LOVELY PROFESSIONAL UNIVERSITY

Assignment of Machine Learning Foundation

Name: Md Arifur Rahman Anik

Registration: 11800320

Section: KM001

Dataset: Breast_cancer

Submitted to

Usha Mittal

Abstract:

Feature engineering is the process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve the performance of machine learning algorithms. Feature engineering can be considered as applied machine learning itself. A feature is an attribute or property shared by all of the independent units on which analysis or prediction is to be done. Any attribute could be a feature, as long as it is useful to the model. The purpose of a feature, other than being an attribute, would be much easier to understand in the context of a problem. A feature is a characteristic that might help when solving the problem.

Introduction:

In classifying objects, the large number of features indicates the complexity of a problem. To solve complex classification problem, good features plays a crucial role, which have a higher predictive power. Though, several dimensional reduction techniques like feature selection exist to help many aspects of learning classification problems, but so far there is little support for a crucial step in the process of engineering features. Engineering good features set is prerequisite to achieve high accuracy in classifying objects and prediction.

As classification systems become an integral part of many domains to take effective decisions or solving prediction problems with least error rate and highest accuracy. It is difficult, however, to compare the accuracy of the techniques and determine the best one because their performance is data dependent. It is therefore necessary to move towards adaptive classification systems that selectively employ appropriate classification method by first engineering the available raw data as features that are basic building blocks or an observation for describing data to an algorithm that implements classification or prediction Features must capture salient aspects of the variables to help the target function learn the target result.. Trained systems are very challenging to build and picking up the right features for training the system is the most critical and time consuming part in developing a successful classification or prediction model.

Feature engineering for classification problem :

- Collection of Raw data : Collection of data from different sources which may be structured, unstructured or textual data.

- Data Preprocessing: Format the raw data by aligning, unions, grouping, intersections. Removing noisy, dirty data(missing, duplicates, ill-formed, wrong values etc.) by pairwise or list wise deletion, by computing imputation(mean substitution, Regression), by stochastic simulation, principle of least harm, sampling error, population parameters, propagation, statistical power etc. as approaches vary quite a lot. Some methods are elabored in the next sext section.

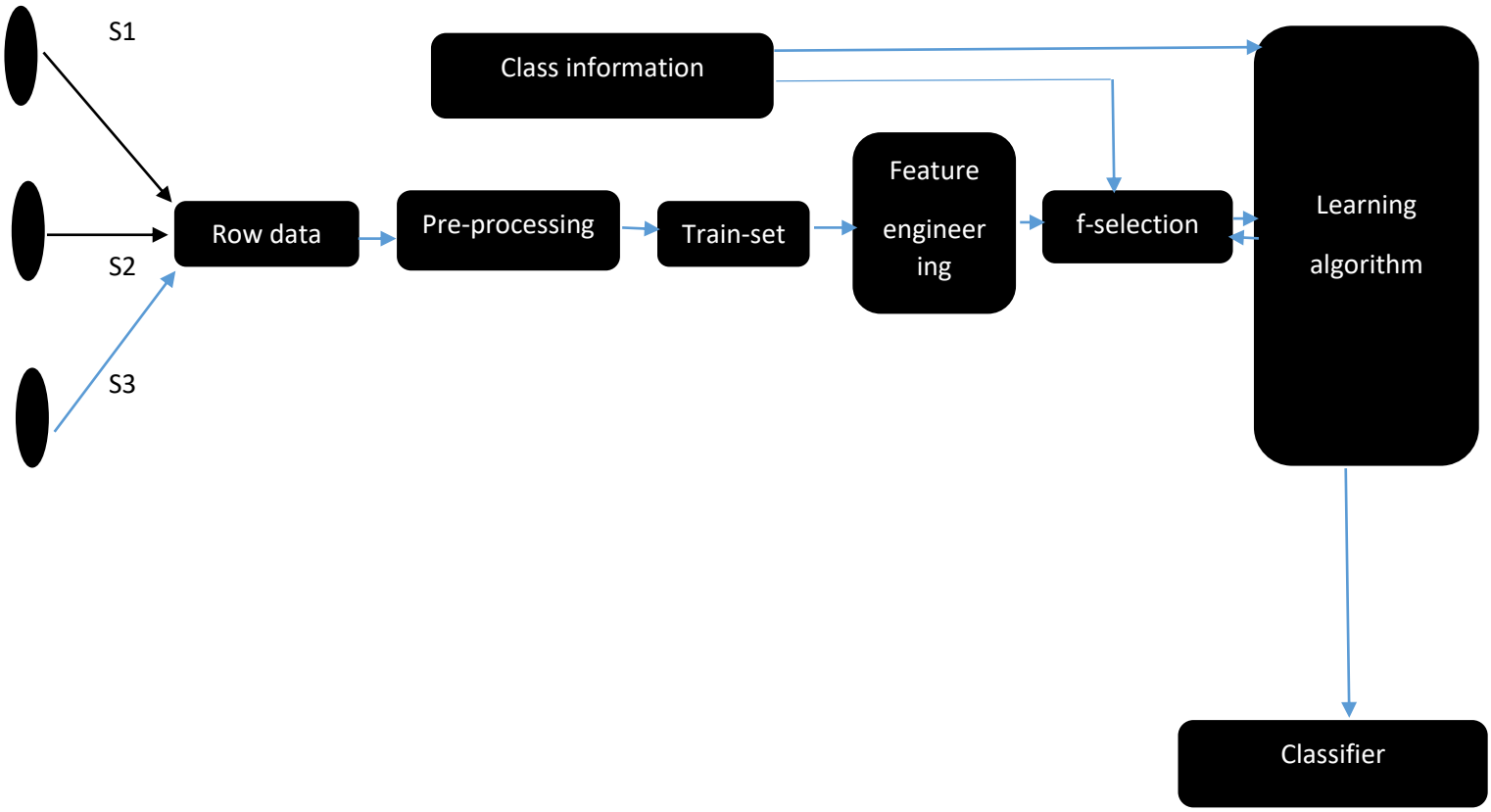
- Feature Engineering (Feature Transformation and Creation of additional features): Clean data is still raw where lot of data is useless, so data is filtered, sliced and transformed to create features for modeling. Feature engineering is hard where analysis and domain knowledge is required. Some common feature engineering methods might be PCA, Kernal PCA.

