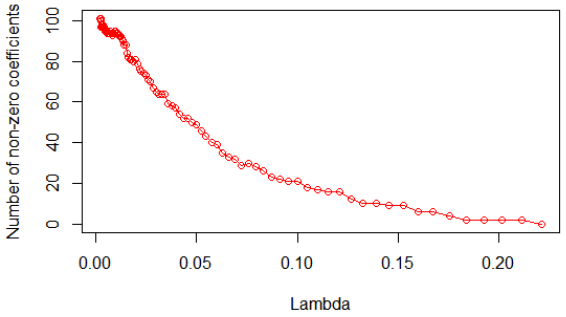

Figure 2: Variation in non-zero coefficients with Lambda in Lasso

Experiment 2: Using Ridge Penalty

Accuracy Model training and error calculation calculated using misclassification error. Optimal value of regularizer weightage gave a testing error of 31.9%

Cross-validation and Optimal Lambda Cross-validation error shown in Fig.3. Value of Lambda minimizing misclassification error: 8.52699

Ridge - Coefficients Grouping Effect Grouping of coefficients (after suitable scaling) when using ridge penalty may be seen in the clustered output shown in Fig.4. Clustering was performed using DBSCAN. For the optimal value of lambda, 3 clusters with 3 noise points were observed (with an epsilon of 0.25).

Important Genes This was calculated considering the absolute value of the feature weights in the trained model. Accordingly, "SLC5A1", "CD74", "VAPA", "BRF2", "PCSK5.1", "ACOX2", "TLE3", "SOX9.1","CYR61.1", "TAOK3.1" were observed to be the most important genes, in that order.

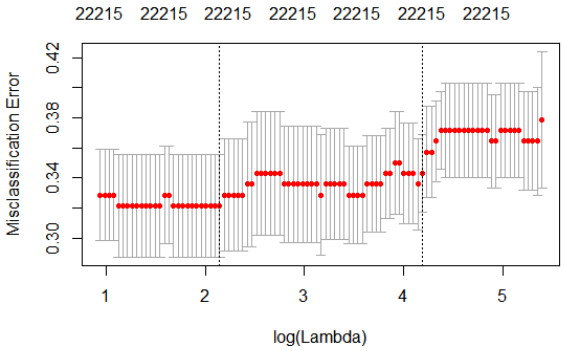

Figure 3: Variation in Validation Miscalculation Error with Lambda

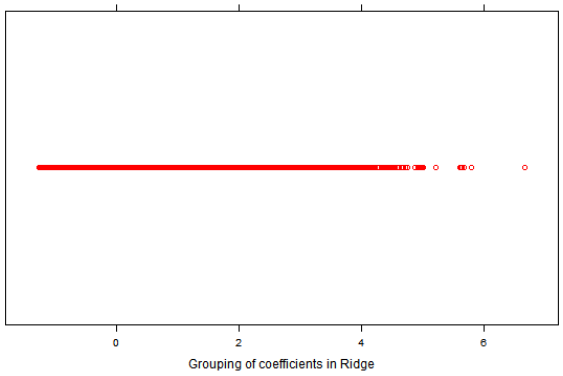

Figure 4: Grouping of coefficients in Ridge Penalty

Experiment 3: Using Elasticnet Penalty

Accuracy Model training and error calculation calculated using mean-square error. Optimal value of regularizer weightage gave a testing error of 23.4%

Cross-validation and Optimal Lambda Cross-validation (10-fold) error for alpha (choosing error-minimizing lambda) is shown in Fig.5. Value of alpha chosen is 0.56.

Cross-validation error of lambda is shown in Fig.6. Value of Lambda minimizing misclassification error: 0.01264153

Elastic-net Non-Zero Coefficients Variation in number of non-zero coefficients with lambda (for optimal alpha) is shown in Fig.7. For the optimal value of lambda, 930 non-zero coefficients were observed.

Ridge - Coefficients Grouping Effect Grouping of coefficients (after suitable scaling) when using ridge penalty may be seen in the clustered output shown in Fig.8. Clustering was performed using DBSCAN. For the optimal value of lambda, 1 clusters with 23 noise points were observed (with an epsilon of 0.30).

