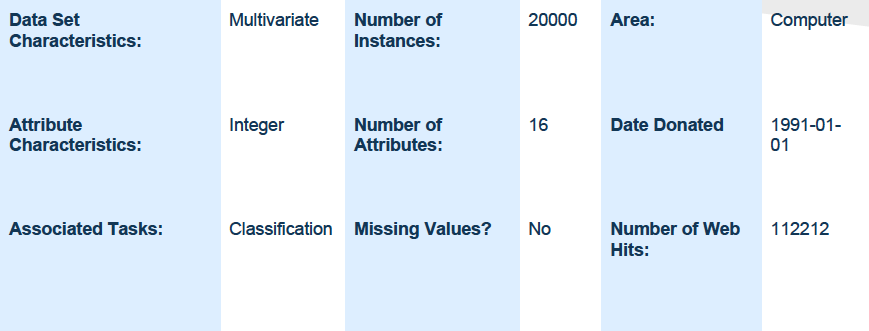
Chathrua J Gunaekara

Predicictive Modeling Project Report

**Letter Image Recognition**

Goal of this project is to compare different classifier systems to learn to correctly guess the letter categories associated with feature of 16 integer attributes extracted from scan images of the letters.

Data set information :



The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet.

Attribute Information:

1. lettr capital letter (26 values from A to Z)

2. x-box horizontal position of box (integer)

3. y-box vertical position of box (integer)

4. width width of box (integer)

5. high height of box (integer)

6. onpix total # on pixels (integer)

7. x-bar mean x of on pixels in box (integer)

8. y-bar mean y of on pixels in box (integer)

9. x2bar mean x variance (integer)

10. y2bar mean y variance (integer)

11. xybar mean x y correlation (integer)

12. x2ybr mean of x \* x \* y (integer)

13. xy2br mean of x \* y \* y (integer)

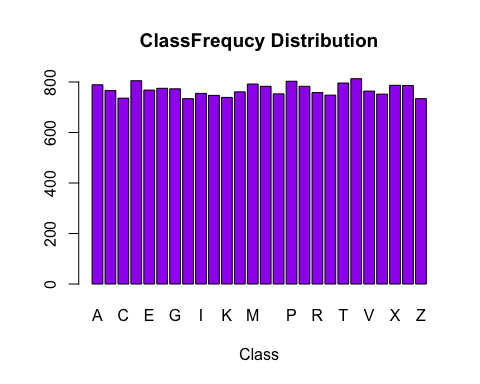
14. x-ege mean edge count left to right (integer)

15. xegvy correlation of x-ege with y (integer)

16. y-ege mean edge count bottom to top (integer)

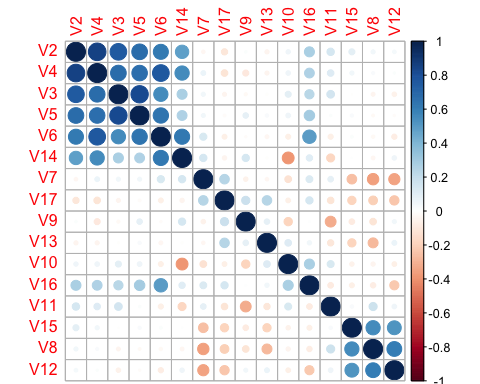
17. yegvx correlation of y-ege with x (integer)

After loading the data, class distributions were observed to see any class imbalances. As the figure below shows class frequency distribution is uniform.



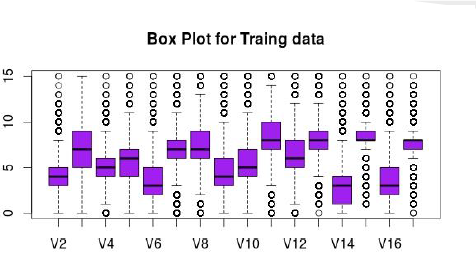
Preprocessing the data set:

Correlations between predictors were visualized to find out if there are correlations greater than 0.75. As the figure below shows there are some predicators, which are correlated. These high correlated predictors were removed using the caret packages, findCorrelation() function.



