Machine Learning

ML - Data Feature Selection

Computer & Health Informatics Department

Introduction

- In the previous chapter, we have seen in detail how to preprocess and prepare data for machine learning.
- In this chapter, let us understand in detail data feature selection and various aspects involved in it.

Importance of Data Feature Selection

- The performance of machine learning model is directly proportional to the data features used to train it.
- The performance of ML model will be affected negatively if the data features provided to it are irrelevant. On the other hand, use of relevant data features can increase the accuracy of ML model especially linear and logistic regression.

Importance of Data Feature Selection

- Now the question arise that what is automatic feature selection?
- It may be defined as the process with the help of which we select those features in data that are most relevant to the output or prediction variable in which we are interested. It is also called attribute selection.

Importance of Data Feature Selection

- The following are some of the benefits of automatic feature selection before modeling the data :
- Performing feature selection before data modeling will reduce the overfitting.
- Performing feature selection before data modeling will increases the accuracy of ML model.
- Performing feature selection before data modeling will reduce the training time

Feature Selection Techniques

- The followings are automatic feature selection techniques that can use to model ML data in Python :

  - Univariate Selection
  - Recursive Feature Elimination
  - Principal Component Analysis (PCA)

Univariate Selection

- This feature selection technique is very useful in selecting those features, with the help of statistical testing, having strongest relationship with the prediction variables.
- Can implement univariate feature selection technique with the help of SelectKBest0class of scikit-learn Python library.

Example

- In this example, we will use Pima Indians Diabetes dataset to select 4 of the attributes having best features with the help of chi-square statistical test.

Example

```
from pandas import read_csv
from numpy import set_printoptions
from sklearn.feature_selection import SelectKBest
 from sklearn.feature_selection import chi2
path = r'C:\pima-indians-diabetes.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
dataframe = read_csv(path, names=names)
array = dataframe.values
```

Example

- Next, we will separate array into input and output components :

```
X = array[:,0:8]
Y = array[:,8]
```

- The following lines of code will select the best features from dataset

```
test = SelectKBest(score_func=chi2, k=4)
fit = test.fit(X,Y)
```

Example

- We can also summarize the data for output as per our choice.
- Here, we are setting the precision to 2 and showing the 4 data attributes with best features along with best score of each attribute

Example

```
set_printoptions(precision=2)
print(fit.scores_)
featured_data = fit.transform(X)
print ("\nFeatured data:\n", featured_data[0:4])
```

Example

Output

[111.52 1411.89 17.61 53.11 2175.57 127.67 5.39 181.3]

Featured data:

[[148. 0. 33.6 50.]

[85. 0. 26.6 31.]

[183. 0. 23.3 32.]

[89. 94. 28.1 21.]]

Recursive Feature Elimination

Thank you