krishnaik06 Add files via upload History

1 contributor

533 lines (533 sloc) | 16.9 KB

Ordinal numbering encoding or Label Encoding

Ordinal categorical variables

Ordinal data is a categorical, statistical data type where the variables have natural, ordered categories and the distances between the categories is not known.

For example:

- Student's grade in an exam (A, B, C or Fail).
- Educational level, with the categories: Elementary school, High school, College graduate, PhD ranked from 1 to 4.

When the categorical variables are ordinal, the most straightforward best approach is to replace the labels by some ordinal number based on the ranks.

In [6]:
```
import pandas as pd
import datetime
```

In [7]:
```
# create a variable with dates, and from that extract the weekday
# I create a list of dates with 20 days difference from today
# and then transform it into a datafame

df_base = datetime.datetime.today()
df_date_list = [df_base - datetime.timedelta(days=x) for x in range(0,
20)]
df = pd.DataFrame(df_date_list)
df.columns = ['day']
df
```

Out[7]:

	day
0	2019-09-12 12:39:01.344998
1	2019-09-11 12:39:01.344998
2	2019-09-10 12:39:01.344998
3	2019-09-09 12:39:01.344998
4	2019-09-08 12:39:01.344998
5	2019-09-07 12:39:01.344998
6	2019-09-06 12:39:01.344998
7	2019-09-05 12:39:01.344998
8	2019-09-04 12:39:01.344998
9	2019-09-03 12:39:01.344998
10	2019-09-02 12:39:01.344998
11	2019-09-01 12:39:01.344998

12	2019-08-31 12:39:01.344998
13	2019-08-30 12:39:01.344998
14	2019-08-29 12:39:01.344998
15	2019-08-28 12:39:01.344998
16	2019-08-27 12:39:01.344998
17	2019-08-26 12:39:01.344998
18	2019-08-25 12:39:01.344998
19	2019-08-24 12:39:01.344998

In [8]:
```
# extract the week day name

df['day_of_week'] = df['day'].dt.weekday_name
df.head()
```

Out[8]:

	day	day_of_week
0	2019-09-12 12:39:01.344998	Thursday
1	2019-09-11 12:39:01.344998	Wednesday
2	2019-09-10 12:39:01.344998	Tuesday
3	2019-09-09 12:39:01.344998	Monday
4	2019-09-08 12:39:01.344998	Sunday

In [9]:
```
# Engineer categorical variable by ordinal number replacement

weekday_map = {'Monday':1,
               'Tuesday':2,
               'Wednesday':3,
               'Thursday':4,
               'Friday':5,
               'Saturday':6,
               'Sunday':7
}

df['day_ordinal'] = df.day_of_week.map(weekday_map)
df.head(20)
```

Out[9]:

	day	day_of_week	day_ordinal
0	2019-09-12 12:39:01.344998	Thursday	4
1	2019-09-11 12:39:01.344998	Wednesday	3
2	2019-09-10 12:39:01.344998	Tuesday	2
3	2019-09-09 12:39:01.344998	Monday	1
4	2019-09-08 12:39:01.344998	Sunday	7

5	2019-09-07 12:39:01.344998	Saturday	6
6	2019-09-06 12:39:01.344998	Friday	5
7	2019-09-05 12:39:01.344998	Thursday	4
8	2019-09-04 12:39:01.344998	Wednesday	3
9	2019-09-03 12:39:01.344998	Tuesday	2
10	2019-09-02 12:39:01.344998	Monday	1
11	2019-09-01 12:39:01.344998	Sunday	7
12	2019-08-31 12:39:01.344998	Saturday	6
13	2019-08-30 12:39:01.344998	Friday	5
14	2019-08-29 12:39:01.344998	Thursday	4
15	2019-08-28 12:39:01.344998	Wednesday	3
16	2019-08-27 12:39:01.344998	Tuesday	2
17	2019-08-26 12:39:01.344998	Monday	1
18	2019-08-25 12:39:01.344998	Sunday	7
19	2019-08-24 12:39:01.344998	Saturday	6

Ordinal Measurement Advantages

Ordinal measurement is normally used for surveys and questionnaires. Statistical analysis is applied to the responses once they are collected to place the people who took the survey into the various categories. The data is then compared to draw inferences and conclusions about the whole surveyed population with regard to the specific variables. The advantage of using ordinal measurement is ease of collation and categorization. If you ask a survey question without providing the variables, the answers are likely to be so diverse they cannot be converted to statistics.

With Respect to Machine Learning

- Keeps the semantical information of the variable (human readable content)
- Straightforward

Ordinal Measurement Disadvantages

The same characteristics of ordinal measurement that create its advantages also create certain disadvantages. The responses are often so narrow in relation to the question that they create or