After removing the high correlated predictors, box plots were used to find out skewness in predictor variables.

The figure below shows boxplots for all the predictors and none of them show too high or low skewness. So transformations of data weren’t necessary.



Data set was split to 75% training and 25% testing using random sampling.

**Linear Models**

**1. Linear Discriminant Analysis**   
   
 15000 samples  
 13 predictor  
 26 classes: 'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y', 'Z'   
   
 Pre-processing: centered, scaled   
 Resampling: Bootstrapped (25 reps)   
   
 Summary of sample sizes: 15000, 15000, 15000, 15000, 15000, 15000, ...   
   
 Resampling results across tuning parameters:  
   
 dimen Accuracy Kappa Accuracy SD Kappa SD   
 1 0.1685297 0.1355126 0.008434755 0.008705092  
 2 0.3497718 0.3236751 0.006712291 0.006945867  
 3 0.4289124 0.4060080 0.015855319 0.016455423  
 4 0.5390830 0.5205938 0.007712305 0.008013558  
 5 0.5833730 0.5666671 0.006934202 0.007205122  
 6 0.6151006 0.5996624 0.008165075 0.008487258  
 7 0.6250795 0.6100264 0.009064244 0.009421349  
 8 0.6726983 0.6595546 0.006922812 0.007192980  
 9 0.6816396 0.6688550 0.006811382 0.007074126  
 10 0.6904524 0.6780292 0.006649750 0.006907253  
 11 0.6867831 0.6742183 0.007103359 0.007377300  
 12 0.6925854 0.6802618 0.007167037 0.007444857  
 13 0.6915845 0.6792207 0.006972897 0.007243935  
   
 Kappa was used to select the optimal model using the largest value.  
 The final value used for the model was dimen = 12.

Confusion Matrix and Statistics  
  
 Overall Statistics  
   
 Accuracy : 0.6882   
  
   
 Kappa : 0.6757

**2. Partial Least Squares**   
   
 15000 samples  
 13 predictor  
 26 classes: 'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y', 'Z'   
   
 Pre-processing: centered, scaled   
 Resampling: Bootstrapped (25 reps)   
   
 Summary of sample sizes: 15000, 15000, 15000, 15000, 15000, 15000, ...   
   
 Resampling results across tuning parameters:  
   
 ncomp Accuracy Kappa Accuracy SD Kappa SD   
 1 0.07420849 0.0372415 0.003271222 0.002700414  
 2 0.17892449 0.1469015 0.010474731 0.010607809  
   
 Kappa was used to select the optimal model using the largest value.  
 The final value used for the model was ncomp = 2.

Overall Statistics  
   
 Accuracy : 0.1908

Kappa : 0.1571

**3. Penalized Models**

glmnet

15000 samples

13 predictor

26 classes: 'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y', 'Z'

Pre-processing: centered, scaled   
Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 15000, 15000, 15000, 15000, 15000, 15000, ...

Resampling results across tuning parameters:

