Dimension Reduction and Feature Selection

Machine Learning 2 | Assignment 5

Colab link:
https://colab.research.google.com/drive/1U3nz97keW1dsOjkpH4nhUj99mxE_6wm4?usp=sharing

Table of Contents

Dimension Reduction & Feature Selection

So far, dimensionality reduction and feature selection methods are applied on the given datasets.

For Classification: I have used 2 main classifiers: XGBoost and Random Forest.

- With that, PCA and LDA are used for Dimensionality Reduction.
- While, L1(Regularization), Random Forest, Recursive Feature Selection and XGBoost are used for feature selection.
- Overall Accuracy and ROC are used to measure classifiers' accuracy.

For Regression: I have used linear and voting regression models.

- PCA and RFE are used for dimensionality reduction and feature selection respectively.
- To measure accuracy of the model, I have used MAE, MSE and RMSE.

Feature Selection and Dimensionality Reduction Methods

Some important details of the methods used:

1) Principal Component Analysis:

The most popular technique for dimensionality reduction in machine learning is Principal Component Analysis, or PCA for short. This is a technique that comes from the field of linear algebra and can be used as a data preparation technique to create a projection of a dataset prior to fitting a model.
It is a popular linear feature extractor used for unsupervised feature selection based on eigenvectors analysis to identify critical original features for principal component.

2) LDA:

Linear Discriminant Analysis, or LDA for short, is a predictive modeling algorithm for multi-class classification. It can also be used as a dimensionality reduction technique, providing a projection of a training dataset that best separates the examples by their assigned class.
The model fits a Gaussian density to each class, assuming that all classes share the same covariance matrix. The fitted model can also be used to reduce the dimensionality of the input by projecting it to the most discriminative directions, using the transform method.

3) L1:

In linear model regularisation, the penalty is applied over the coefficients that multiply each of the predictors. From the different types of regularisation, Lasso or L1 has the property that is able to shrink some of the coefficients to zero. Therefore, that feature can be removed from the model. The logistic regression is used for the Lasso regularisation to remove non-important features from the dataset. Keep in mind that increasing the penalization c will increase the number of features removed.

4) Random Forest:

Feature selection using Random forest comes under the category of Embedded methods. Embedded methods combine the qualities of filter and wrapper methods. They are implemented by algorithms that have their own built-in feature selection methods. Some of the benefits of embedded methods are: highly accurate, generalize better and interpretable.
For classification, the measure of impurity is either the Gini impurity or the information gain/entropy.
For regression the measure of impurity is variance.

5) XGBoost:

A benefit of using ensembles of decision tree methods like gradient boosting is that they can automatically provide estimates of feature importance from a trained predictive model.
Generally, importance provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance. The performance measure may be the purity (Gini index) used to select the split points or another more specific error function. The feature importance is then averaged across all of the decision trees within the model.

6) Recursive Feature Elimination (RFE):

RFE is popular because it is easy to configure and use and because it is effective at selecting those features (columns) in a training dataset that are more or most relevant in predicting the target variable.
There are two important configuration options when using RFE: the choice in the number of features to select and the choice of the algorithm used to help choose features. Both of these hyperparameters can be explored, although the performance of the method is not strongly dependent on these hyperparameters being configured well.
For classification, DecisionTreeClassifier is used to choose features and set the number of features to five. We will then fit a new *DecisionTreeClassifier* model on the selected features.
For classification, DecisionTreeRegressor is used.
This method is slow as compared to other feature selection methods but effective. Technically, RFE is a wrapper-style feature selection algorithm that also uses filter-based feature selection internally.

Algorithms for Classification

1) XGBoost:

XGBoost stands for eXtreme Gradient Boosting specially designed to improve speed and performance. It is a faster algorithm when compared to other algorithms because of its parallel and distributed computing. XGBoost is developed with both deep considerations in terms of systems optimization and principles in machine learning. The goal of this library is to push the extreme of the computation limits of machines to provide a scalable, portable and accurate library. The main features are: Regularized Learning, Gradient Tree Boosting and Shrinkage and Column Subsampling.

2) Random Forest:

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

3) Linear Regression:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

4) Voting Regression:

Voting is an ensemble machine learning algorithm. For regression, a voting ensemble involves making a prediction that is the average of multiple other regression model. RandomForest Regressor and Linear Regressor both are used as an implementation of ensembling technique.

