Prospeção de Dados 2021/2022

Second Home Assignment

Group 4
Catarina Canastra (57766) - 25h contribution
Daniel Dias (59056) - 27h contribution
João Raimundo (57454) - 28h contribution

April 10, 2022

Introduction

Data sets are often high dimensional, containing a large number of features, although the relevancy of each feature for analysing this data is not always clear. So, because the data must be treated and the training process cannot be relied on to determine which features are really useful, it is imperative to perform feature selection and/or dimensionality reduction to reduce the number of features in a dataset. **Feature selection** is the simple process of selecting and excluding given features, whereas **dimensionality reduction** is the transformation of data from a high-dimensional space into a low-dimensional space. Choosing the best feature set is essential to both reduce the computation time and storage space required and improve the performance of the model.

1 Exploratory Data Analysis (EDA)

The sample dataset is available from UC Irvine Machine Learning Repository [1], which consists of superconductivity data. It is a multivariate physical dataset to predict the critical temperature of a superconductor. Dataset comprises of 21263 rows (superconductors) and has 82 numerical columns representing relevant features. The dependent variable is called "critical_temp". We've concluded that all variables are of the correct data type and no feature has null/missing values. There are some duplicate values that need to be deleted (so they do not bias the results). A descriptive table provides various summary statistics, where it can be observed a large difference between the 75th quantile and the max values of the predictors. This observations suggests that there are extreme value outliers in the dataset, however, the proportion of outliers in not very significant in this context/problem. The variables "wtd_gmean_Density" and "wtd_gmean_TermalConductivity" have high variance and three others stand out. 50% of the variables in the dataset have zero variance and therefore, feature selection methods are expected to remove these variables. It is also inferred that some features have a high correlation (above 90%) with other's, for example, "number_of_elements" and "entropy_atomic_mass". In order to transform the raw data into an understandable format, it needs to standardise the features so that the values are on a similar scale.

2 Methodology

Basic on a generic data mining workflow, the following methodology were defined. The methodology is divided into four main tasks: data preprocessing and data partition; test different feature selection procedures with different algorithms to identify the most important features in descending order of importance (**Objective 1**); perform dimensionality reduction and determine the best number of components (**Objective 2**); and, finally, use an exhaustive search method with 5-fold cross validation to evaluate the score of five algorithms with different set of parameters to identify the best model with the minimal set of features or dimensions (**Objective 3**).

Two different subsets of the data are used: initially, a training set of scaled independent variables, a test set of scaled independent features and a training and test dependent variable are used to identify the best features of each of the three models under test and analyze the explained variance ratio always between these three regressors. The methods for selecting the best set of features are the *SelectFromModel* class and the *SequentialFeatureSelection* from *Scikit-learn*[1]. The scaled training set of independent features is used to evaluate the variance explained with a different number of components. Then, a training set of scaled independent variables and a test set of scaled

[1] https://scikit-learn.org/stable/

independent features are used to analyze the data components. In the last objective, the training set is used to perform the grid search method over specified parameter values for 5 estimators. This method internally divides the input set into a training set and a validation set. train scaled e train pca After providing the ten best models/features and their corresponding hyperparameters, the test set is used to test the best model of all.

The code developed to implement the methodology described was written in *Python* programming language, in the *Jupyter Notebook* format, using the *Anaconda* package manager and, in the *Visual Studio Code* editor.

3 Results and Discussion

XXXX

Method	Algorithm	Column Indexes	RVE		
			RFs	DTs	LRs
Select From Model	RF Regressor	67,64,9,74,66,72	0.905	0.764	0.537
	DT Regressor	67,64,74,9,78,47,10,31,62	0.911	0.769	0.606
	Linear Regressor	22,72,75,49,1,14,4,62,25,73,54,17 26,44,19,47,76,52,12,2,74,15,24	0.920	0.753	0.650
Sequential Feature Selection	RF Regressor	3,9,10,15,17,38,47,52,61,70,71,72	0.914	0.734	0.584
	DT Regressor	2,11,15,19,37,38,46,47,49,52,77,79	0.912	0.670	0.493
	Linear Regressor	5,6,27,30,42,44,50,60,66,69,70,80	0.914	0.779	0.666

Table 1: Your caption.

Figure **??** shows the model F1-Score according to the alpha and the best set of variables for that value, for the two classification tasks.

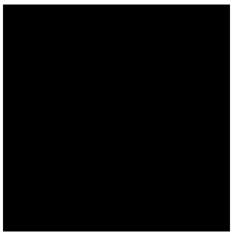

Figure 1: A figure

Author	Title
Knuth	The TeXbook
Lamport	LaTeX

Table 2: A table

discussion about the results and chosen set of features

4 Conclusions

This document has presented the use of

References

[1] Kam Hamidieh. Superconductivty Data. UCI Machine Learning Repository, 2018.