Rakamin Academy

Virtual Internship Experience

id/x partners

Automate ML Workflow with Pipeline

Disclaimer

Outline

1. Otomasi Workflow ML
2. Pipeline Data Preparation dan Pemodelan
3. Pipeline Ekstraksi Fitur dan Pemodelan

Otomatisasi Workflow
Machine Learning

Dalam praktik Machine Learning, terdapat suatu standard workflow yang dapat diotomatisasi. Disebut sebagai standar karena karena prosedur tersebut mampu mengatasi permasalahan umum yang biasa terjadi pada aplikasi Machine Learning.

Package Scikit-learn pada Python memiliki fitur yang bisa membantu dalam pembuatan Pipeline untuk melakukan otomatisasi workflow Machine Learning.

Data Preparation dan Modeling

Salah satu tahap penting dalam Pipeline Data Preparation adalah mencegah terjadinya data leakage, yaitu ketika ada informasi dari training set yang “bocor” ke test ataupun validation set.

Berikut contoh pipeline untuk tahap Data Preparation dan Modeling terdiri dari langkah-langkah berikut:

1. Standarisasi Data
2. Model Linear Discriminant Analysis

workflow data preparation dan model evaluation

```
# Create a pipeline that standardizes the data then creates a model
from pandas import read_csv
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
# load data
filename = 'pima-indians-diabetes.data.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
dataframe = read_csv(filename, names=names)
array = dataframe.values
X = array[:,0:8]
Y = array[:,8]
# create pipeline
estimators = []
estimators.append(('standardize', StandardScaler()))
estimators.append(('lda', LinearDiscriminantAnalysis()))
model = Pipeline(estimators)
# evaluate pipeline
kfold = KFold(n_splits=10, random_state=7)
results = cross_val_score(model, X, Y, cv=kfold)
print(results.mean())
```

Feature Extraction dan Modeling

Selain proses Data Preparation, proses Feature Extraction juga rentan terhadap fenomena Data Leakage. Oleh karena itu, pipeline yang baik harus dipastikan agar tidak terjadi hal tersebut. Contoh langkah-langkah yang dapat diterapkan dalam Pipeline Feature Extraction dan Modeling:

1. Feature Extraction dengan PCA
2. Feature Extraction dengan Statistical Selection
1. Feature Union
2. Membuat model Logistic Regression

workflow data preparation dan model evaluation

```
# Create a pipeline that extracts features from the data then creates a model
from pandas import read_csv
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.pipeline import Pipeline
from sklearn.pipeline import FeatureUnion
from sklearn.linear_model import LogisticRegression
from sklearn.decomposition import PCA
from sklearn.feature_selection import SelectKBest
# load data
filename = 'pima-indians-diabetes.data.csv'
names = ['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class']
dataframe = read_csv(filename, names=names)
array = dataframe.values
X = array[:,0:8]
Y = array[:,8]
# create feature union
features = []
features.append(('pca', PCA(n_components=3)))
features.append(('select_best', SelectKBest(k=6)))
feature_union = FeatureUnion(features)
# create pipeline
estimators = []
estimators.append(('feature_union', feature_union))
estimators.append(('logistic', LogisticRegression()))
model = Pipeline(estimators)
# evaluate pipeline
kfold = KFold(n_splits=10, random_state=7)
results = cross_val_score(model, X, Y, cv=kfold)
print(results.mean())
```

Bagaimana penerapannya di ID/X Partners?


Thank You!
id/x partners