Statistical Data Mining I - Homework 2

Tuesday, October 1, 2019 12:17 PM

Vimal Kumarasamy
UBIT name: vimalkum
E-mail: Vimalkum@buffalo.edu
Collaborators: Nikita Goswami

Q 3 - Creating dataset - Muting estimates - Generating response - Exhaustive search - Inference

Dataset Creation

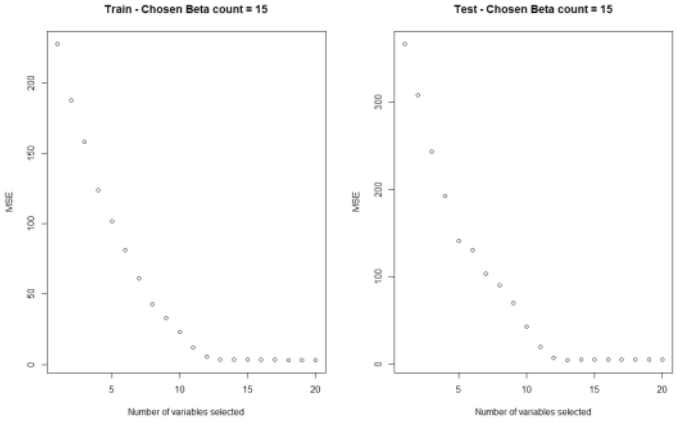
- Created dataset with 1000 observations and 20 features - normally distributed
- Defined variables that would choose the number of beta estimates to be considered based on which the response variable is generated

Beta estimate Muting

- The muted estimates are randomly chosen, so fixed set of estimates are not always muted
- Used matrix cross product to generate the response and added an error that is normally distributed (lower magnitude compared to the features)
- Randomly sampled 100 datapoints as train and 900 as test

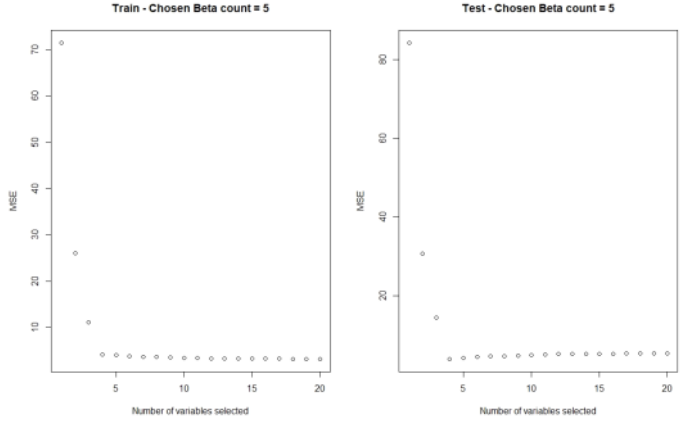
Response geraration and introducing error

- For the a given beta_count = 15 (considering 15 beta estimates, muting 5), an exhaustive search model is built using regsubsets
- For different values of I between 1 and 20, the best models are chosen and the respective train and test error are computed
- The error metrics are plotted against the number of variables chosen as per the exhaustive search



Inference

- We can observe that the test error is increasing when we consider more than 15 variables
- This is the expectation, when we add more meaningless variables in the model, test error increase after a point
- The similar characteristics was observed when beta_count was set as 5 and 10 etc.



Beta Coefficients

- The below snip shows that the model was able to arrive at the estimates which was pretty close to the random values that were assigned to the beta while response generation

- There are some inaccuracies which could be due to the normally distributed error that was introduced on purpose
- There are instances where the used beta was not part of the model, which would have been compensated by the intercept value learned in the model
- In the below snip, best_coef_5 was learned in the model where as beta_matrix was the beta based on which response was generated

```
# > best_coef_5
# (Intercept)              V7           V9                                V15           V17          V20
# 0.42207165      0.32465291      0.71449835                          0.38201244    0.04590932   0.94904069
# > beta_matrix
# V1 V2 V3 V4 V5 V6        V7 V8         V9 V10 V11 V12 V13         V14        V15 V16       V17 V18 V19  V20
# [1,]  0  0  0  0  0  0 0.3250954  0    0.7277053   0   0   0   0 0.001136587 0.3912033 0 0 0   0        0.9516588
```

Q1 College dataset - LM, Lasso, Ridge, PCR and PLS

LM

- College dataset is split into Train (80%) and Test (20%)
- Linear model is fit with all the variables
- RMSE is 790.853

Ridge Regression

- Model.matrix function is used to convert the factors to one-hot encoded variables
- Ridge model is built and based on CV fold, the best lambda is identified as 377.924
- The high value of lambda could be due to higher magnitude of the features
- RMSE is 783.519

Lasso Regression

- Process similar to Ridge is followed for Lasso and best lambda is identified based on CV fold
- Lambda is identified as 2.213509
- RMSE is 787.6368
- The variables such as Outstate, Personal and Expend variables have been almost removed from the model

PCR

- The plot with different components and the respective RMSE shows that when # of components = 8, RMSE is almost low and doesn't reduce significantly after that
- RMSE when 8 components are chosen : 1105.95
- This is significantly higher than the previous models built so far

PLS

- The plot with different components and the respective RMSE shows that when # of components = 8, there is lower RMSE and doesn't reduce significantly after that
- RMSE at number of components 8 : 922.3273

