Feature Engineering

Papangkorn Inkeaw, Ph.D.

Feature Combining

Lab 4

Combining Features with Statistical Operations

Dataset: Breast Cancer Wisconsin (Diagnostic) Data Set (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))

- Import libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.datasets import load_breast_cancer
```

- Load the Breast Cancer dataset from scikit-learn

```
data = load_breast_cancer()
print(data.DESCR)
df = pd.DataFrame(data.data, columns=data.feature_names)
df['target'] = data.target
```

Combining Features with Statistical Operations

- Creating a list with the subset of features to which we will apply:

```
features = ['mean smoothness', 'mean compactness','mean concavity', 'mean concave points',
    'mean symmetry']
```

- Create a new feature with the sum of the selected variables:

```
df['added_features'] = df[features].sum(axis=1)
```

- Derive a new feature using the product of the selected features:

```
df['prod_features'] = df[features].prod(axis=1)
```

- Obtain a new feature corresponding to the mean value of the variables:

```
df['mean_features'] = df[features].mean(axis=1)
```

- Capture the standard deviation of the features in a new variable:

```
df['std_features'] = df[features].std(axis=1)
```

Combining Features with Statistical Operations

- Find the maximum value across the selected variables:

```
df['max_features'] = df[features].max(axis=1)
```

- Find the minimum value across the selected features:

```
df['min_features'] = df[features].min(axis=1)
```

- Find the minimum value across the selected features:

```
df['min_features'] = df[features].min(axis=1)
```

- Create a violin plot of the newly created feature.

```
sns.violinplot(x="target", y="added_features", data=df)
plt.title('Added Features')
plt.show()
```

Combining Features with Mathematical Functions

- Capture the difference between two features in a new variable:

  ```
  df['difference'] = df['worst compactness'].sub(df['mean compactness'])
  ```

- Create a new feature with the ratio between two variables:

  ```
  df['quotient'] = df['worst radius'].div(df['mean radius'])
  ```

- Make a list of the features we want to compare:

  ```
  features = ['mean smoothness', 'mean compactness', 'mean concavity', 'mean concave points',
      'mean symmetry']
  ```

- Make a list of the features we want to aggregate:

  ```
  worst_f = ['worst smoothness', 'worst compactness’, 'worst concavity', 'worst concave points’,
      'worst symmetry’]
  ```

- Create a new feature with the sum of the worst features:

  ```
  df['worst'] = df[worst_f].sum(axis=1)
  ```

- Obtain the ratio between each one of the feature and the feature created in the previous step:

  ```
  df[features] = df[features].div(df['worst'], axis=0)
  ```

Combining Features with Mathematical Functions

- Obtain the ratio between each one of the feature and the feature created in the previous step:

```
df[features] = df[features].div(df['worst'], axis=0)
```

Your work!

1. Download the SARS-CoV-2 variants in Thailand from https://data.go.th/dataset/sars-cov-2-variants
2. Investigate the dataset details
3. Construct new features (one or more) by combining some features with statistical operations. Describe why you perform the statistical operation on the selected features in a comment.
4. Construct new features (one or more) by combining two features with mathematical functions. Describe the meaning of the new feature in a comment.
5. Submit your program to the assignment submission system (http://hw.cs.science.cmu.ac.th/).

Note:

- Put your name and student ID in the first cell using comment tag.
- Name your python notebook file with the pattern Lab_04_XXXXXXXXX.py (XXXXXXXXX is your student ID)

References & Study Resources

- Soledad Galli. (2020). *Python Feature Engineering Cookbook*. Packt Publishing.
- https://archive-beta.ics.uci.edu/ml/datasets/credit+approval
- http://lib.stat.cmu.edu/datasets/boston
- https://archive.ics.uci.edu/ml/datasets/Adult