EE 622 Advanced Machine Learning
Assignment 1 : HDLSS analysis of gene expression data

Name: Immidisetti Rakhil
Roll No: 130102026
Department: ECE

Models worked on: Ridge, LASSO, Elastic net, SVM

First, the given data was cleaned and formatted into readable inputs for training the model. The persons with 'Suspect cancer' were removed from the data. The Na values which were all lying in the last columns/features were removed. The data whose values were of type 'character' were converted into a numeric matrix.

The model target values were converted into a two level factor '1' and '2' which are respectively 'cancer' and 'no cancer'. The model is trained on the values of the different genes/features.

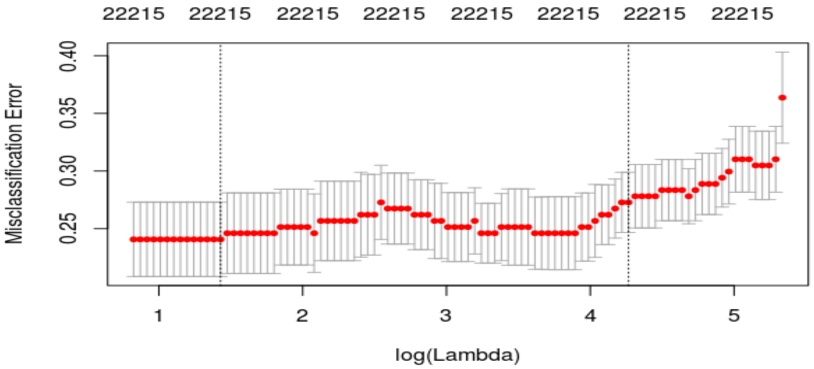
Note- All the models were trained using cross-validation and I am reporting the validation accuracies. There was no manual splitting of train and test data performed.

Packages: Glmnet for Ridge and LASSO models. Caret for comparison between models. E1071 for svm.

RIDGE

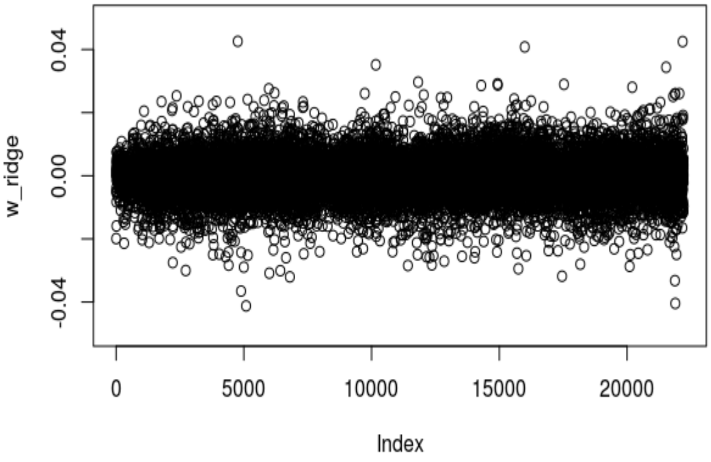
Alpha= 0 ; Best Validation Accuracy= 75.93% ; Lambda.min= 4.17
Lamda.1se= 71.2493

Plot of Error vs log(Lambda)



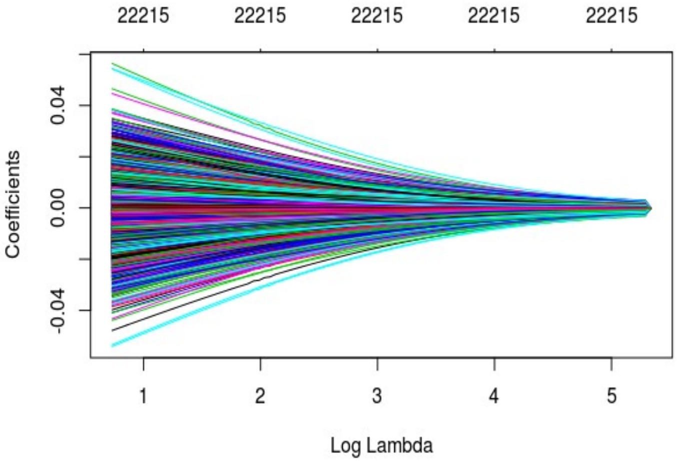
#The red dots represent the average accuracy across all validations. And the line over it represents the variance
#The first vertical line represents the lambda.min for best average accuracy/least error and the next line represents the lambda.1se for least variant accuracy across all validations.
#Accuracy=(1-Misclassification_Error)*100
#The best accuracy is the maximum average accuracy obtained by training. The corresponding value of lambda is the lambda.min
#The upper number in the graph indicates the number of non-zero co-efficients.

