University of Glasgow

Machine Learning & Artificial Intelligence for Data Scientist M (COMPSCI5100)

Case Study:
Feature Engineering

Student: Ching Hsuan Lin
Student ID: 2702329L

Student: Weijian Ning
Student ID: 2821551N

Table of Contents

1. Introduction

Feature selection is a process of selecting a subset of features from a given set of features that are relevant to the current learning task. As a data pre-processing process, feature selection is usually an important prerequisite to determine the superiority of model training in practical machine learning tasks.

The main reasons are that, First, in the actual task of research, the research object will often have a large amount of attribute information. But in the case with high dimensionality, problems such as sparse data samples and difficulty in distance calculation are likely to emerge. Therefore, to prevent the problem of a dimensional curse, it is necessary to select the appropriate key features from the attributes.

Second, by removing irrelevant features, the difficulty of training machine learning models can be significantly reduced, while the stability of the outcomes can also be improved.

This paper first provides a methodological description of the classifiers for validation, cross-validation strategies, evaluation metrics, and feature selection methods used in the design of the feature engineering strategy. Then, the tuning method of the classifier hyperparameters, the process of searching the optimal feature numbers to retain, and the results of specific experimental data in the implementation of the model are shown. Next, by analysing the results, the comparison and evaluation of the performance of different feature selection methods are done. Finally, with the evaluation results, suggestions are given for the selection of feature engineering strategies.

2. Methods Introduction

2.1 Classifier

• 2.1.1 K-nearest Neighbour

K-nearest neighbour(KNN) is one of the most accessible supervised learning, meaning targets or classes are contained within the datasets. The principle of KNN is that when predicting a new data point x, the algorithm will select n points, which are a hyperparameter as n_neighbour, and calculate the distance between x and the n points. And then, assign the group of the nearest n to x. Therefore, the number of neighbours and the measurement of distance are the significant factors determining prediction accuracy.

• 2.1.2 Support Vector Machine

Support Vector Machine(SVM) is a supervised learning method whose objective is to find a decision boundary among the given classes in the dataset. The method maximises the margins of the classes and will finally reach an optimal hyperplane that best distinguishes the classes.

• 2.1.3 Logistic Regression

Logistic regression is a classification method that uses a logistic function for binary classification problems, which can be derived from linear regression. The unit-step function and the sigmoid function are implemented to find a line that best separates the two classes.

2.2 Leave One Group Out Cross-Validation

In normal machine learning, the dataset is divided into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate the metrics of the model. Cross-validation is an approach to avoid overfitting by dividing the training into two subsets that one for training and one for validation. When the training set is composed of groups of data, leave one group out cross-validation will be the best way to determine which group are chosen to be the validation subset. If the size of the data set is N, then N-1 pieces of data will be used for training, and the remaining one group of data will be used for validation. Take the dataset in this report as an example; all ten samples from a subject of the dataset are considered as the validation set in each training section.

2.3 Hypterparamater Optimization

- ### 2.3.1 Grid Search

The function of Grid Search can be automated hyper-parameter tuning. The optimised results and parameters can be evaluated as long as the parameters are input. In grid search, parameters are searched, that is, parameters are adjusted in turn according to the step size within the specified parameter range, and the modified parameters are used to train the learner to find the parameter with the highest evaluation metrics score from all the parameters, which is a process of training and comparison.

In this report, we are using leave one group out cross-validation accuracy to evaluate model performance since the ground truth label is given in the dataset. It is a completive way to evaluate the overall performance of models.

2.4 Evaluation Metrics

- ### 2.4.1 Accuracy

The accuracy rate is the most commonly used classification performance index, which is the correct proportion of the prediction, as shown in the former below.

$$Accurany = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives}$$

- ### 2.4.2 Sensitivity - True Positive Rate

Sensitivity, also known as true positive rate, is a judgment index to quantify and avoid false negatives. It can be thought of as finding out how many patients are really sick, also known as Recall. The higher the sensitivity, the lower the probability of missed true positives.