Important Genes This was calculated considering the absolute value of the feature weights in the trained model. Accordingly, "CYR61.1", "SLC5A1", "BRF2", "GSDMB.1", "ARAP1.1",

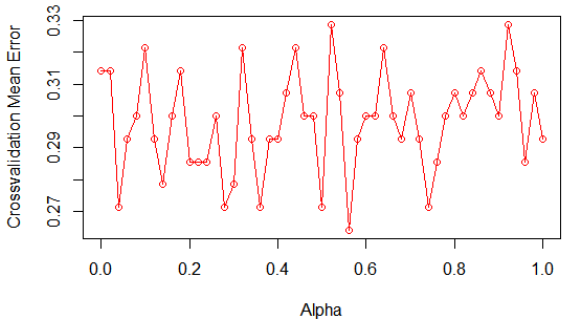

Figure 5: Variation in Validation Mean-Square Error with Alpha (for optimal lambda)

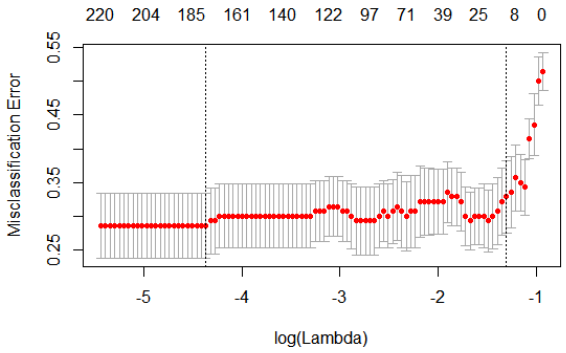

Figure 6: Variation in Validation Mean-Square Error with lambda (for optimal alpha)

"IL13RA1.3", "RFC2", "RPL6", "RNPS1", "ARF1" were observed to be the most important genes, in that order.

Model 2: Gaussian Family (Linear with threshold)

Experiment 4: Using Lasso Penalty

Accuracy Model training and error calculation calculated using mean-sqaure error. Optimal value of regularizer weightage gave a testing error of 22.3%

Cross-validation and Optimal Lambda Cross-validation error shown in Fig.9. Value of Lambda minimizing misclassification error: 0.1003466

Lasso Non-Zero Coefficients Variation in number of non-zero coefficients with lambda is shown in Fig.10. For the optimal value of lambda, 22 non-zero coefficients were observed.

Important Genes This was calculated considering the absolute value of the feature weights in the trained model. Accordingly, "SLC5A1", "CYR61.1", "BRF2", "VAPA", "GSDMB.1",

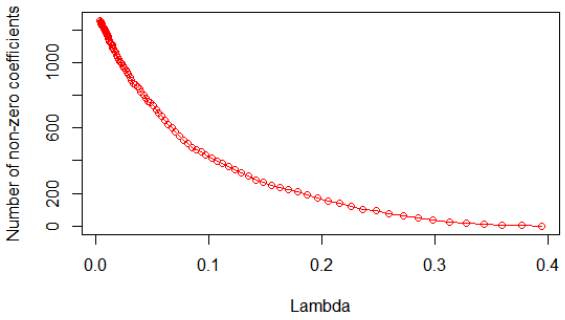

Figure 7: Variation in non-zero coefficients with Lambda in Elasticnet (for optimal alpha)

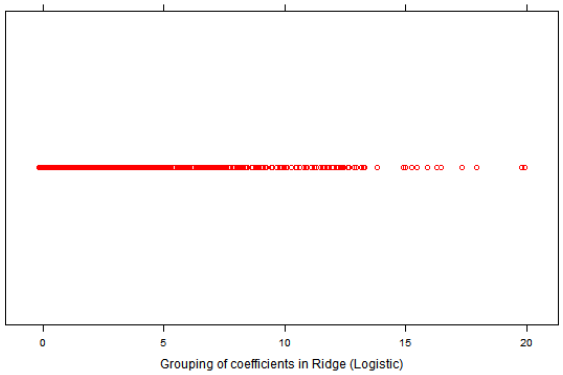

Figure 8: Grouping of coefficients in Ridge Penalty (Linear Elasticnet)

"HUWE1", "NNT.1", "TMEM147-AS1", "HSPA1L", "ARAP1.1" were observed to be the 10 most important genes, in that order.