alpha lambda Accuracy Kappa Accuracy SD Kappa SD

0.0 0.1000000 0.5768891 0.5600010 0.01298131 0.01342248

0.0 0.1111111 0.5682126 0.5509860 0.01323037 0.01367192

0.0 0.1222222 0.5601518 0.5426123 0.01334577 0.01378078

0.0 0.1333333 0.5533433 0.5355393 0.01346102 0.01389497

0.0 0.1444444 0.5470622 0.5290141 0.01383478 0.01427529

0.0 0.1555556 0.5410713 0.5227907 0.01451011 0.01496821

0.0 0.1666667 0.5353392 0.5168374 0.01481343 0.01527790

0.0 0.1777778 0.5304621 0.5117712 0.01490745 0.01537096

0.0 0.1888889 0.5251520 0.5062547 0.01485169 0.01530962

0.0 0.2000000 0.5203919 0.5013103 0.01492080 0.01537833

0.1 0.1000000 0.5014833 0.4816663 0.01289188 0.01327896

0.1 0.1111111 0.4852512 0.4648082 0.01358775 0.01398503

0.1 0.1222222 0.4703423 0.4493276 0.01356431 0.01394621

0.1 0.1333333 0.4548711 0.4332666 0.01519515 0.01561544

0.1 0.1444444 0.4402404 0.4180760 0.01596798 0.01640747

0.1 0.1555556 0.4254912 0.4027650 0.01725252 0.01773493

0.1 0.1666667 0.4094354 0.3860944 0.01900821 0.01954999

0.1 0.1777778 0.3940862 0.3701506 0.01849510 0.01903835

0.1 0.1888889 0.3794613 0.3549606 0.02072140 0.02135320

0.1 0.2000000 0.3638619 0.3387574 0.02230321 0.02300903

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were alpha = 0 and lambda = 0.1.

Overall Statistics

Accuracy : 0.5808

95% CI : (0.567, 0.5945)

No Information Rate : 0.0456

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5639

Mcnemar's Test P-Value : NA

**4. Nearest Shrunken Centroids**

15000 samples

13 predictor

26 classes: 'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y', 'Z'

Pre-processing: centered, scaled

Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 15000, 15000, 15000, 15000, 15000, 15000, ...

Resampling results across tuning parameters:

threshold Accuracy Kappa Accuracy SD Kappa SD

0 0.6080019 0.5923025 0.006248540 0.006473857

1 0.6036960 0.5878307 0.006435820 0.006662524

2 0.5948069 0.5785900 0.007077063 0.007325056

3 0.5842482 0.5676138 0.008558737 0.008860115

4 0.5717130 0.5545881 0.010274424 0.010647078

5 0.5579820 0.5403286 0.011095785 0.011497738

6 0.5403542 0.5220250 0.012315319 0.012748859

7 0.5183379 0.4991683 0.012693090 0.013130464

8 0.4929050 0.4727736 0.011473305 0.011865651

9 0.4612455 0.4399211 0.013049754 0.013482382

10 0.4308509 0.4083684 0.014772277 0.015266688

11 0.4022141 0.3786233 0.014943307 0.015434551

12 0.3711531 0.3463274 0.015064321 0.015529815

13 0.3455720 0.3197119 0.014186495 0.014597165

14 0.3228725 0.2960926 0.014237422 0.014671842

15 0.3043471 0.2767924 0.011561100 0.011919913

16 0.2915882 0.2634689 0.013995954 0.014411181

17 0.2846178 0.2561714 0.017144759 0.017666049

Kappa was used to select the optimal model using the largest value.

The final value used for the model was threshold = 0.

Overall Statistics

Accuracy : 0.614

Kappa : 0.5985

Non - Linear models

1. **K- Nearest Neighbour**

15000 samples

13 predictor

26 classes: 'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y', 'Z'

Pre-processing: centered, scaled

Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 15000, 15000, 15000, 15000, 15000, 15000, ...

Resampling results across tuning parameters:

k Accuracy Kappa Accuracy SD Kappa SD

3 0.9188318 0.9155698 0.003479140 0.003618114

5 0.9170265 0.9136917 0.003572591 0.003715878

7 0.9163788 0.9130183 0.003412833 0.003549829

9 0.9150894 0.9116772 0.003584294 0.003727636

11 0.9134054 0.9099255 0.003621076 0.003766167

Kappa was used to select the optimal model using the largest value.

The final value used for the model was k = 3.

Overall Statistics for Testing set:

Accuracy : 0.9544   
 Kappa : 0.9526

Improved KNN. For the problem of nearest neighbor classification, a simpler approach called "leave-out-one" cross-validation can be used, and this is provided by the knn.cv function. Using this technique, the observation itself is ignored when looking for its neighbors.