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- ### 2.4.3 Specificity - True Negative Rate

Specificity, also known as true negative rate, is a judgment index to quantify and avoid false positives. The characteristics of non-morbidity (we call it health here) are different from the characteristics of morbidity. We use these differences to avoid misdiagnosis. The higher the specificity, the higher the probability of diagnosis.

$$Specificity = \frac{True\ Negatives}{True\ Positives + False\ Negatives}$$

2.5 Feature Selection

- ### 2.5.1 Filtering Method - Chi-Square

The filtering method will first decide which feature is more important in the data set and then select those features to form a new training set which can improve model performance. The feature selection process has nothing to do with the subsequent learner. This is equivalent to using the feature selection process to "filter" the initial features first, then using the filtered features to train.

The classic chi-square test is a hypothesis-testing method which aims to evaluate the correlation between two independent variables. This method is also used to check the importance of features in chi-square feature filtering. And then, we can select the top k features as the new training set, where k is a parameter decided by ourselves.

- ### 2.5.2 Wrapper Method - Backward feature elimination

The wrapper method differs from filtering feature selection in that it selects features directly by evaluating the performance of the final model metrics generated by fitting the new feature set. The method uses a basic model for multiple rounds of training. After each round of training, features with less importance are eliminated, and then the next round of training is performed based on a new feature set.

In other words, the objective of wrapped feature selection is to select the feature subset that is most beneficial to its performance as a "tailor-made" for a given learner. Generally speaking, since the wrapper feature selection method is directly optimised for training, the performance of the final score will be better than filtering feature selection. But on the other hand, since the training set needs to be trained many times during the feature selection process, the computation overhead of wrapping feature selection is usually much larger than that of filtering feature selection.

- ### 2.5.3 Embedding Method - L1 Regularization

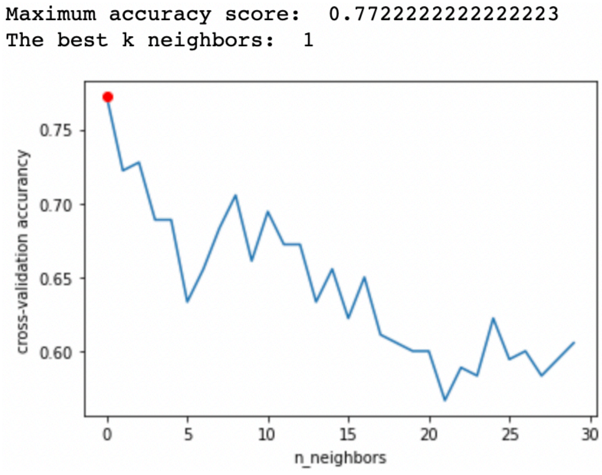
In filtering and wrapper methods, the feature selection process is obviously different from the training process; in contrast, the embedding method combines the feature selection process and the training process into the same process. By using the base model with penalty items to filter out features, dimensionality reduction is also performed. This is done in an optimisation process; that is, feature selection is implemented automatically during the training process. The principle of L1 penalty dimensionality reduction is to retain one of the features with the same correlation with the target value, so the unselected feature does not mean it is not essential.

3. Model Implementation and Feature Selection

3.1 Model Implementation

• 3.1.1 KNN (Grid Search)

In KNN, we grid search from 1 neighbour to 30 neighbours within integer steps and find the number of neighbours with a maximum leave one subject out cross-validation accuracy score, as shown in Figure 1. We can see how the number of neighbours affects the leave one subject out cross-validation accuracy score, where the score is on the y-axis, and the number of neighbours is on the x-axis; the red dot is the highest score where the k neighbour is one and the accuracy is 0.7722.