Experiment 5: Using Ridge Penalty

Accuracy Model training and error calculation calculated using mean-square error. Optimal value of regularizer weightage gave a testing error of 27.65%

Cross-validation and Optimal Lambda Cross-validation error shown in Fig.11. Value of Lambda minimizing misclassification error: 11.80891

Ridge - Coefficients Grouping Effect Grouping of coefficients (after suitable scaling) when using ridge penalty may be seen in the clustered output shown in Fig.12. Clustering was performed using DBSCAN. For the optimal value of lambda, 2 clusters with 7 noise points were observed (with an epsilon of 0.20).

Important Genes This was calculated considering the absolute value of the feature weights in the trained model. Accordingly, "SLC5A1", "CD74", "VAPA", "BRF2", "TSFM", "CYR61.1", "CCDC81", "MTCH1", "SLC39A14", "TMEM147-AS1" were observed to be the 10 most important genes, in that order.

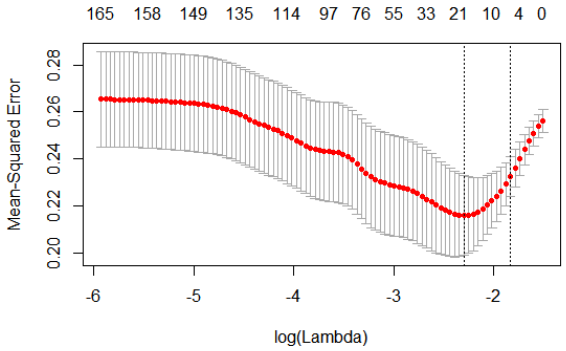

Figure 9: Variation in Validation Mean-Square Error with Lambda

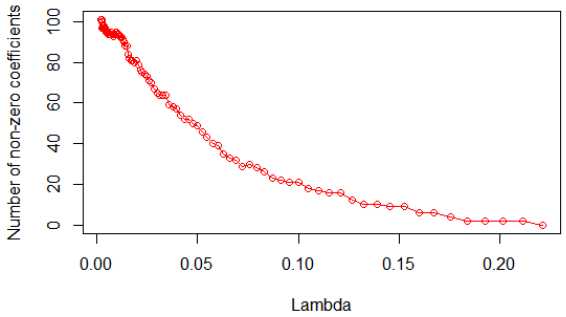

Figure 10: Variation in non-zero coefficients with Lambda in Lasso

Experiment 6: Using Elasticnet Penalty

Accuracy Model training and error calculation calculated using mean-square error. Optimal value of regularizer weightage gave a testing error of 21.9%

Cross-validation and Optimal Lambda Cross-validation (10-fold) error for alpha (choosing error-minimizing lambda) is shown in Fig.**??**. Value of alpha chosen is 0.02.

Cross-validation error of lambda is shown in Fig.**??**. Value of Lambda minimizing misclassification error: 0.817700

Elastic-net Non-Zero Coefficients Variation in number of non-zero coefficients with lambda (for optimal alpha) is shown in Fig.15. For the optimal value of lambda, 930 non-zero coefficients were observed.

Ridge - Coefficients Grouping Effect Grouping of coefficients (after suitable scaling) when using ridge penalty may be seen in the clustered output shown in Fig.16. Clustering was performed using DBSCAN. For the optimal value of lambda, 1 clusters with 23 noise points were observed (with an epsilon of 0.30).

Important Genes This was calculated considering the absolute value of the feature weights in the trained model. Accordingly, "SLC5A1", "BRF2", "VAPA", "ACOX2", "PCSK5.1",

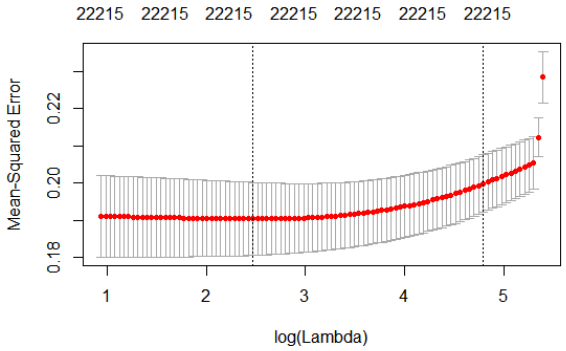

Figure 11: Variation in Validation Mean-Square Error with Lambda

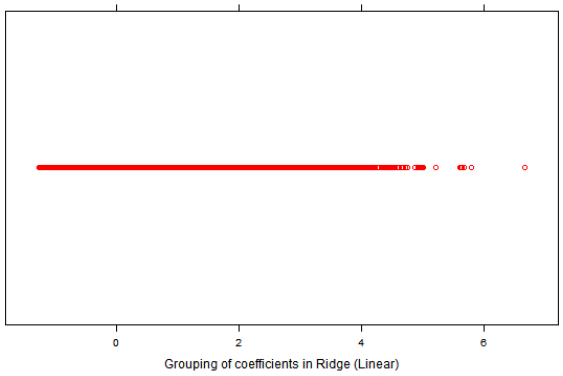

Figure 12: Grouping of coefficients in Ridge Penalty

"TLE3", "MT1G", "TAOK3.1", "SERF1B", "CD74" were observed to be the most important genes, in that order.

