Ana Clara Tupinamba Freitas

Metro College of Technology, Data Science and Application

[anaclarat@gmail.com](mailto:anaclarat@gmail.com)

Abstract

The application of Encoding is vital for the creation of models since most variables, although can be counted, are categorical by nature. This paper focus on describing some of the transformations that can be applied when such a task is required.

Keywords: encoding, one hot encoding, dummy coding, ordinal coding, sum coding, deviation coding, multicollinearity

Introduction

Machine Learning heavily relies on Statistical concepts that is based on the application of Mathematical calculations, thus, to be able to analyse and create models with categorical variables one must convert them to numerical values. This task is not as easy as it seems because the loss of information may be present, specially when the variable is not ordinal.

Methods

Consulting of papers and python documentation.

Discussion

The conversion of categorical to numerical variables implies an understanding of the meaning and distribution of the categorical variables, besides the goal of analysis to correctly apply the proper technique.

The most used technique is **One Hot Coding** (also known as dummy coding) where the n levels of a categorical variable are transformed into n new variables, for example:

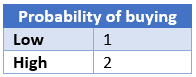
Table

Description automatically generatedA picture containing background pattern

Description automatically generated

With this conversion the levels are transformed into new Boolean variables. In the example the first row refers to a Winter while the second to a Summer register. Notice that when applied to a linear regression model in order to avoid multicollinearity, one of the new variables can be dropped, if relevant to the problem ( [When can you safely ignore multicollinearity](https://statisticalhorizons.com/multicollinearity), by Paul Allison and [Think twice before dropping that first one-hot encoded column](https://inmachineswetrust.com/posts/drop-first-columns/) by Red Huq explains more of when and not drop a variable due to multicollinearity). If this is the case, if we drop Summer, and encounter an observation with a value of zero for Winter, it means that the observation is related to Summer.

When the variable is ordinal, conversion is simpler (**Ordinal Coding**), one can assign numbers to the categories, for example:

Graphical user interface, text, application

Description automatically generated

This kind of conversion utilizes the order, so Low being first in ascending order gets 1 followed by 2 for High. This kind of coding has an entrapment: it implies order that may not exist and must be applied with caution.

The third coding technique is the **Sum Coding**: although similar to the first technique (dummy encoding) , it takes a level of the variable as reference using -1, instead of 0. It contrasts the mean of the dependent variable for a given level to the overall mean of mean of the dependent variable over all the levels. This kind of coding is useful when you are trying to find the deviances of levels compared to the reference and because of this application, this technique is also known as deviation coding:

Table

Description automatically generatedA picture containing background pattern

Description automatically generated

Conclusion

The use of coding permits data scientists adds to their model categorical variables, expanding the power of predictions and explanation of such models by translating to numerical values what before could not be explained in such way. But there are some considerations to be aware when applying the techniques to not lose predictions or explanation power, especially when multicollinearity is a preoccupation.

References

<https://www.researchgate.net/publication/320465713_A_Comparative_Study_of_Categorical_Variable_Encoding_Techniques_for_Neural_Network_Classifiers>

<https://www.investopedia.com/terms/s/statistics.asp>

<https://phillipalday.com/stats/coding.html>

<https://www.youtube.com/watch?v=WRxHfnl-Pcs>

<https://www.algosome.com/articles/dummy-variable-trap-regression.html>

<https://inmachineswetrust.com/posts/drop-first-columns/>