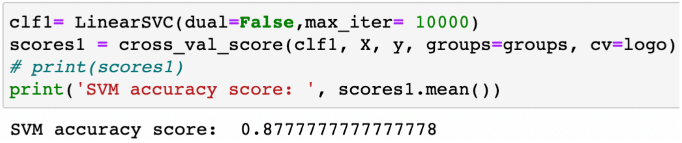


```
Maximum accuracy score:  0.7722222222222223
The best k neighbors:  1
```

Figure 1. KNN grid search results

• 3.1.2 SVM

In SVM, the linear SVM model is implemented to satisfy the requirements of some feature selections package we are introducing below. The max_iter parameter is set to 10000, and the leave one group one cross-validation accuracy is 0.8788, as shown in Figure 2.



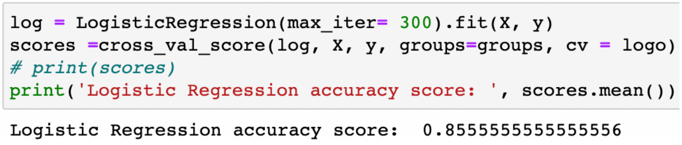
```
clf1= LinearSVC(dual=False,max_iter= 10000)
scores1 = cross_val_score(clf1, X, y, groups=groups, cv=logo)
# print(scores1)
print('SVM accuracy score: ', scores1.mean())
```

```
SVM accuracy score:  0.8777777777777778
```

Figure 2. SVM Loocv accuracy

• 3.1.3 Logistic Regression

In logistic regression, we set the max_iter to 300, and the leave one group one cross-validation accuracy is 0.8788, as shown in Figure 3.



```
log = LogisticRegression(max_iter= 300).fit(X, y)
scores =cross_val_score(log, X, y, groups=groups, cv = logo)
# print(scores)
print('Logistic Regression accuracy score: ', scores.mean())
```

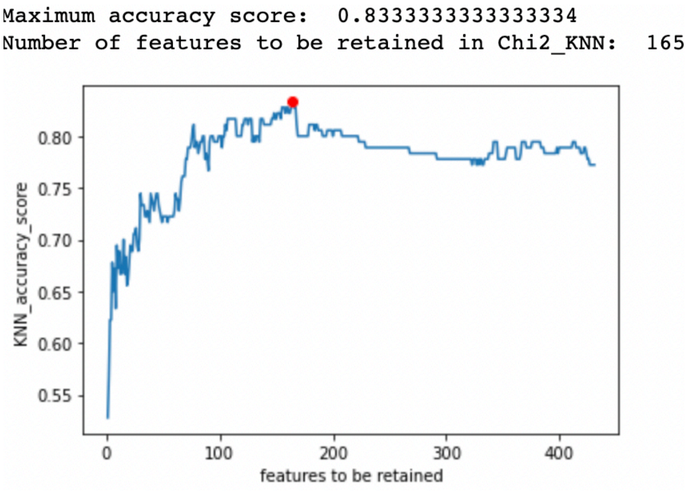
```
Logistic Regression accuracy score:  0.8555555555555556
```

Figure 3. Logistic Regression Loocv accuracy

3.2 Filtering Method - Chi-Square

• 3.2.1 KNN

In KNN, we grid search from 1 to 432 features and evaluate the model's performance by calculating the leave one group cross-validation accuracy score to find the best k features to remain in the KNN training set. The results are shown in Figure 4 below. The plot indicates that the accuracy score increases with the number of features and reaches the peak score of 0.8333 at 165 features remain, then slowly decreases after 165 to 432 features.