Plot of Co-efficients vs Index/feature (for best accuracy)



-Most of the co-efficients in ridge are in the range of -0.1 to +0.1 except zero.
-But one of the co-efficients which is of the first gene/feature was an out-lier whose value is 6.5
-Ridge has a grouping effect that is why all the co-efficients are near to each other and it shrinks all the co-efficients.

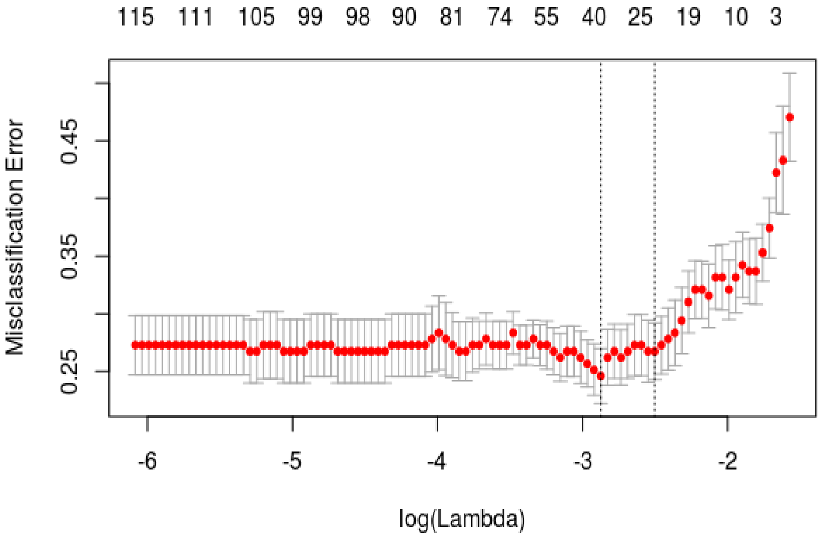
Plot of Co-efficeints across varying lambda



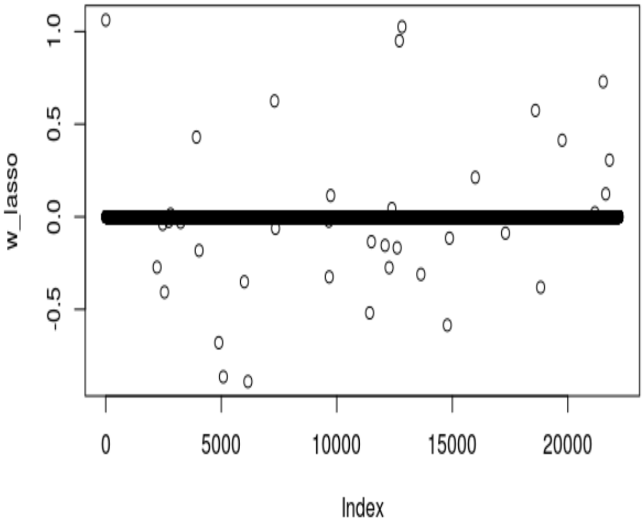
-As we can see ridge has a smooth curve across log(lambda)
-The lambda values are very high and most of the co-efficients are tending to zero.

