

Classifying car price ranges with neural networks

Sivert M. Skarning

March 2020

Contents

1	Introduction	3
1.1	Related work	3
2	Method	3
2.1	Data quality	3
2.1.1	missing values	3
2.1.2	duplicate values	3
2.2	Encoding	4
2.3	Artificial Neural Networks(ANN)	4
2.3.1	Opening remarks	4
2.3.2	Training data	4
2.3.3	Test data	5
2.3.4	K-fold cross-validation	5

1 Introduction

This project will try to find data pre-processing methods and a neural network that best predicts the buying price of a car, based on the car evaluation dataset. It will also compare performance and anccuracy between decision trees and neural networks on this dataset.

1.1 Related work

There are numerous articles that have studied the performance of different modeling techniques with respect to the car evaluation dataset. The article by Sameer Singh[4] discusses the performance of varying training set sizes for different classification methods for the car evaluation sets. Sameer used artificial neural networks, K-nearest neighbour, decision trees and support vector machines in order to classify the acceptability of each car.

An article[3] also explored the performance of data mining classification methods. Here the authors also focus on the pre-processing of the data. They discuss concepts like data-cleaning, data-transformation and splitting of the data-set.

2 Method

2.1 Data quality

In order to ensure data quality it is necessary to assess the dataset. In this report we will clean the dataset by removing duplicates and fill missing values. We do this to give the neural network algorithms quality data to analyze[1].

2.1.1 missing values

After running an r script that counts missing values, we foundt out that there are no missing values. This will save us from doing interpolation or fill with mean method to fill missing values.

2.1.2 duplicate values

Duplicate values might give the modeling algorithm an idea that the date counts more than other data. This migh contribute to over-fitting. After running a script that counted duplicates in r, we found out that the dataset did not have any duplicate data.

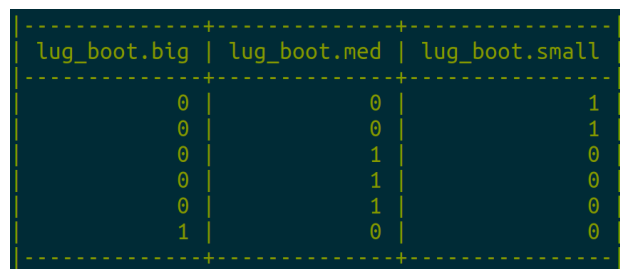
2.2 Encoding

There are three main encodings when working with classification of categorical data[2].

- Integer encoding
- One Hot encoding
- Learned embedding encoding

In this project i will explore the performance of One Hot encoding on the car evaluation data set. This encoding method is used when the machine learning algorithm might not be able to understand the relationship between the data. Integer encoding on this dataset gave poor performance.

In figure 1 you can see how one hot encoding encodes the lug_boot size category into numerical values by splitting the categories into seperate columns of data.



lug_boot.big	lug_boot.med	lug_boot.small
0	0	1
0	0	1
0	1	0
0	1	0
0	1	0
1	0	0

Figure 1: Excerpt of One Hot Encoding on car evaluation dataset

2.3 Artificial Neural Networks(ANN)

2.3.1 Opening remarks

In this project i decided to use neuralnet in r for the neural network prediction. This is a well know package and it has good resources.

2.3.2 Training data

After cleaning the dataset and applying one-hot encoding I fit the model to the training data. The calculation of the accuracy of the model was done with the with the result matrix and the original data. This showed an error rate of 64% wich is worse than the error rate of the decition tree. I decided to split the dataset into to different sets. One training set and one testing

