Московский Государственный Технический Университет имени Н.Э.Баумана

Факультет Информатика и системы управления

Кафедра ИУ-5 «Системы обработки информации и управления»

Отчёт по рубежному контролю № 1

По дисципление

«Методы Машинного Обучения»

Группа ИУ5И-23М

Ли Хао

Москва 2024г

Номер варианта: 17

Номер задачи №1: 17

Для набора данных проведите нормализацию для одного (произвольного) числового признака с использованием преобразования Йео-Джонсона (Yeo-Johnson transformation).

Загрузить данные

```
import numpy as np
import pandas as pd
from sklearn.preprocessing import PowerTransformer

# Создание примерного набора данных
np.random.seed(0)
data = pd.read_csv('Movies.csv')

data.head()
```

	index	Title	Release Date	Year	Description	URL	Rating	Runtime	Genres	Votes	Directors	Series	Order
0	0	101 Dalmatians	18-11-1996	1996	NaN	https://www.imdb.com/title/tt0115433/	5.7	103.0	Adventure, Comedy, Crime, Family	98439.0	Stephen Herek	101 Dalmatians	1
1	1	102 Dalmatians	22-11-2000	2000	NaN	https://www.imdb.com/title/tt0211181/	4.9	100.0	Adventure, Comedy, Family	33823.0	Kevin Lima	101 Dalmatians	2
2	2	12 Rounds	19-03-2009	2009	NaN	https://www.imdb.com/title/tt1160368/	5.6	108.0	Action, Crime, Thriller	26828.0	Renny Harlin	12 Rounds	1
3	3	12 Rounds 2: Reloaded	04-06-2013	2013	NaN	https://www.imdb.com/title/tt2317524/	5.3	95.0	Action, Adventure, Thriller	5141.0	Roel Reiné	12 Rounds	2
4	4	21 Jump Street	12-03-2012	2012	NaN	https://www.imdb.com/title/tt1232829/	7.2	109.0	Action, Comedy, Crime	498876.0	Christopher Miller, Phil Lord	21 Jump Street	1

后续步骤: 查看推荐的图表

Yeo-Johnson transformation

```
df = pd.DataFrame(data)

# Выбор одного числового признака для нормализации
feature = df['Runtime']

# Применение преобразования Йео-Джонсона
pt = PowerTransformer(method='yeo-johnson')
feature_transformed = pt.fit_transform(feature.values.reshape(-1, 1))

# Результаты до и после преобразования
df['Runtime_transformed'] = feature_transformed

df[['Runtime', 'Runtime_transformed']].head()
```

	Runtime	Runtime_transformed
0	103.0	-0.058320
1	100.0	-0.254277
2	108.0	0.239961
3	95.0	-0.613542
4	109.0	0.295760

Номер задачи №2: 37

Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectPercentile для 5% лучших признаков, и метод, основанный на взаимной информации.

Загрузить данные

```
import numpy as np
import pandas as pd
from sklearn.preprocessing import PowerTransformer

# Создание примерного набора данных
np.random.seed(0)
data = pd.read_csv('Movies.csv')

data.head()
```

	index	Title	Release Date	Year	Description	URL	Rating	Runtime	Genres	Votes	Directors	Series	Order
0	0	101 Dalmatians	18-11-1996	1996	NaN	https://www.imdb.com/title/tt0115433/	5.7	103.0	Adventure, Comedy, Crime, Family	98439.0	Stephen Herek	101 Dalmatians	1
1	1	102 Dalmatians	22-11-2000	2000	NaN	https://www.imdb.com/title/tt0211181/	4.9	100.0	Adventure, Comedy, Family	33823.0	Kevin Lima	101 Dalmatians	2
2	2	12 Rounds	19-03-2009	2009	NaN	https://www.imdb.com/title/tt1160368/	5.6	108.0	Action, Crime, Thriller	26828.0	Renny Harlin	12 Rounds	1
3	3	12 Rounds 2: Reloaded	04-06-2013	2013	NaN	https://www.imdb.com/title/tt2317524/	5.3	95.0	Action, Adventure, Thriller	5141.0	Roel Reiné	12 Rounds	2
4	4	21 Jump Street	12-03-2012	2012	NaN	https://www.imdb.com/title/tt1232829/	7.2	109.0	Action, Comedy, Crime	498876.0	Christopher Miller, Phil Lord	21 Jump Street	1

后续步骤: 查看推荐的图表

Использование класса SelectPercentile для 5% лучших признаков, и метод, основанный на взаимной информации.

```
import numpy as np
import pandas as pd
from sklearn.feature_selection import SelectPercentile, mutual_info_regression

# Создание примерного набора данных
np.random.seed(0)
data = {
    'feature1': np.random.normal(loc=0, scale=1, size=100),    # нормальное распределение
    'feature2': np.random.exponential(scale=1, size=100),    # экспоненциальное распределение
    'feature3': np.random.uniform(low=-1, high=1, size=100),    # равномерное распределение
    'target': np.random.normal(loc=0, scale=1, size=100)    # целевая переменная
}

df = pd.DataFrame(data)

# Определение признаков и целевой переменной
X = df.drop(columns='target')
y = df['target']

# Применение SelectPercentile с использованием mutual_info_regression
selector = SelectPercentile(mutual_info_regression, percentile=5)
X_selected = selector.fit_transform(X, y)

# Получение маски отобранных признаков
selected_features = X.columns[selector.get_support()]

selected_features
```

```
Index(['feature3'], dtype='object')
```