STATISTICAL LEARNING

Home Assignment 1

MAY 9, 2022

KRISHNA BHARAT PASUMARTHY
h21pasbh@du.se

Question 1.A

Introduction:

The ANES2016.xlsx data has the survey data collected during the 2016 US prudential election campaign. The dataset has 4271 rows with 18 columns which comprises of the data related to each individual and is uniquely identified by the ID column, their personal characteristics like education, income, etc., columns Trump & Hillary which defines how likely a person is willing to vote for them and Party ID which gives information about how they consider themselves viz., a Democrat, a Republican or an Independent. Note that ID column is a just used for identifying the individuals uniquely, so that column was not considered in any part of the analysis.

Loaded the excel file into R, created a data frame 'df' and analyzed the summary and structure of the dataset. The text documents help us to understand that all the columns except ID are coded to numeric values for computation.

Method:

For this task, I have made a copy of the data frame with name df_trump. The first step is to convert Trump column into a categorical column with the values being only Liberal or conservative based on their levels. As the columns Trump and Hillary contain data in similar levels and structure, converting Hillary also into liberal and conservative will estimate the parameters much accurate than using Hillary column with 7 categories. For example, if there is row which says Trump as conservative and Hillary having value as Liberal, or conservative would be better to correlate than comparing with a value from 1 to 7.

Once transformed both the fields to values 0 (Liberal) or 1 (Conservative), and dropping all the rows having NA, we have 4017 rows left for further processing. Out of these 3062 rows has opposite values when compared between Trump and Hillary which means if Trump is Liberal then Hillary in Conservative and vise versa. Further transformed, 0 and 1 values to Liberal and Conservative for the Trump column as required and converted into a categorical variable. Below we can find the counts of each value in Trump column in the dataset.

Conservative	Liberal
3237	780

Further explored the values in all the fields to identify if they have data which can be invalid or ignorable values which does not contribute for the analysis and drop those rows. Used table() function to explore all the columns to check the unique values in each field. As per the number of occurrences of each numeric value in each field, description of the column provided the text file, we can come to a conclusion that for all the columns having values lesser than 0 can be ignored as they do not seem to add value to estimate the response, except for the columns Partner and SpouseEdu, as these columns has value -1 which has a significant count and the description helps us to understand -1 represents individuals who are married.

Dropping the unnecessary rows based on the above reasoning, we are left with 3756 rows. As part of understanding the influence of personal characters on determining Trump if its Liberal or Conservative, I have tried to analyze in below 2 approaches.

- Ignoring interactions within the features
- Considering Interactions within the features

Discussion:

i. Ignoring Interactions:

Created a generalized linear model to perform a logistic regression, using each column at a time under binomial family, and Trump as response as has only 2 categories. Below table gives the information about the coefficients by combining information of all the coefficients from 16 summaries to make it easy for analyzing.

Coefficients:					
	Estimate	Std.Error	z value	Pr(>|z|)	
Media	-0.08846	0.02005	-4.411	1.03E-05	***
FamSize	0.07031	0.02585	2.72	0.00652	**
Hillary	1.88811	0.08866	21.3	<2e-16	***
Age	-0.00538	0.00236	-2.279	0.0227	*
Partner	0.12687	0.0288	4.405	1.06E-05	***
Education	-0.18899	0.01813	-10.42	<2e-16	***
SpouseEdu	-0.03798	0.00644	-5.899	3.67E-09	***
Employment	0.04254	0.01775	2.396	0.0166	*
Birthplace	0.12517	0.03893	3.215	0.0013	**
GBirth	-0.08908	0.03889	-2.291	0.022	*
Dependent	0.15091	0.03586	4.209	2.57E-05	***
Housing	-0.16024	0.04728	-3.389	0.0007	***
Income	-0.0487	0.00515	-9.451	<2e-16	***
Education2	-0.32292	0.02822	-11.44	<2e-16	***
PartyID	-0.00365	0.04504	-0.081	0.935	
Marital	0.1003	0.02394	4.19	2.79E-05	***

We can clearly observe that all the columns except PartyID have a significant Pr(>|Z|) value lesser than 5 percent when interactions are ignored and simulating column by column.

ii. Considering Interactions:

As we have observed earlier that Hillary column has stronger correlation as 76% of rows have values opposite to each other, we can analyze the coefficients by considering interactions among the features including and excluding Hillary as value in Hillary column will have higher influence on estimate of Trump.

	Pr(>|z|)		Pr(>|z|)	
(Intercept)	0.000527	***	0.83175	
Media	0.238338		0.082552	.
FamSize	0.771957		0.663958	
Age	0.239937		0.238553	
Hillary	<2e-16	***		
Partner	0.05308	.	0.076197	.
Education	0.665844		0.548452	
SpouseEdu	0.018493	*	0.013709	*
Employment	0.893427		0.820216	
Birthplace	0.110543		0.004146	**
GBirth	0.041165	*	0.07205	.
Dependent	0.057252	.	0.015579	*
Housing	0.399688		0.029559	*
Income	0.026749	*	0.000391	***
Education2	0.000104	***	6.03E-09	***
PartyID	0.450331		0.676498	
Marital	0.141915		0.093499	.