LASSO

Alpha= 1 ; Best Validation Accuracy= 75.40% ; Lambda.min= 0.056
Lamda.1se= 0.0819
Number of co-efficients whose magnitude is greater than zero= 38
(for best accuracy)

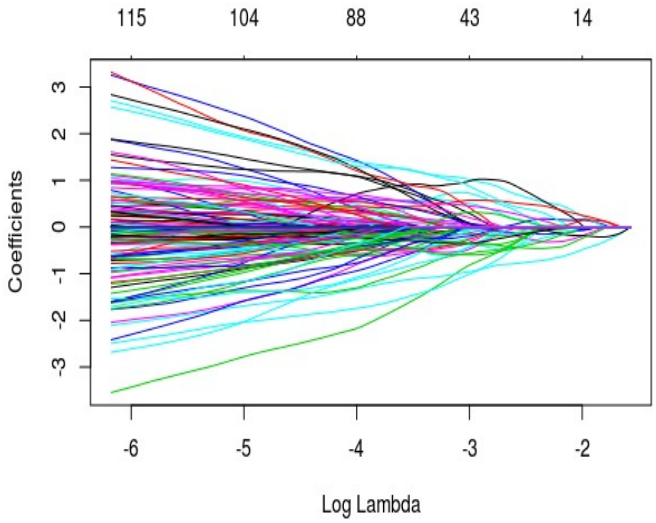


Plot of Co-efficients corresponding to each features
(for best accuracy)
-38 co-efficients are non-zero
-LASSO is a sparse model as it shrinks most of the co-efficients to zero. It does not have a grouping effect as LASSO selects only one among a correlated group and drops the others.



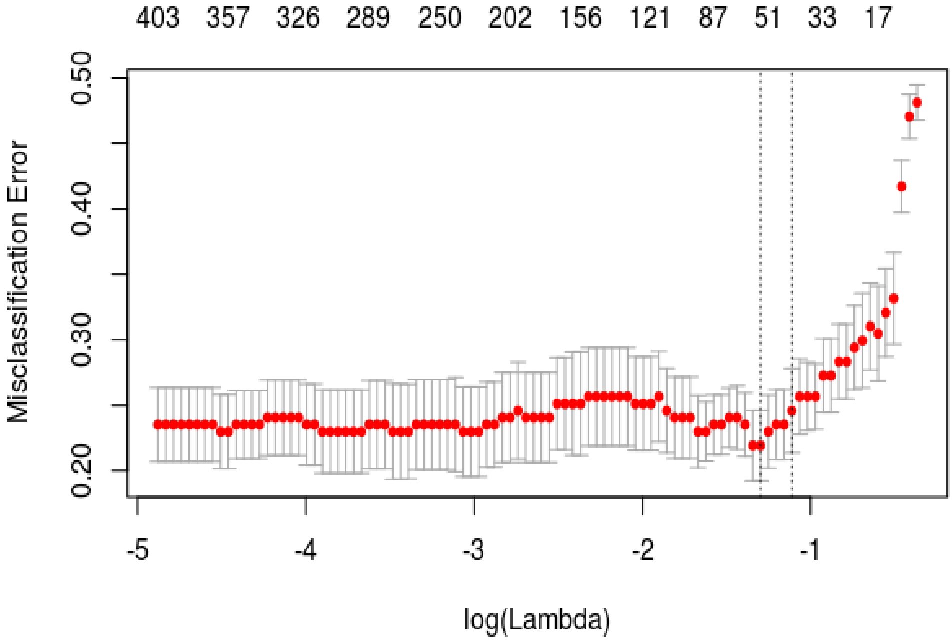
Plot of Co-efficeints across varying lambda

-As we can see the regularization path is not stable as compared to ridge.
-And the co-efficients decrease for increasing penalty factor(lamda) which is a general trend among all the models.

