20 March Assignment

May 28, 2023

[ ]:
```
Q1. What is data encoding? How is it useful in data science?
```

ANS -

[ ]:
```
 Data encoding is the process of converting data from one form to another. It␣
 ↪is useful in data science because it helps in converting
categorical data into numerical data which can be used for analysis. There are␣
 ↪different types of encoding techniques such as One-Hot Encoding,
Label Encoding, Ordinal Encoding and Dummy Encoding.

For example, One-Hot Encoding is used when there are multiple categories in a␣
 ↪column and we want to convert them into separate columns.
Label Encoding is used when there is a hierarchy in the columns which can be␣
 ↪misleading to nominal features present in the data. Ordinal
Encoding is used when there is a hierarchy in the columns which can be␣
 ↪misleading to nominal features present in the data. Dummy Encoding is
used when there are multiple categories in a column and we want to convert them␣
 ↪into separate columns.
```

[ ]:
```

```

[ ]:
```
Q2. What is nominal encoding? Provide an example of how you would use it in a␣
 ↪real-world scenario.
```

ANS -

[ ]:
```
 Nominal encoding is a type of categorical data encoding where no notion of␣
 ↪order is present. For example, the city a person lives in. It is
important to retain where a person lives but we do not have any order or␣
 ↪sequence. One way to encode nominal data is using one-hot encoding
which is similar to binary coding considering there are in less number.

    example of how you would use nominal encoding in a real-world scenario:␣
 ↪Suppose you have a dataset containing information about different
types of fruits such as apple, banana, and orange. You want to use this dataset␣
 ↪to train a machine learning model that can predict the type of
fruit based on its features such as color, size, and weight. Since the type of␣
 ↪fruit is a categorical variable with no notion of order, you can
```

```
use nominal encoding to convert it into numerical values that can be used by␣
 ↪the machine learning model.
```

[ ]:

[ ]:
```
Q3. In what situations is nominal encoding preferred over one-hot encoding?␣
 ↪Provide a practical example.
```

ANS -

[ ]:
```
Nominal encoding is preferred over one-hot encoding when the categorical␣
 ↪features don't have any order (nominal). In One-Hot Encoding each
categorical column is split into many columns depending on the number of␣
 ↪categories present in the column. Each column is mapped with 0 or 1s.

For example, if we have a dataset with a column called "color" and the values␣
 ↪are "red", "green", and "blue", we can use one-hot encoding to
create three new columns called "color_red", "color_green", and "color_blue".␣
 ↪Each row will have a 1 in the column corresponding to its color
value and 0s in all other columns.

On the other hand, nominal encoding can be used when we have categorical data␣
 ↪that doesn't have any order. For example, if we have a dataset
with a column called "fruit" and the values are "apple", "banana", and␣
 ↪"orange", we can use nominal encoding to map each value to a unique
integer.
```

[ ]:

[ ]:
```
Q4. Suppose you have a dataset containing categorical data with 5 unique values.
 ↪ Which encoding
technique would you use to transform this data into a format suitable for␣
 ↪machine learning algorithms?
Explain why you made this choice.
```

ANS -

[ ]:
```
There are several encoding techniques that can be used to transform categorical␣
 ↪data into a format suitable for machine learning algorithms.
Some of the popular encoding techniques are:

Label Encoding: In this technique, each label is converted into an integer␣
 ↪value. This technique is useful when the labels have some order or
hierarchy.

1.Ordinal Encoding: This technique is similar to label encoding but is␣
 ↪generally used when we are intended for input variables that are
```

```
organized into rows and columns (e.g., matrix).

2.One-Hot Encoding: In this technique, each label is mapped to a binary vector.␣
 ↪This technique is useful when there is no order or hierarchy
among the labels.

3.Learned Embedding: In this technique, a distributed representation of the␣
 ↪categories is learned.

The choice of encoding technique depends on the nature of the data and the␣
 ↪machine learning algorithm being used. For example, one-hot
encoding is preferred when there is no order or hierarchy among the labels.
```

[ ]:

[ ]:
```
Q5. In a machine learning project, you have a dataset with 1000 rows and 5␣
 ↪columns. Two of the columns
are categorical, and the remaining three columns are numerical. If you were to␣
 ↪use nominal encoding to
transform the categorical data, how many new columns would be created? Show␣
 ↪your calculations.
```

ANS -

[ ]:
```
Nominal encoding is a technique used when the features are nominal (do not have␣
 ↪any order). In one hot encoding, for each level of a
categorical feature, we create a new variable. Each category is mapped with a␣
 ↪binary variable containing either 0 or 1. Here, 0 represents the
absence, and 1 represents the presence of that category 1.

So if you have two categorical columns in your dataset and you use nominal␣
 ↪encoding to transform them, you will create two new columns.
```

[ ]:

[ ]:
```
Q6. You are working with a dataset containing information about different types␣
 ↪of animals, including their
species, habitat, and diet. Which encoding technique would you use to transform␣
 ↪the categorical data into
a format suitable for machine learning algorithms? Justify your answer.
```

ANS -

[ ]:

[ ]:
```
Q7.You are working on a project that involves predicting customer churn for a␣
 ↪telecommunications
```

```
company. You have a dataset with 5 features, including the customer's gender,␣
 ↪age, contract type,
monthly charges, and tenure. Which encoding technique(s) would you use to␣
 ↪transform the categorical
data into numerical data? Provide a step-by-step explanation of how you would␣
 ↪implement the encoding.
```

ANS -

[ ]:
```
There are several encoding techniques for categorical data. The most common␣
 ↪ones are One-Hot Encoding, Dummy Encoding, and Ordinal Encoding.

One-Hot Encoding is used when there is no ordinal relationship between the␣
 ↪categories. In this technique, each category is converted into a
binary vector of length equal to the number of categories. Each vector has a␣
 ↪single 1 and the rest of the values are 0s.

Dummy Encoding is similar to One-Hot Encoding but it creates one less column␣
 ↪than the number of categories. This technique is used to avoid
multicollinearity in the dataset.

Ordinal Encoding is used when there is an ordinal relationship between the␣
 ↪categories. In this technique, each category is assigned an integer
value based on its order.
```

[1]:
```
import pandas as pd

# create a dataframe with categorical data
df = pd.DataFrame({'gender': ['Male', 'Female', 'Male', 'Female'],
                   'contract_type': ['Month-to-month', 'One year', 'Two year',␣
 ↪'Month-to-month']})

# apply one-hot encoding
df_encoded = pd.get_dummies(df)
```

[4]:
```
df_encoded
```

[4]:

	gender_Female	gender_Male	contract_type_Month-to-month	\
0	0	1	1	
1	1	0	0	
2	0	1	0	
3	1	0	1	

	contract_type_One year	contract_type_Two year
0	0	0
1	1	0
2	0	1

3	0	0