Conclusion:

We can clearly observe that while interactions are considered the features Partner, SpouseEdu, GBirth, Dependent, Income, Education has a significant Pr(>|Z|) value lesser than 10 percent in common irrespective of Hillary being considered or not. When Hillary column is considered, the significance of Media, Birthplace, Housing and Marital features was over 10 percent, which says that Hillary column can change the odds of determining Trump Liberal or Conservative. While Interactions are not considered all the fields except PartyID has a significant impact on Trump.

Question 1.B

Introduction:

Required to build a suitable prediction model to predict an individual's party identification using the respective individual's personal, and family characteristics. Since Party ID column has 4 different values (Classes) we need to build classification models to identify the individual's PartyID.

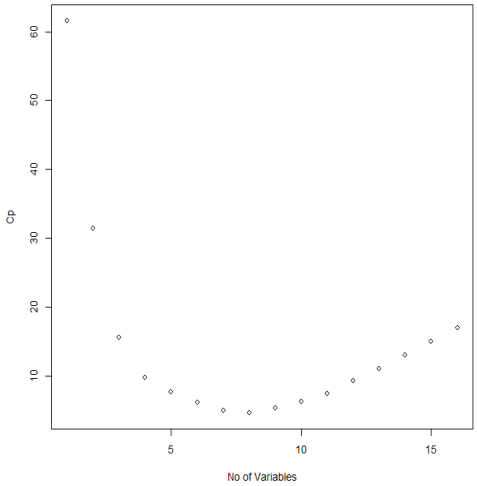
Method:

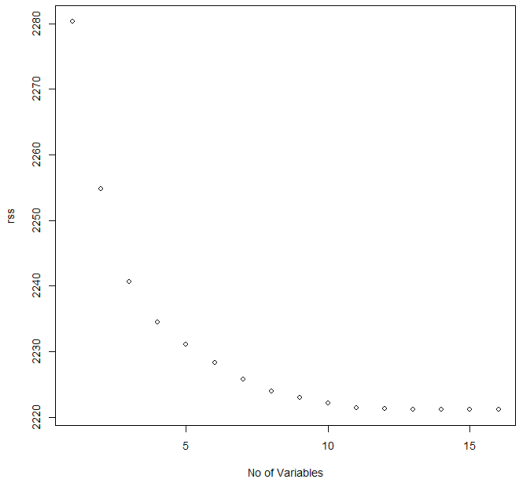
For this task, I have made a copy of the data frame with name df_party. Similar to the previous problem dropped rows with negative vales except for SpouseEdu and Partner as -1 is significant. After dropping missing, unnecessary rows and ID column. Based on the validation set approach, train and test data sets are created with 75 percent train and 25 percent test. Targeted response test$PartyID is stores in response variable "Y" for further evaluations. As part of building model for predicting the response, different models and methods are performed.

I. Multinomial Logistic Model with all columns:
   - Considering all the 16 fields multinomial logistic model is created to predict the response.
   - Below image shows the confusion matrix and the accuracy of the model.

```
                Y
predict_fit   1   2   3   4
          1 189  50 113  16
          2  67 184  99  25
          3  74  33  81   8
          4   0   0   0   0
> mean(Y == predict_fit)
[1] 0.4834931
```

II. Multinomial Logistic Model with best subset selected columns:
   - In the process of improving the performance of the model, I have tried to implement the best subset method to identify the significant variables.
   - Using the regsubsets() function with nvmax as 16, identified number of predictors based on RSS and CP ranking.





- Proceeded further considering the ran of Cp and minimum number of variables as per Cp ranking is 8, and below are the 8 predictors identified by best subset prediction.

```
"Media"     "Hillary"   "Trump"     "Age"       "Partner"   "SpouseEdu"  "Birthplace" "GBirth"
```

- Created new train and test data frames, using the existing data frames by extracting only the required 8 predictors and response variables as it avoids providing list of columns again and again further.
- Now performed Multinomial Logistic regression with the new set of columns and below the results.

```
                Y
predict_fit_bs   1   2   3   4
             1 193  45 112  17
             2  66 193 106  25
             3  71  29  75   7
             4   0   0   0   0
> mean(Y == predict_fit_bs)
[1] 0.4909478
```

III. K Nearest Neighbors Model with best subset selected columns and K = 5:

- Using the columns in the best subset selection, created a k nearest neighbors' model with K as 5.
- Below is the confusion matrix and accuracy of KNN model with K as 5.

```
   party_knn
Y       1      2      3      4
  1 0.7061 0.2212 0.0727 0.0000
  2 0.1760 0.6255 0.1948 0.0037
  3 0.0990 0.3003 0.5973 0.0034
  4 0.0408 0.1633 0.7959 0.0000
> mean(party_knn==test_bs$PartyID)
[1] 0.6123536
```