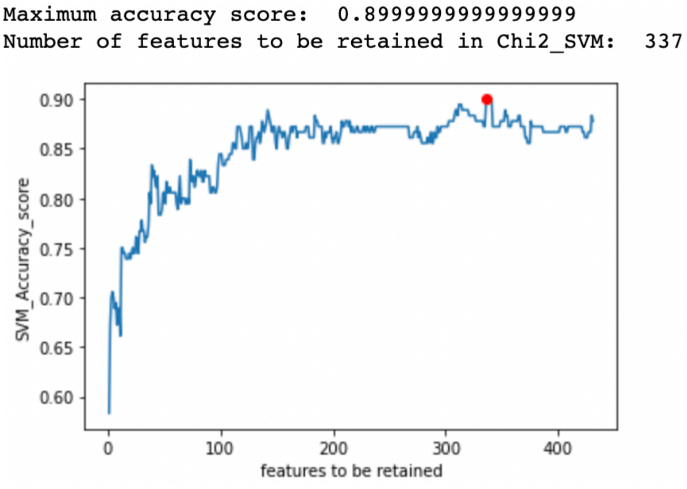


```
Maximum accuracy score:  0.8333333333333334
Number of features to be retained in Chi2_KNN:  165
```

Figure 4. KNN k features grid search results

• 3.2.2 SVM

In SVM, we also grid search from 1 to 432 features and evaluate the model's performance by calculating the leave one group cross-validation accuracy score to find the best k features to remain in the SVM training set. The results are shown in Figure 5 below. The plot demonstrates that the accuracy scores logarithmically grow with the number of features at first within 150 features and remains slightly over 0.85 accuracies after 150 to 432 features. The best score can be found at 337 features, and the score is 0.9.



```
Maximum accuracy score:  0.8999999999999999
Number of features to be retained in Chi2_SVM:  337
```

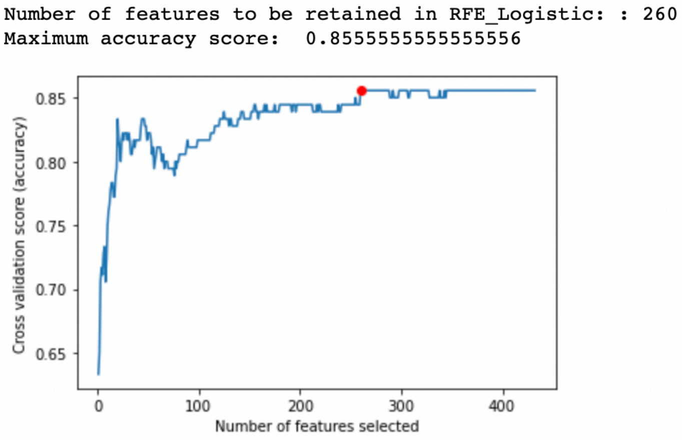
Figure 5. SVM k features grid search results

3.3 Wrapper Method - Backward feature elimination (BFE)

In this section, we implement the recursive feature elimination with cross-validation(RFECV) function from Scikit-Learn, an achievement of the wrapper method feature section.

• 3.3.1 Logistic Regression

In logistic regression, the RFECV function grid searches from 1 to 432 features with integer steps and evaluates the model metrics by leave one subject out cross-validation accuracy, as shown in Figure 6. The plot shows that the accuracy dramatically increases at first but drops to less than 0.8 accuracy at around 80 features. And then steadily grows to reach the highest score of 0.8556 as 260 features remain in the training set. After that, it fluctuates but remains the same score after the peak.

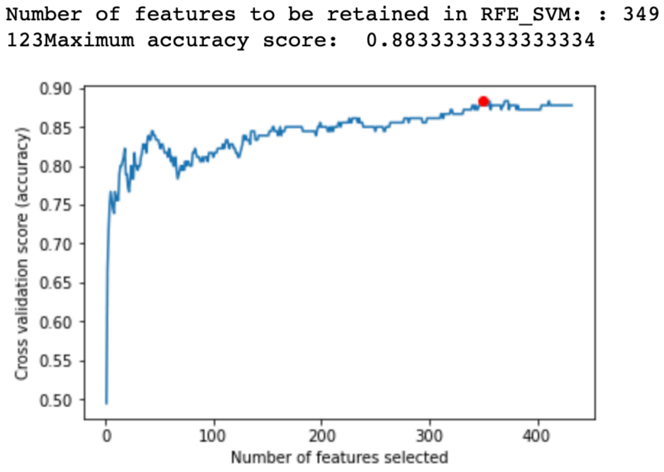


```
Number of features to be retained in RFE_Logistic: : 260
Maximum accuracy score:  0.8555555555555556
```

