assignment=41

August 1, 2023

1 Q1. What is the difference between Ordinal Encoding and Label Encoding? Provide an example of when you might choose one over the other

*Ordinal encoding and label encoding are two common methods for encoding categorical variables

*Ordinal encoding:
-It is a technique where the categories of a categorical variable are assigned integers base
Example
-If we have a categorical variable "Size" with categories "Small", "Medium" and "Large", we

*Label encoding:
-It is a technique where the categories of a categorical variable are assigned a unique int
Example
-If we have a categorical variable "Color" with categories "Red", "Green" and "Blue", we mi
-One major difference between the two encoding methods is that ordinal encoding considers t
-Therefore, ordinal encoding is suitable for variables with a clear order or ranking, such

*Example:
-If we have a dataset of students with their grades in different subjects, we can use ordinal
-If we have a dataset of animals with their colors, we can use label encoding to represent th

2 Q2. Explain how Target Guided Ordinal Encoding works and provide an example of when you might use it in a machine learning project.

*Target guided ordinal encoding is a method of encoding categorical variables that combines th
* It involves encoding the categories of a categorical variable based on the mean of the target

*The process involves the following steps:
1. Group the data by the categorical variable and calculate the mean of the target variable
2. Sort the categories based on the mean of the target variable in ascending or descending

```
   3.  Assign an integer value to each category based on their rank.
   4.  Replace the original categorical variable with the encoded variable.
```

```
 *Example:
     -Consider a dataset of credit card applications, where we want to predict whether an appl
```

3 Q3. Define covariance and explain why it is important in statistical analysis. How is covariance calculated?

```
 *Covariance is a measure of the relationship between two random variables.
 *It describes the degree to which the two variables move together or apart from their expecte
 *A positive covariance indicates that the two variables tend to increase or decrease togethe


 *  Covariance is an important concept in statistical analysis because it helps to understand
 Example
*If we are trying to predict the price of a house based on its size, we would want to know whe
*By calculating the covariance, we can see if the two variables move together or in opposite d


*Covariance is calculated as follows:
 cov(X,Y) = E[(X-E[X])(Y-E[Y])]
     where
      -X and Y are the two random variables
      -E[X] and E[Y] are their expected values
      -cov(X,Y) is the covariance between them.
```

4 Q4. For a dataset with the following categorical variables: Color (red, green, blue), Size (small, medium, large), and Material (wood, metal, plastic), perform label encoding using Python's scikit-learn library. Show your code and explain the output.

[1]:

```python
from sklearn.preprocessing import LabelEncoder
import pandas as pd

# Create a sample dataset
data = {'Color': ['red', 'green', 'blue', 'red', 'green', 'blue'],
        'Size': ['small', 'medium', 'large', 'small', 'medium', 'large'],
        'Material': ['wood', 'metal', 'plastic', 'wood', 'metal', 'plastic']}
df = pd.DataFrame(data)

# Initialize the LabelEncoder
le = LabelEncoder()
```

```
# Apply label encoding on the categorical variables
df['Color_encoded'] = le.fit_transform(df['Color'])
df['Size_encoded'] = le.fit_transform(df['Size'])
df['Material_encoded'] = le.fit_transform(df['Material'])

# Show the encoded dataset
df
```

```
[1]:    Color    Size Material  Color_encoded  Size_encoded  Material_encoded
     0    red   small     wood              2             2                 2
     1  green  medium    metal              1             1                 0
     2   blue   large  plastic              0             0                 1
     3    red   small     wood              2             2                 2
     4  green  medium    metal              1             1                 0
     5   blue   large  plastic              0             0                 1
```

```
*The LabelEncoder encodes each category of the categorical variables with a unique integer. I
 *The encoded variables are added to the original dataset as new columns with the suffix '_enc
*Label encoding is a simple and effective method of converting categorical variables into nume
```

5 Q5. Calculate the covariance matrix for the following variables in a dataset: Age, Income, and Education level. Interpret the results.

```
[2]: import numpy as np
     import pandas as pd

     # Create a sample dataset
     age = np.random.normal(40, 10, 100)
     income = np.random.normal(50000, 10000, 100)
     education = np.random.normal(12, 2, 100)
     df = pd.DataFrame({'Age': age, 'Income': income, 'Education': education})

     # Calculate the covariance matrix
     cov_matrix = np.cov(df.T)

     # Show the covariance matrix
     print(cov_matrix)
```

```
[[ 8.11596119e+01 -8.07364736e+02  4.11478618e+00]
 [-8.07364736e+02  1.15462575e+08  3.76845946e+02]
 [ 4.11478618e+00  3.76845946e+02  3.80386055e+00]]
```

6 Q6. You are working on a machine learning project with a dataset containing several categorical variables, including “Gender” (Male/Female), “Education Level” (High School/Bachelor’s/Master’s/PhD), and “Employment Status” (Unemployed/Part-Time/Full-Time). Which encoding method would you use for each variable, and why?

```
*For the categorical variable "Gender" (Male/Female), I would use binary encoding, as there a
*In this case, we could represent "Male" as 0 and "Female" as 1.
```

```
 *For the categorical variable "Education Level" (High School/Bachelor's/Master's/PhD), I woul
  *In this case, we could assign "High School" a value of 1, "Bachelor's" a value of 2, "Maste
```

```
 *For the categorical variable "Employment Status" (Unemployed/Part-Time/Full-Time), I would u
  *In this case, we would create three new binary features, one for each category.
```

```
 *Note that the choice of encoding method may also depend on the specific machine learning alg
```

7 Q7. You are analyzing a dataset with two continuous variables, “Temperature” and “Humidity”, and two categorical variables, “Weather Condition” (Sunny/Cloudy/Rainy) and “Wind Direction” (North/South/East/West). Calculate the covariance between each pair of variables and interpret the results.

[3]:

```python
import numpy as np
import pandas as pd

# Create a sample dataset
temperature = np.random.normal(25, 5, 100)
humidity = np.random.normal(50, 10, 100)
weather_condition = np.random.choice(['Sunny', 'Cloudy', 'Rainy'], 100)
wind_direction = np.random.choice(['North', 'South', 'East', 'West'], 100)
df = pd.DataFrame({'Temperature': temperature, 'Humidity': humidity, 'Weather␣
 ↪Condition': weather_condition, 'Wind Direction': wind_direction})

# Calculate the covariance matrix
cov_matrix = np.cov(df[['Temperature', 'Humidity']].T)
print('Covariance between Temperature and Humidity:\n', cov_matrix)
```

```
cov_matrix = pd.crosstab(df['Weather Condition'], df['Wind Direction'],␣
 ↪normalize='index').values
print('Covariance between Weather Condition and Wind Direction:\n', cov_matrix)
```

```
Covariance between Temperature and Humidity:
 [[ 25.87404836   5.2789205 ]
 [  5.2789205  105.65327512]]
Covariance between Weather Condition and Wind Direction:
 [[0.08       0.36       0.36       0.2       ]
 [0.18918919 0.24324324 0.37837838 0.18918919]
 [0.26315789 0.26315789 0.21052632 0.26315789]]
```

```
[ ]:
```