IV. K Nearest Neighbors Model with best subset selected columns and K = 10:

- Using the columns in the best subset selection, created a k nearest neighbors' model for prediction with K as 10.
- Below is the confusion matrix and accuracy of the model with K as 10

```
   party_knn2
Y       1      2      3      4
  1 0.7152 0.2182 0.0667 0.0000
  2 0.1985 0.6667 0.1348 0.0000
  3 0.1024 0.3140 0.5836 0.0000
  4 0.0204 0.1633 0.7959 0.0204
> mean(party_knn2==test_bs$PartyID)
[1] 0.6240682
```

Conclusion:

Analyzing the accuracy percentages of 4 models created, the most accurate model was KNN with K as 10 which has 62.41% accuracy when performed on the columns identified using best subset selection. Next is the KNN model with the similar columns but with K as 5 had accuracy as 61.24 which is close to that of k =10. The best subset selection multinomial model had lesser accuracy compared to the above 2 models with 49.09% and the least accurate model was simple multinomial model with all columns used has accuracy of 48.34%.

Finally, though the KNN2 model has 62.41 accuracy, if we observe class wise, the performance is very poor with respect to class 4 as only 2% of accurate prediction and almost 80% prediction towards class 3. Further analysis can be performed on other classification algorithms like LDA, QDA and K-fold cross validation approach also can be implemented to attain a minimal variance model.

Question 2

Summary:

Large-scale prediction algorithms have become popular and treated as heirs to regression algorithms. When analyzed with larger datasets, it was surprising that how do these algorithms compare to typical regression approaches like ordinary least squares or logistic regression? Several major distinctions will be investigated, with a focus on the differences between prediction and estimate or prediction and attribution. Statistical regression methods may be traced back to Gauss and Legendre in the early 1800s, and in particular to Galton in 1877.

Regression theories were used to a range of significant statistical problems over the twentieth century. Regression is aligned a lot of most significant ideas in twentieth-century statistics. Pure prediction algorithms may function at massive sizes, with millions of data points and predictor variables. Algorithms majorly deep learning have become the game changers because of their effectiveness in automating jobs such as online shopping, machine translation, and flight information. A quick web search yields deep learning in biology, computational linguistics with deep learning and deep learning for regulatory genomics.

Pure Prediction Algorithms are a valuable addition to the statisticians, but they require significant additional improvement before they can be used in normal scientific applications. Such progress is already being made in the statistical world, and it has offered a much-needed boost to our profession.

The purpose of this journal was to be descriptive for the present practice rather than prescriptive of how things should be. There is no assumption of prior understanding of the different prediction techniques and so it helps readers without much background in it also. This was not a scientific work, and majority of the argument is supported by numerical examples.

Traditional regression models, according to the author, outperform pure prediction models since they are based on a scientific formula rather than a black box. Each variable in a classical model, such as linear regression, has a coefficient that describes how much each variable impacts the result. The author concludes that "the emperor has lovely clothes, but they aren't appropriate for every occasion," implying that prediction models can be effective in certain areas but not in others.

Critical Analysis with Example:

In the context of analyzing how machine learning techniques helps in solving various problems, I have tried to understand how traditional methodologies and advanced techniques impacted a Big Data classification problem which analysis the challenges in Network intrusion prediction using Machine Learning. Traditional Machine Learning (ML) approaches have been created and utilized for extracting valuable information from data utilizing labeled datasets through training and validation.

Three fundamental issues render ML approaches inadequate for handling Big data categorization problems: A ML technique trained on a specific labeled dataset or data domain may not be suitable for another dataset or data domain, and thus the classification may not be robust across different datasets or data domains; A ML technique is generally trained using a limited number of class types, and thus a large variety of class types found in a dynamically growing dataset will result in inaccurate classification results. ML techniques are built around a single learning objective, they are unsuitable for today's many learning tasks and knowledge transfer needs of Big Data analytics.

The Author stated that, in order to solve traffic intrusion prediction problem a supervised algorithms like classification can be used to classify the data. Support Vector Machines is one of the highly performing algorithms for this problem, but the computational costs are higher than other classification models.

Questions to Author:

1. How a highly imbalanced dataset, say 100:1 proportionate 2 class classification model be handled in terms of splitting into train and test dataset, with giving a considerable significance of the lower proportionate class.

2. How does cross-validation is performed on a time-series dataset, if new classes were added after a certain period? For example, if there is a tournament like Olympics with games A, B, and C being held every year, say we have data for the past 30 years of each country. Assume after the 9th year a new game D was introduced into the tournament which was expected to be held in further years and game B is stopped after year 7. So, what kind of cross-validation can be applied if we wish to predict which game a country would win in the next year’s tournament.

References:

- Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. Performance Evaluation Review, 41(4), 70-73.doi: 10.1145/2627534.2627557
- Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. 2013.