- Selection of Features: Append additional features with training data sets and perform feature selection and extraction and then process the selected features to the learning algorithm. The feature engineering and selection phase may be independent of the learning algorithm.

- Modeling and performance measure: Create models and it may iteratively utilize the performance of the learning algorithms to evaluate the quality of the selected features, like wrapper models, cross validation etc.

- Classifier: Finally selected features, a classifier is induced for the classification or prediction phase.

A General Framework of Feature Engineering for Classification



Prior to feature engineering methods, raw data need to be preprocessed as preprocessing also affects performance of classifiers. Some of the data preprocessing methods are as follows:

A. Variable Identification: Identify input and output(target) features(variable), their data type and category.Data type like character, Numeric, Variable category like Categorical, Continuous.

B. Univariate Analysis: Exploring features one by one with their type. If the type is continuous, can be measured using various statistical metrics visualization methods and for categorical, frequency table may be used to understand distribution of each category.

Dataset:

The name of the dataset is Breast_cancer . This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. (See also lymphography and primary-tumor.)

This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are linear and some are nominal.

	0	1	2	3	4	5	6	7	8	9
0	no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
1	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	right_up	no
2	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	left	left_low	no
3	no-recurrence-events	60-69	ge40	15-19	0-2	no	2	right	left_up	no
4	no-recurrence-events	40-49	premeno	0-4	0-2	no	2	right	right_low	no
...	...	...	...	...	...	...	...	...	...	...
281	recurrence-events	30-39	premeno	30-34	0-2	no	2	left	left_up	no
282	recurrence-events	30-39	premeno	20-24	0-2	no	3	left	left_up	yes
283	recurrence-events	60-69	ge40	20-24	0-2	no	1	right	left_up	no
284	recurrence-events	40-49	ge40	30-34	3-5	no	3	left	left_low	no
285	recurrence-events	50-59	ge40	30-34	3-5	no	3	left	left_low	no

There is some null values

```
Out[6]: 0    0
        1    0
        2    0
        3    0
        4    0
        5    0
        6    0
        7    0
        8    0
        9    0
        dtype: int64
```

After perform imputing on it

```
Out[57]: 0    0
         1    0
         2    0
         3    0
         4    0
         5    0
         6    0
         7    0
         8    0
         9    0
         dtype: int64
```

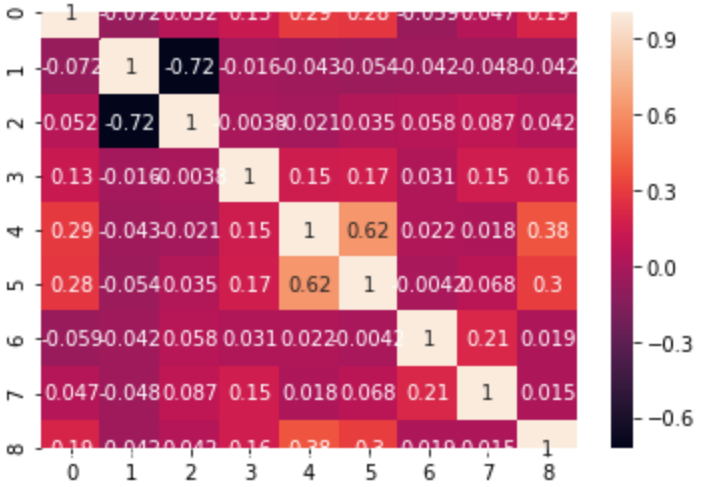
Performing Label Encoder :

In this dataset we can find many type of data like integer ,categorical and other . for this reason we need to perform label encoding by scikit learn . after performing this the dataset is like this .

