Data Encoding

Tushar B. Kute,
http://tusharkute.com

Data Encoding

- In the field of data science, before going for the modelling, data preparation is a mandatory task.
- There are various tasks we require to perform in the data preparation. Encoding categorical data is one of such tasks which is considered crucial.
- As we know, most of the data in real life come with categorical string values and most of the machine learning models work with integer values only and some with other different values which can be understandable for the model.

Data Encoding

- All models basically perform mathematical operations which can be performed using different tools and techniques.
- But the harsh truth is that mathematics is totally dependent on numbers.
- So in short we can say most of the models require numbers as the data, not strings or not anything else and these numbers can be float or integer.

Data Encoding

- Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the models to give and improve the predictions.

Categorical Data

- Categorical data can be considered as gathered information that is divided into groups.
- For example, a list of many people with their blood group: A+, A-, B+, B-, AB+, AB-,O+, O- etc. in which each of the blood types is a categorical value.
- There can be two kinds of categorical data:
  - Nominal data
  - Ordinal data

Types of encoding

- Label Encoding or Ordinal Encoding
- One-Hot Encoding
- Effect Encoding
- Binary Encoding
- Base-N Encoding
- Hash Encoding
- Target Encoding

Label Encoding or Ordinal Encoding

- This type of encoding is used when the variables in the data are ordinal, ordinal encoding converts each label into integer values and the encoded data represents the sequence of labels.

One-Hot Encoding

- In One-Hot Encoding, each category of any categorical variable gets a new variable. It maps each category with binary numbers (0 or 1).
- This type of encoding is used when the data is nominal. Newly created binary features can be considered dummy variables.
- After one hot encoding, the number of dummy variables depends on the number of categories presented in the data.

Effect Encoding

- In this type of encoding, encoders provide values to the categories in -1,0,1 format. -1 formation is the only difference between One-Hot encoding and effect encoding.

Hash Encoding

- Just like One-Hot encoding, the hash encoder converts the category into binary numbers using new data variables but here we can fix the number of new data variables.
- Before going to the implementation we should know about hashing.
- So hashing is used for the transformation of arbitrary size input in the form of a fixed-size value.

Thank you

This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com