Credit Card Dataset (CLASSIFICATION PROBLEM):

Type	No. of Features	Method	Algorithm	Accuracy	ROC Accuracy	Precision
DR	3	PCA	XGBoost	0.801	0.7396	0.617
DR	3	PCA	RANDOM_FOREST	0.776	0.678	0.117
DR	1	LDA	XGBoost	0.8066	0.714	0.638
DR	1	LDA	Random Forest	0.808	0.713	0.656
FS	7	L1	XGBoost	0.815	0.761	0.680
FS	7	L1	Random Forest	0.805	0.749	0.702
FS	13	RF	XGBoost	0.814	0.768	0.678
FS	13	RF	Random Forest	0.775	0.751	0.689
FS	3	RFE	XGBoost	0.774	0.599	0.033
FS	3	RFE	Random Forest	0.775	0.557	0.0
FS	3	XG	XGBoost	0.815	0.731	0.684
FS	3	XG	Random Forest	0.805	0.716	0.7
DR	10	PCA	XGBoost	0.8	0.744	0.62
DR	10	PCA	RANDOM_FOREST	0.775	0.697	0.0
FS	10	RFE	XGBoost	0.815	0.762	0.689
FS	10	RFE	Random Forest	0.787	0.744	0.684
DR	15	PCA	XGBoost	0.808	0.757	0.642
DR	15	PCA	RANDOM_FOREST	0.775	0.712	0.0
FS	15	RFE	XGBoost	0.815	0.770	0.682
FS	15	RFE	Random Forest	0.776	0.752	0.608

On credit card default dataset, I have multiple DR and FS methods.

Q: How does the classification performance compare across the 7 DR/FS methods?

Here what I have found, results can be relate from the table:

For Minimum Feature/Component Selection:

- Using PCA for DR with XGBoost Classifier gives 80% accuracy with 61% precision. While using LDA with Random Forest, auto-selected 1 feature for DR with 90% variance and gives almost 81% accuracy with 65% precision (best) so far.
- Similarly, using XG for FS with XGBoost Classifier gives 81.5% accuracy with 68.4% precision (best with 3 features).
- While RFE has given 0% precision.

For 10 feature selection:

- Using PCA for DR with XGBoost Classifier also gives 80% accuracy with 62% precision.
- Using RFE for FS with XGBoost Classifier gives 81.5% accuracy with 69% precision.

- While PCA using Random Forest has given worst results with 0% precision which means the results using this model will not be accurate at all.

For 15 feature selection:

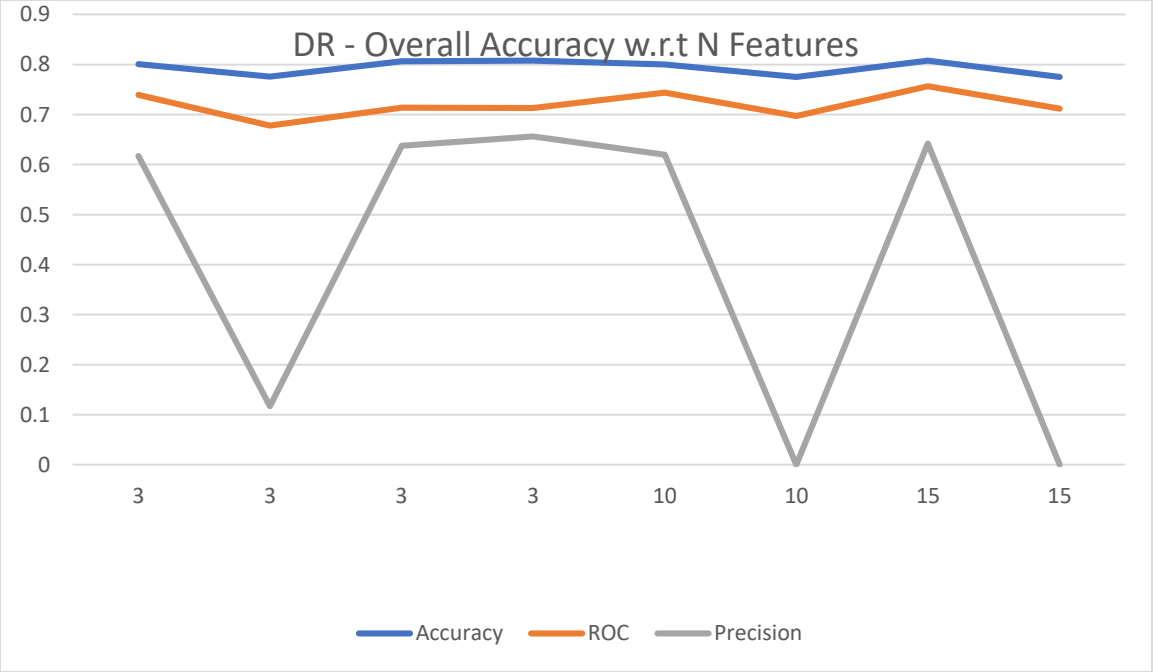
- Using PCA for DR with XGBoost Classifier also gives 80% accuracy with 64% precision.
- Using RFE for FS with XGBoost also gives 81.5% accuracy with 68% precision.
- Again, PCA performance is worst with the increase in number of features.