Inference

- LM observes the least RMSE, however we have used all the 18 features in the model, and few of the variables are not significant in that model
- With Ridge and Lasso methods (penalized models), the error is reduced further, however the number of components required in that case is also high
- The number of variables which has almost been removed from the model in Lasso, is 3. This shows that about 15 variables are present in the model still
- PCR has higher RMSE when we use 8 components, where as PLS for the same number of components RMSE is 922, which is far lesser than PCR
- The reason for PCR to perform better is because, PLS considers the correlation between the residual of the model and the feature that is not currently used in the model
- Based on the correlation, it chooses the variable that explains the residual the maximum, so it considers the relationship between features and Y to decide on the components
- Whereas PCR only considers the components within X, based on the maximum variance, in other words it doesn't consider Y
- So the better model here would be PLS with 8 components (parsimonious model and with fairly lesser RMSE)

Q2 - Insurance company benchmark data - Variable selection experiment

Dataset study

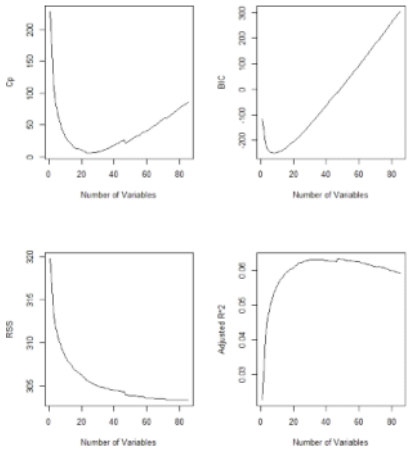
- Quick summary on the response variable in Training Dataset shows that there is a heavy class imbalance
- About 5% of the overall observations pertain to positive class and the rest are negative class
- This might lead to most of the predicted scores leaning towards the negative class

Linear Regression

- Least squared model is built and we need to arrive at a cutoff based on which we can define 1s and 0s
- The summary of the predicted scores show that the scores are crowded near 0.05 and thus we can't have 0.5 as the threshold
- Reduced the threshold to 0.1 and recall is 52.9% with a very low precision of 14.4%
- There are few insignificant variables here in the least square model
- It doesn't make sense to look at accuracy here, as if the model predicts every datapoint as 1, it would result in a 95% accuracy which is misleading

Forward variable selection

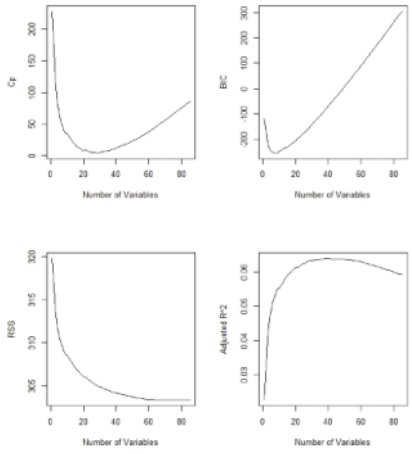
- Using reg subset selection with forward method, a model for every step has been built
- Let's look at the model statistics



- The plot shows that even with all the variables included, the adjusted R^2 is around 6% and this is a very poor model
- Given the options, the model we can leverage here would be with 22 variables, where adjusted R^2 is almost maximum with least number of variables
- The Recall that we arrived at with this model is about 51.2% and Precision at 14%

Backward variable selection

- Using backward selection also shows a similar performance
- Threshold of 0.1 is selected to qualify 1s which was decided based on the summary of the predicted value



- Model with 22 variables is chosen as the best model, and the recall and precision are mentioned below
- Recall: 47.47% and Precision: 13.4%

Lasso Model

- Best lambda is chosen based on CV fold and similar threshold of 0.1 is set to qualify 1s
- The recall is at 44% and Precision at 14%

Ridge Model

- Ridge model resulted in a recall of 42.8% with 14% precision

Inference

- Who are interested in buying Caravan insurance policy?
  - We can interpret the results based on the LR model estimates and the results are as follows
  - The below features are significant in the least square model and are having positive correlation with the number of caravan policies in the household
  - This denotes that if these metrics are higher for a household, they tend to have higher likelihood to purchase a caravan policies (Assuming a causation based on correlation here)
    - PPERSAUT Contribution car policies
    - MGEMLEEF Avg
    - PBRAND Contribution fire policies
    - APLEZIER Number of boat policies
  - The below feature is significant and has negative relationship with the number of caravan policies
    - GEZONG Number of family accidents insurance policies
  - However, for better interpretability we can build a model with just the significant variables, by that approach the estimates are credible

- Comparison of different algorithms

Method	*Recall*	*Precision*
Linear Regression	52.9%	14.4%
Forward subset selection	51.2%	14.2%
Backward subset selection	47.5%	13.4%
Lasso	44.1%	14.2%
Ridge	42.8%	14.7%

  - We see that Recall is ranging between 52.9% to 42.8% across algorithms with a similar precision
  - Given that Forward subset selection method has one of the highest Recall and comparatively lesser features - 22, This is the best model that we have now
  - The problem that is identified in this dataset is 'Class Imbalance' as there are only 5% positive class in the dataset
  - If we want to improve the performance further, we can probably over sample the positive classes and increase their representation in the model
  - We can also build an ensemble model, where the first model is the current model, the second model can have the False Negatives from the current model as positive class
  - So we can finally qualify the cases as positive class, when it qualifies as positive class in any one of the ensemble models