Out[9]:

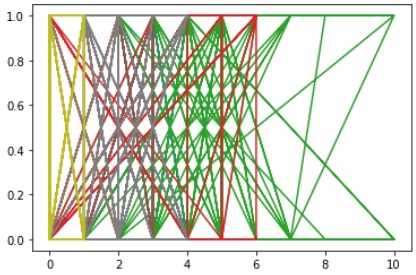
	0	1	2	3	4	5	6	7	8	9
0	0	1	2	5	0	0	3	0	1	0
1	0	2	2	3	0	0	2	1	4	0
2	0	2	2	3	0	0	2	0	1	0
3	0	4	0	2	0	0	2	1	2	0
4	0	2	2	0	0	0	2	1	3	0

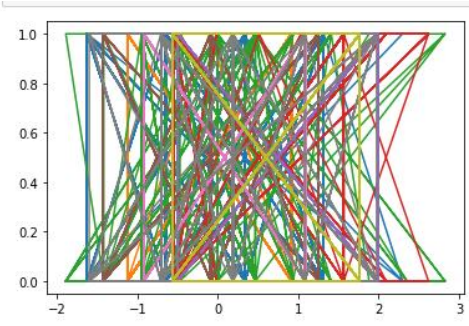
To see the correlation coefficient of the dataset and heat map ,



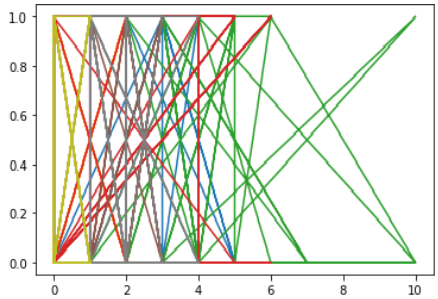
These are some graph of the dataset after doing the split method .

This graph is for x_train and y_train and x_test and y_test





X_train_sc and y_train_sc



In this experiment I have implemented many of regression method , SVM, KNN , Logistic Regression, Bagging Classifier , Ada Boost Classifier , Perceptron, Random Forest Classifier . and all the accuracy score are shown below.

Accuracy score of perceptron model :

```
0.5233644859813084
0.5138888888888888
```

In this model we can observe that this model is not over fitting because training and testing values are quite same .

Accuracy score of SVM model :

```
0.7336448598130841
0.8055555555555556
```

In this model also we can observe that this model is not over fitting because training and testing values are almost same .

Accuracy score of KNN model :

```
0.780373831775701
0.7916666666666666
```

This is not over fitting .

Accuracy score of logistic regression model :

```
0.7336448598130841
0.8055555555555556
```

This is also not overfitting .

Accuracy score of Random forest classifier model :

```
0.9672897196261683
0.7361111111111112
```

Accuracy score of Bagging Classifier model :

```
0.7336448598130841
0.7916666666666666
```

Accuracy score of AdaBoost Classifier model :

```
0.7476635514018691
0.75
```

This is also not overfitting .

Finally we performed Principal component analysis(PCA) ,Linear discriminant analysis(LDA) and kernel principal component analysis on Logistic Regression . and we got very good model .

```
accuracy score 0.7616822429906542
accuracy score 0.7222222222222222
```

```
Training accuracy score 0.7476635514018691
testing accuracy score 0.7638888888888888
```

```
Training accuracy score 0.7523364485981309
testing accuracy score 0.75
```

Conclusion:

Building an efficient classification model for classification problems with some additional relevant features along with different data sets and different sample size is important. The main tasks are construction of additional relevant features from the existing feature sets, and then feature selection to discard redundant, irrelevant or highly correlated

features. Selected features are minimal set of independent features which explain the patterns in the data and then classified outcomes successfully

In this research paper i tried to show all the operations of this dataset and also tried to show all the regression model and their accuracy score .

Reference :

Michalski,R.S. , Mozetic,I., Hong,J., Lavrac,N, Clark,P. & Niblett,T, Tan, M., & Eshelman, Cestnik,G., Konenenko,I, & Bratko.

Quinlan, J.R., C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets.html Weka: http://www.cs.waikato.ac.nz/~ml/weka/

https://www.sas.com/en_us/insights/analytics/machine-learning.html

https://en.wikipedia.org/wiki/Machine_learning

https://www.researchgate.net/post/How_to_get_data_set_for_breast_cancer_using_machine_learning

THANK YOU