Comparison

Model Comparison

Table 1 shows the comparison between the error rates of the different models. It can be seen that the Gaussian family consistently outperforms the binomial family (for the current parameter settings).

Table 1: Error Variation across Models and Penalties

(Error Percentages)	Lasso	Ridge	Elasticnet
Binomial	29.7%	31.9%	23.4%
Gaussian	22.3%	27.65%	21.9%

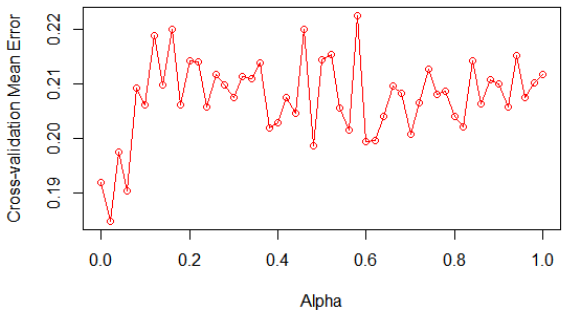

Figure 13: Variation in Validation Mean-Square Error with Alpha (for optimal lambda)

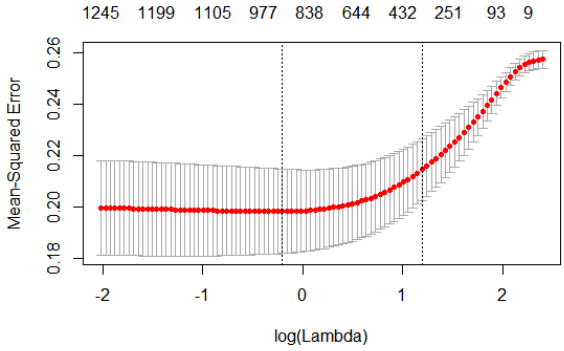

Figure 14: Variation in Validation Mean-Square Error with lambda (for optimal alpha)

Consistently-occurring genes

The genes "SLC5A1", "CYR61.1", "BRF2" are found to occur consistently in the set of important genes across all the models. Other than these, genes such as "PCSK5.1", "VAPA" and "RFC2" are also frequently occuring in the models.

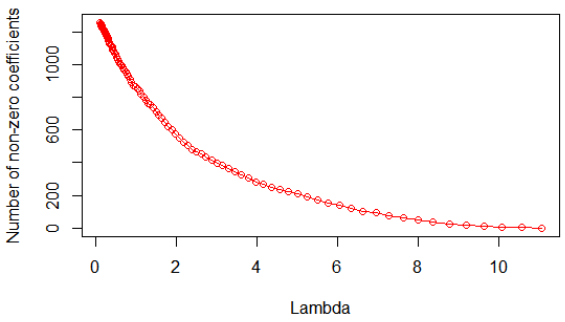

Figure 15: Variation in non-zero coefficients with Lambda in Elasticnet (for optimal alpha)

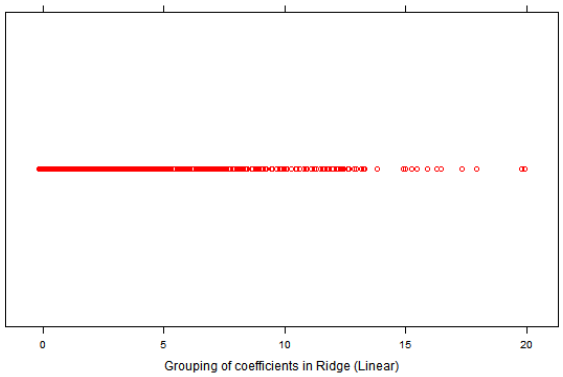

Figure 16: Grouping of coefficients in Ridge Penalty (Linear Elasticnet)