Figure 6. Logistic regression RFECV results

• 3.2.2 SVM

In SVM, the RFECV function also grid searches from 1 to 432 features with integer steps and evaluates the model metrics by leave one subject out cross-validation accuracy, as shown in Figure 7. The figure indicates that the accuracy scores increase dramatically at first until around 50-60 features are selected. There is also a drop at around 80 feature, which grows steadily and reaches the peak of 0.8833 accuracy score when 349 feature is selected. After the peak, the score fluctuates and remains almost the same.

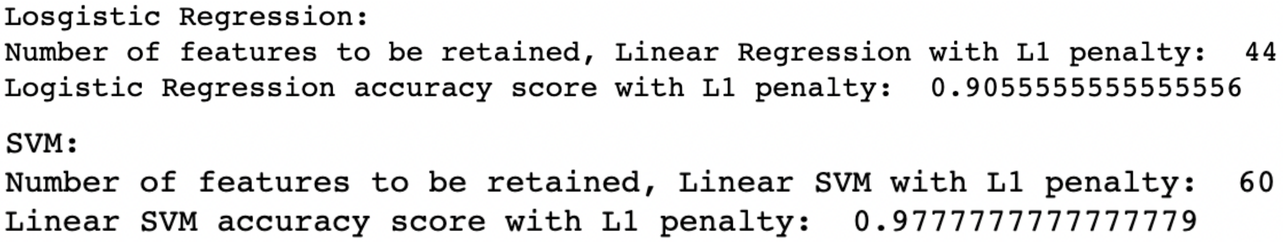


```
Number of features to be retained in RFE_SVM: : 349
123Maximum accuracy score:  0.8833333333333334
```

Figure 7. SVM RFECV results

3.3 Embedding Method - L1 Regularisation

In this section, we fit the L1 regularisation model separately in a logistic regression model and an SVM model. We also calculate the leave one group out cross-validation score as the models' metrics. The results are displayed in the python script below as Figure 8. In logistic regression, the number of features selected is 44, and the score is 0.9056; in SVM, the number of features is 60, and the score is 0.9778.



```
Losgistic Regression:
Number of features to be retained, Linear Regression with L1 penalty:  44
Logistic Regression accuracy score with L1 penalty:  0.9055555555555556

SVM:
Number of features to be retained, Linear SVM with L1 penalty:  60
Linear SVM accuracy score with L1 penalty:  0.9777777777777779
```

Figure 8. L1 regularisation results

4. Result and Discussion

The results of all methods above are shown in Table 1, where the leave one subject out cross-validation accuracy scores are evaluated within three models and four datasets (one original data and three feature selection methods). In chi-square feature selection, the KNN model greatly improves, while the SVM model also slightly increases. In backward feature elimination, the logistic regression model has the same accuracy, which might be because of the limitation of the algorithm itself, and the SVM model shows a slight improvement. In L1 regularisation, both logistic regression and SCM models have a conspicuous growth in accuracy.

	KNN	Logistic Regression	SVM
Original data	0.77222	0.85556	0.87778
Chi-square	0.83333	-	0.9
BFE	-	0.85556	0.88333
L1 Regulariztion	-	0.90556	0.97778

Table 1. Loocv accuracy

5. Conclusion

To summarise the analysis above, all three feature selection methods contribute to improvements in most of the models. The filtering method increases the accuracies of the models slightly more than the wrapper method. On top of that, regarding the time complexity of the feature selection methods, the wrapper methods need to grid search the data size and evaluate the model metrics in every iteration, which is time-consuming. The embedded method has the most distinctive result of increasing the leave one group out cross-validation accuracy among all other methods, so obviously, this is the most sagacious feature selection strategy in this case.