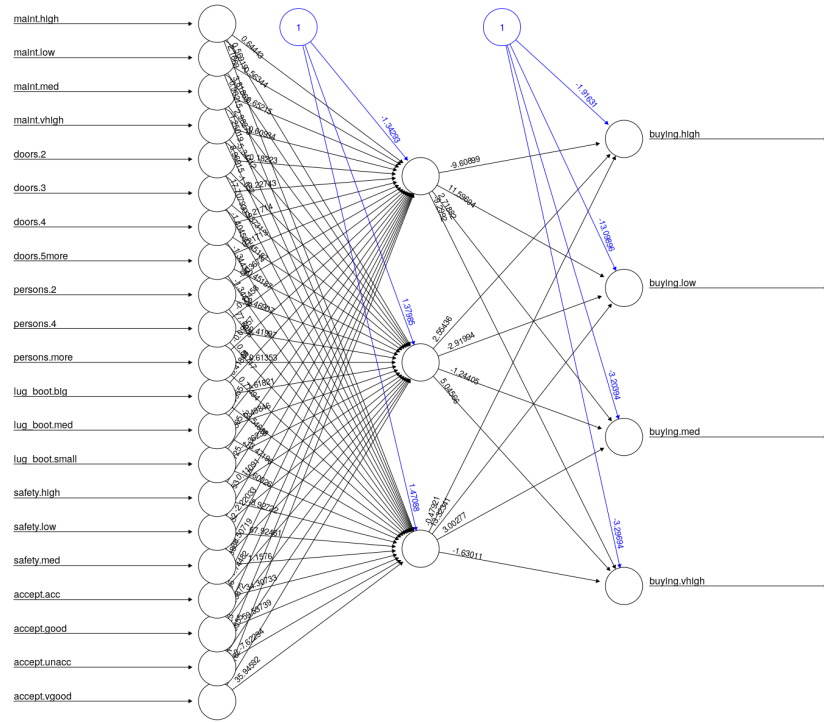


Figure 2: Artificial Neural Network

set. This prediction was done on the training set. This is not accurate of how the model will perform in real life since the model can be highly tuned for this set (over-fitted).

2.3.3 Test data

The prediction on the test data will display the accuracy of the model more realistic. A huge gap between accuracy in test data-set and training data-set will be signs of over-fitting.

When running the ANN on the validation data we got an almost identical accuracy. On the training set we got 35.57% accuracy, when running the model on the validation data we got a 36.67%. The fact that the two results are so simulare to eachother is a good indicator that the model is not over-fitted.

2.3.4 K-fold cross-validation

Even tough the model is not over-fitting i decided to try k-fold cross-validation in order to find the accuracy of the model. Cross-validation is a method for

	[,1]	[,2]	[,3]	[,4]
1383	0.23159957	2.579760e-01	0.25722731	0.231791022
411	0.24153519	2.454995e-01	0.25416921	0.241860548
431	0.55299105	1.079132e-02	0.13549577	0.579161748
1193	0.26355184	2.203555e-01	0.24796110	0.263166388
799	0.46675712	8.613516e-02	0.21365853	0.387412042
1260	0.26652479	2.172073e-01	0.24715686	0.266039653
1071	0.25005984	2.353632e-01	0.25171477	0.250093363
1578	0.24406715	2.424337e-01	0.25343128	0.244312533

Figure 3: ANN results

testing the model that used all of the data for testing and all of the data for validation. The way it works is that it splits the dataset into k different parts. It then uses k - 1 parts for training and the last part for validation. Then it repeats it k times in such a way that all the k-parts has been used in both testing and validation. Then it takes the average accuracy of these models.

In my project I used 10 fold cross validation. Cross validation

References

- [1] Data preprocessing concepts. <https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825>. Accessed: 2020-03-02.
- [2] Why one-hot encode data in machine learning? <https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>. Accessed: 2020-03-02.
- [3] Jamilu Awwalu, Anahita Ghazvini, and Azuraliza Abu Bakar. Performance comparison of data mining algorithms: A case study on car evaluation dataset. *Int. Jour. of Computer Trends and Technology (IJCTT)*, 13(2), 2014.
- [4] Sameer Singh. Modeling performance of different classification methods: deviation from the power law. *Project Report, Department of Computer Science, Vanderbilt University, USA*, 2005.