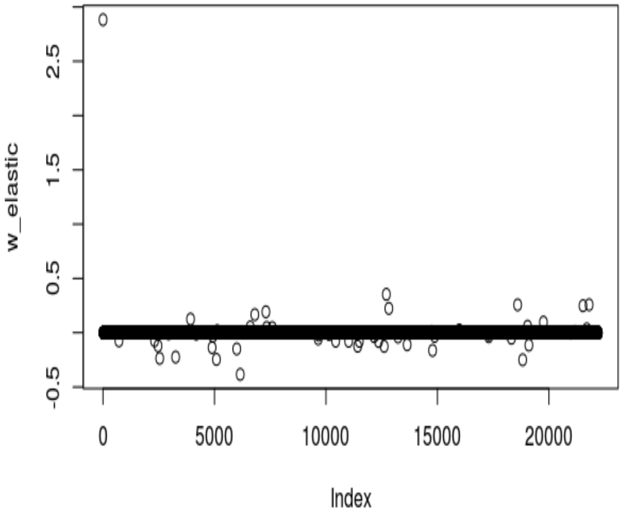


Elastic Net

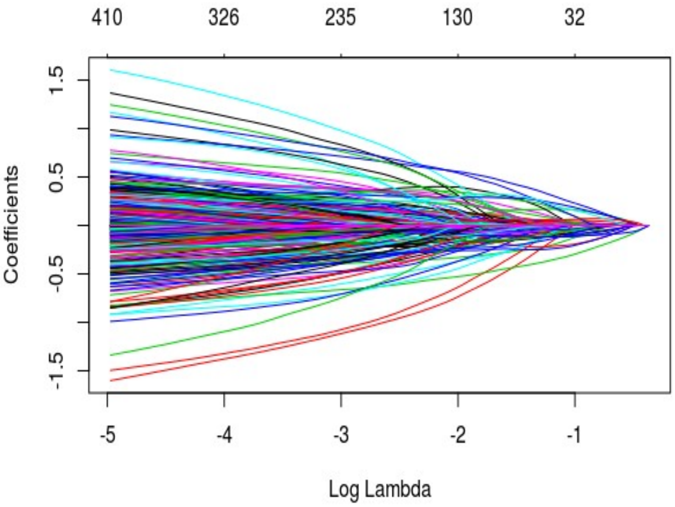
Alpha= 0.3 ; Best Validation Accuracy= 78.07% ; Lambda.min= 0.273
Lamda.1se= 0.328
Number of co-efficients whose magnitude is greater than zero= 63
(for best accuracy)



Plot of Co-efficients corresponding to each features
(for best accuracy)
-***63*** co-efficients are non-zero
-Elastic net which is a combination of L1 and L2 penalty has features of both of them. It has a moderate grouping phenomenon and behaves like LASSO.
-Similar to Ridge it has a out-lier.



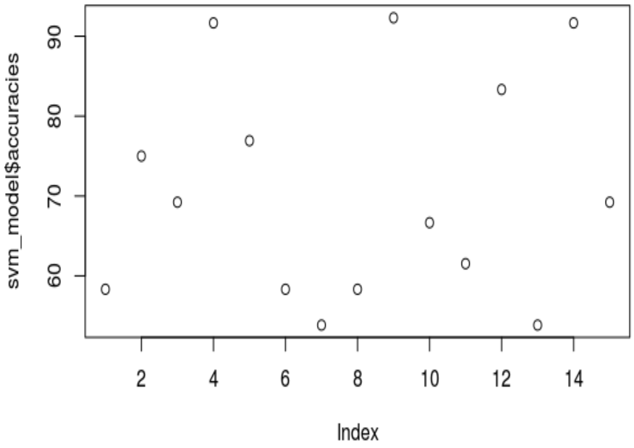
Plot of Co-efficeints across varying lambda
-Elastic net stabilizes the L1 regularization path.
-Many co-efficients are shrunken to zero like LASSO.