**2. K- Neareset Neighbout with CV**

myknn<-knn.cv(x, y, k = 3, l = 0, prob = FALSE, use.all = TRUE)  
confusionMatrix(myknn, y)

Overall Statistics for testing set.  
   
 Accuracy : 0.9614   
 Kappa : 0.9599

**3. SVM**

Support Vector Machines with Radial Basis Function Kernel   
   
 15000 samples  
 13 predictor  
 26 classes: 'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y', 'Z'   
   
 Pre-processing: centered, scaled   
 Resampling: Bootstrapped (25 reps)   
   
 Summary of sample sizes: 15000, 15000, 15000, 15000, 15000, 15000, ...   
   
 Resampling results across tuning parameters:  
   
 C Accuracy Kappa Accuracy SD Kappa SD   
 0.25 0.8116716 0.8041039 0.005517341 0.005736462  
 0.50 0.8429431 0.8366312 0.004796872 0.004988122  
 1.00 0.8686856 0.8634081 0.004181792 0.004350780  
 2.00 0.8903813 0.8859756 0.003851433 0.004006828  
 4.00 0.9079643 0.9042654 0.003491611 0.003631531

8.00 0.9224697 0.9193564 0.002909209 0.003025576

16.00 0.9336818 0.9310184 0.002754771 0.002865363

32.00 0.8263113 0.8359756 0.002861032 0.00425156

Tuning parameter 'sigma' was held constant at a value of 0.0223395

Kappa was used to select the optimal model using the largest value.

The final values used for the model were sigma = 0.0223395 and C = 16.

Overall Statistics  
   
 Accuracy : 0.9392   
 Kappa : 0.9368

4. Neural Network   
   
 15000 samples  
 13 predictor  
 26 classes: 'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y', 'Z'   
   
 Pre-processing: centered, scaled   
 Resampling: Bootstrapped (25 reps)   
   
 Summary of sample sizes: 15000, 15000, 15000, 15000, 15000, 15000, ...

Resampling results across tuning parameters:  
   
 size decay Accuracy Kappa Accuracy SD Kappa SD   
 1 0.0 0.1721552 0.13925465 0.012104369 0.012588550  
 1 0.1 0.1645228 0.13141653 0.008313172 0.008761456  
 1 1.0 0.1366458 0.10276125 0.006194946 0.006369236  
 1 2.0 0.1243592 0.09018002 0.008899931 0.008694901  
 2 0.0 0.3392758 0.31291087 0.039539183 0.041060387  
 2 0.1 0.3389576 0.31261560 0.011208979 0.011558657  
 2 1.0 0.2963639 0.26850521 0.011969575 0.012334720  
 2 2.0 0.2724686 0.24372387 0.014093756 0.014460833  
 3 0.0 0.5012276 0.48124893 0.022398048 0.023284106  
 3 0.1 0.4965047 0.47636823 0.008824991 0.009161699  
 3 1.0 0.4581531 0.43656710 0.008548176 0.008836588  
 3 2.0 0.4261495 0.40337037 0.011090971 0.011372202

Kappa was used to select the optimal model using the largest value.  
 The final values used for the model were size = 3 and decay = 0.

Overall Statistics for Testing set  
   
 Accuracy : 0.4882   
 Kappa : 0.4675

**Base on Testing data metrics(Kappa and Accuracy) KNN with CV is chosen to do the predictive model based on the below table**.

|  |  |  |
| --- | --- | --- |
| Model(blue=linear, red=Non linear) | Kappa | Accuracy |
| LDA | 0.68 | 0.67 |
| PLSDA | 0.15 | 0.19 |
| Penalize Models | 0.5962 | 0.5808 |
| Nearest Shrunken Centroids | 0.5921 | 0.6193 |
| KNN | 0.9529 | 0.9544 |
| **KNN-CV** | **0.9599** | **0.9614** |
| SVM | 0.9392 | 0.9396 |
| NeuralNet | 0.4882 | 0.4246 |