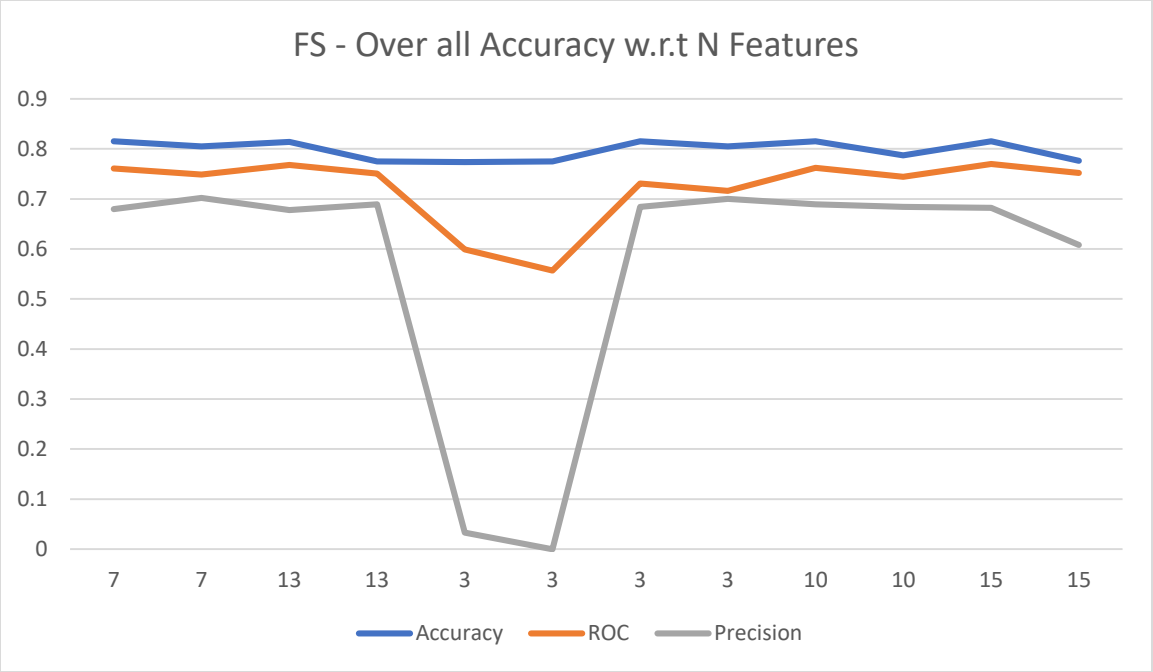
Overall XGboost has performed really well as a classifier and for dimension reduction LDA has given Best result while RFE gives best accuracy for Feature Selection.

Q: What is the effect of increasing or decreasing the total number of your desired features on the classification performance?

I can assume that there is trade-off between number of features and precision as we can see that with 3 no. of features PCA and RFE have given a very low precision value while with the increase in features (10), RFE has performed better and has given 81.5% accuracy with 69% precision.

However, Random Forest using PCA has given a poor performance with 0% precision.





Super Conduct Dataset (REGRESSION PROBLEM):

Method	Features	MAE	MSE	RMSE
Linear Regression	5	18.514	544.27	23.329
Voting Regression	5	11.882	254.928	15.966
RFE	5	7.558	210.251	14.5
Linear Regression	10	17.794	500.763	22.377
Voting Regression	10	11.301	226.089	15.036
RFE	10	7.155	189.359	13.760
Linear Regression	20	16.674	438.671	20.944
Voting Regression	20	10.668	205.401	14.331
RFE	20	7.021	186.993	13.674

On super conduct dataset, PCA is used as a Feature Selection for Linear and Voting regression while RFE has used Decision Tree Regressor as a Regression Algorithm.

Q: How does the regression performance compare across the 7 DR/FS methods?

Here what I have found, results can relate from the table:

For 5, 10 and 20 feature selection:

- Linear Regression didn't perform well as giving highest error.

- While RFE and Voting Regression, they both give a close performance where RFE has won by giving lowest error rate.

Overall we can say that both voting and RFE are ensembling methods and have given a better performance than simple linear regression.

Q: What is the effect of increasing or decreasing the total number of your desired features on the classification performance?

Since the dataset has 81 features, the result shows that features have correlation with each other as with the increase in number of features we can see that the error decrease rapidly. This change is shown below that the relationship among no. of features and error is inversely proportional to each other for all methods.