Support Vector Machine(SVM)

Validation Accuracy: **70.58824**
Gamma: **4.501463e-05**; SVM-kernel: Radial;
Number of Support Vectors: **175 (83,92)**

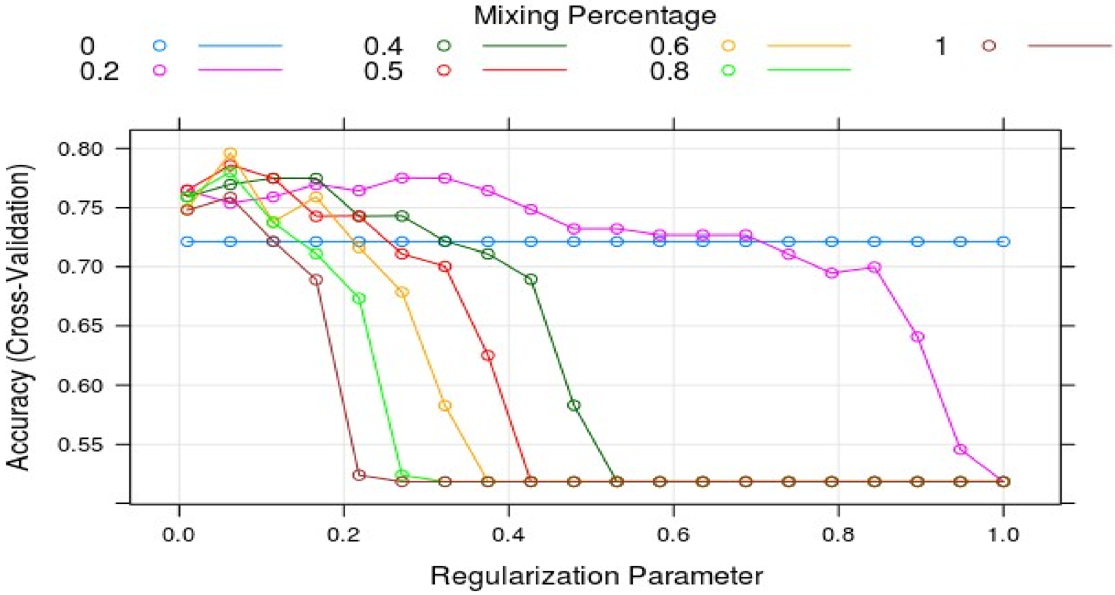
Plot of Accuracy across Validations
-We can observe that the accuracies are random.



Comparison of LASSO, Ridge and Elastic Net

-Used caret and glmnet package. Created a grid of combinations of different alpha and lamda for training a model.

Plot of Accuracy vs lamda



#The Regularization parameter is nothing but the penalty factor(lamda)
#Each color indicates a line for a particular 'Alpha' given above the figure.

-It can be observed that for large values of lambda/penalty factor the accuracy decreases.
-Best Accuracy: 79.619; Alpha= 0.6 ; Lambda= 0.0621
-The top 20% most important variables(out of 22215)
(The numbers correspond to the respective genes in given data)

```
       Overall
V5087   100.00
V4891    97.13
V21528   82.11
V12707   81.17
V6154    76.97
V11421   61.78
V12820   55.11
V15995   54.45
V7304    51.72
V10167   51.47
V12266   50.55
V9733    43.56
V5998    43.32
V3925    42.44
V2217    40.97
V12041   39.87
V17529   39.60
V18597   38.18
V2542    34.73
V19753   34.35
```

-For the models given in the starting, following are the important genes

-Most important genes in Ridge:

```
"SLC16A4"       "TBCCD1"        "HUWE1"         "TRIM5"         "TRIM33"
  "SMCHD1"        "VAMP1"         "PRKAA1"        "FCF1"          "AL050032"      "AF198444"
"ARAP1"         "PPARD"         "ING4"          "TLDC1"
```

-Most important genes in LASSO:

```
"EML2"          "RPL23AP32"     "HUWE1"         "VAMP1"         "RUNDC3A"
 "GSDMB"         "CCDC81"        "AF198444"      "ARAP1"
```

-Most important genes in Elastic net(alpha=0.3):

```
"EML2"          "MED6"          "RPL23AP32"     "HUWE1"         "208082_x_at"
 "LOC101060275"  "VAMP1"         "RUNDC3A"       "TXN"           "GSDMB"         "DCLRE1C"
"CCDC81"        "AF198444"      "222339_x_at"   "ARAP1"
```