21 March Assignment

May 28, 2023

[ ]:

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

ANS -

[ ]:

```
Ordinal Encoding and Label Encoding are both techniques used to convert categorical data into numerical data. The difference between them is
that Ordinal Encoding is used when the categorical data has an inherent order, while Label Encoding is used when the categorical data has no
inherent order.

For example, if we have a dataset with a column called "Size" that contains values "Small", "Medium", and "Large", we can use Ordinal Encoding
to convert these values into 1, 2, and 3 respectively. However, if we have a column called "Color" that contains values "Red", "Green", and
"Blue", we can use Label Encoding to convert these values into 1, 2, and 3 respectively.

In general, Ordinal Encoding should be used when the categorical data has an inherent order, such as in the case of clothing sizes
(Small < Medium < Large). On the other hand, Label Encoding should be used when the categorical data has no inherent order, such as in the
case of colors (Red != Green != Blue).
```

[ ]:

```

```

[ ]:

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

ANS -

[ ]:

```
Target Guided Ordinal Encoding is a technique used to encode categorical data in machine learning projects. It is a type of ordinal encoding
that uses the target variable to encode categorical data. In this technique, the labels are ordered based on their target variable.
```

```
The technique replaces the categorical data with a blend of the posterior probability of the target given a particular categorical value and
the prior probability of the target over all the training data.

For example, let’s say you have a dataset with a categorical feature called "City" and you want to predict salaries. You can use Target Guided
Ordinal Encoding to encode the "City" feature by calculating the mean salary for each city and then ranking them based on their mean salary.
```

[ ]:

[ ]:

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

ANS -

[ ]:

```
Covariance is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other
words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable. This is the
property of a function of maintaining its form when the variables are linearly transformed. Covariance is measured in units, which are
calculated by multiplying the units of the two variables.

Covariance can have both positive and negative values. Based on this, it has two types:

Positive Covariance
Negative Covariance

Covariance is a statistical tool used to determine the relationship between the movements of two random variables. It measures the joint
variability of two random variables and can take any positive or negative value. A positive covariance means that the two variables tend to
move in the same direction, while a negative covariance means that they move in opposite directions. Covariance is different from the
correlation coefficient, which measures the strength of a correlative relationship.

The formula for covariance is:

Cov(X,Y) = E[(X - E[X])(Y - E[Y])]

where X and Y are random variables, E[X] and E[Y] are their expected values.
```

[ ]:

[ ]:

```
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.
```

ANS-

[28]:

```
from sklearn.preprocessing import LabelEncoder

# Create a dictionary of the categorical variables
data = {'Color': ['red', 'green', 'blue'], 'Size': ['small', 'medium',␣
  ↪'large'], 'Material': ['wood', 'metal', 'plastic']}

# Create an instance of the LabelEncoder class
le = LabelEncoder()

# Encode the categorical variables
for col in data:
    data[col] = le.fit_transform(data[col])

print(data)
```

```
{'Color': array([2, 1, 0]), 'Size': array([2, 1, 0]), 'Material': array([2, 0,
1])}
```

[ ]:

```

```

[ ]:

```
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?
```

ANS -

[ ]:

```
There are several ways to encode categorical variables in machine learning. The␣
  ↪three most common methods are:

1.Integer Encoding: Where each unique label is mapped to an integer.

2.One Hot Encoding: Where each label is mapped to a binary vector.

3.Learned Embedding: Where a distributed representation of the categories is␣
  ↪learned.
```

```
For the "Gender" variable, you can use integer encoding since there are only␣
 ↪two categories (Male/Female). For "Education Level," you can use
one hot encoding since there are multiple categories with no particular order␣
 ↪(High School/Bachelor’s/Master’s/PhD). For "Employment Status,"
you can also use one hot encoding since there are multiple categories with no␣
 ↪particular order (Unemployed/Part-Time/Full-Time).
```

[ ]:

```
```

[ ]:

```
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.
```

ANS -

[ ]:

```
The covariance between two variables is a measure of how much they vary␣
 ↪together. It is calculated by taking the product of the difference
between each variable and its mean, then averaging over all observations.␣
 ↪Covariance can be calculated between two continuous variables or
between a continuous and a categorical variable. However, it is not meaningful␣
 ↪to calculate covariance between two categorical variables.
 "Temperature" and "Humidity", and two categorical variables "Weather␣
 ↪Condition" (Sunny/Cloudy/Rainy) and "Wind Direction"
(North/South/East/West). You can calculate the covariance between each pair of␣
 ↪variables using the following formula.
Covariance(X,Y) = (1/n) * Σ(xi - x̄)(yi - ȳ)

where X and Y are the two variables, xi and yi are the individual observations,␣
 ↪x̄ and ȳ are the means of X and Y respectively,
and n is the number of observations.
```