Linear LDA for classification

Predictive modeling

ITESO, Universidad Jesuita de Guadalajara

Gregorio Alvarez

Introduction

This document aims to provide a comprehensive comparison of different variable selection methods for the classification problem, with a focus on the use of linear models. The methods under consideration include the benchmark model, correlation-based selection, Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA). By comparing these methods, we aim to determine the most effective approach for variable selection in the context of classification using linear models. The evaluation will consider factors such as model performance and interpretability. The insights gained from this analysis has the purpose of finding the strengths and weaknesses of each method, and to provide a recommendation for the most suitable method for the classification problem.

Data

The data used for this analysis is the Heart Disease UCI dataset from UC Irvine repository. The dataset contains 303 observations and 14 variables, including the target variable. The target variable is a binary variable indicating the presence of heart disease in the patient. The remaining variables are a mix of categorical and continuous variables, and are described in the table below.

Variable	Description	Type
age	age in years	categorical
sex	sex	categorical
cp	chest pain type	categorical
trestbps	resting blood pressure	continuous
chol	serum cholestoral in mg/dl	continuous
fbs	fasting blood sugar > 120 mg/dl	categorical
restecg	resting electrocardiographic results	categorical
thalach	maximum heart rate achieved	continuous
exang	exercise induced angina	categorical
oldpeak	ST depression. relative to rest	continuous
slope	slope of peak exercise ST segment	categorical
ca	number of major vessels	categorical
thal	type of defect	categorical
num	Presence of heart disease (target)	categorical

Methods

The data was characterized using clasical nalysis order to get insight of the original format. From this analysis missing data, representing 1.9% of the total data, was found. These observations were drawn from the datsaset.

id	ca	thal	num
87	0.0	?	0
166	?	3.0	0
192	?	7.0	1
266	0.0	?	2
287	?	7.0	0
302	?	3.0	0

Table 1. Missing data

A second descriptive analysis was performed to get a better understanding of the relation of the dependent variables and the target variable. The results of this analysis are shown in the following two tables.

	sex	cp	exang	slope	thal
sex	1.000000	0.078621	0.190396	0.029293	0.397816
cp	0.078621	1.000000	0.438351	0.151966	0.294815
exang	0.190396	0.438351	1.000000	0.214334	0.303364
slope	0.029293	0.151966	0.214334	1.000000	0.181306
thal	0.397816	0.294815	0.303364	0.181306	1.000000

Table 2. Spearman correlation

	age	trestbps	chol	fbs	restecg	thalach	oldpeak	ca
age	1.000000	0.242166	0.187714	0.162945	0.103325	0.407587	0.205622	0.301078
trestbps	0.242166	1.000000	0.090187	0.199315	0.122061	0.010049	0.077950	0.055723
chol	0.187714	0.090187	1.000000	0.030658	0.135911	0.031801	0.054819	0.132022
fbs	0.162945	0.199315	0.030658	1.000000	0.079521	0.015820	0.038117	0.184629
restecg	0.103325	0.122061	0.135911	0.079521	1.000000	0.005288	0.038169	0.123223
thalach	0.407587	0.010049	0.031801	0.015820	0.005288	1.000000	0.367592	0.259173
oldpeak	0.205622	0.077950	0.054819	0.038117	0.038169	0.367592	1.000000	0.282241
ca	0.301078	0.055723	0.132022	0.184629	0.123223	0.259173	0.282241	1.000000

Table 3. Pearson correlation

From the previous correlation tables, it can be seen that the correlation between the variables si low, therefore the use of all the variables in the model is justified.

The categorical variables were transformed into dummy variables, and the ordinal variables were standarized after the train-test split. The data was split into a 70/30 sets. The training set was used to fit the models, and the testing set was used to evaluate the performance of the models.

The models were evaluated using the accuracy metric, which is defined as the proportion of correct predictions. Before the training process, a distribution of 137 positive and 160 negative entries on the target variable was found. This distribution was mantained in the training and testing sets.

Benchmark model

For this method, all the variables were used to fit a logistic regression model. The results of the model are shown in the table below.

Train	Test
0.874	0.822

Table 4. Benchmark model accuracy

Correlation based

For this method, the variable pairs with a correlation greater than 0.4 were checked. The most correlated variable was removed from every pair. The selected variables were age, sex, trestbps, chol, fbs, restecg, exang, oldpeak, ca, cp_2, cp_3, slope_2, slope_3 and thal_6.0. The results of the model are shown in the table below.

Train	Test
0.860	0.789

Table 5. Correlation based selection accuracy

PCA

For this method, a cumulative explained variance plot was obtained to determine the number of components to be used. Since the variables were standarized before this process, the incremental explained variance of every principal component is negligible. For this reason sets of 1 to 6 components were used to fit the model. The chosen model was the one with 3 components, since the addition of the following components did not improve the accuracy of the model. The results of the model are shown in the table below.

Train	Test
0.807	0.722

Table 6. PCA selection accuracy

LDA

The LDA transformation was performed using the the train set. Since there are only two classes in the target variable, only one component was obtained. The results of the model are shown in the table below.

Train	Test
0.870	0.789

Table 7. LDA selection accuracy

Results

Below is a table with the accuracy of every model.

Model	Features	Train	Test
Benchmark	17	0.874	0.822
Correlation based	14	0.860	0.789
PCA	3	0.807	0.722
LDA	1	0.870	0.789

Table 8. Model accuracy

Discussion

In terms of interpretability the LDA model is the best, since it only uses one feature and it's possible to get the anti-transformed values of the feature. The correlation based model is more interpretable than the benchmark model, since it uses less features. The PCA model and the Benchmark model are the least interpretable, since the PCA model uses a linear combination of the original features, and the benchmark model has some correlated features, which would lead to a biased interpretation of the coefficients.

In terms of performance, the benchmark model is the best, followed by the correlation based model which has a similar number of features. The PCA and LDA models have a similar performance, with the LDA model having a slightly better performance for a smaller number of features.

Conclusion

The LDA transformation has proven to be a suitable method for the classification problem, providing sparse feature reduction and effectively representing the data with a smaller set of features. It performs comparably or better than other methods, indicating its effectiveness in reducing dimensionality while preserving discriminative information. Overall, LDA is a valuable tool offering a balance between feature reduction and classification accuracy. Further research can explore its application in different classification problems and compare its performance with other techniques.