BÀI THỰC HÀNH SỐ 4+5

Tiền xử lý dữ liệu

Một số hàm thường sử dụng để tiền xử lý dữ liệu

- Xóa tất cả các dòng có dữ liệu thiếu:

  df_no_missing = df.dropna()

  Xóa các dòng có nhiều hơn 2 giá trị thiếu.

  df.dropna(thresh=2)

- Xóa cột có giá trị rỗng

  df.dropna(axis=1, how='all')

- Thay thế giá trị rỗng bằng một giá trị nào đó sử dụng lệnh

  df.fillna(giatrithaythe)

df2['one'].fillna('missing') điền những ô có giá trị NaN bằng từ missing cho cột one

Thay tất cả các giá trị thiếu bằng giá trị trung bình của cột tương tứng.
df.fillna(df.mean())
Điền giá trị thiếu của cột A với giá trị trung bình
df['A'].fillna(df['A']. mean(),
Thay trực tiếp vào trường dữ liệu thì dùng lệnh sau:
df['A'].fillna(df['A']. mean(), inplace=True)

Ví dụ Thay các giá trị thiếu của trường Giữa kỳ bằng giá trị trung bình

df["Giữa kỳ "].fillna(df["Giữa kỳ "].mean(), inplace=True)

Điền giá trị thiếu của cột A bằng giá trị thường xuyên xuất hiện ở cột A

df['A'].fillna(df['A'].mode())

Đếm số mẫu dữ liệu trong mỗi thuộc tính thỏa mãn một điều kiện nào đó

ví dụ: đếm những sinh viên có điểm Giữa kỳ =0

c=((df[['Giữa kỳ']]=0).sum())

Ví dụ xóa các cột dữ liệu có số lỗi >=50%

loans_2007 = pd.read_csv('lending_club_loans.csv', skiprows=1, low_memory=False)
half_count = len(loans_2007) / 2
loans_2007 = loans_2007.dropna(thresh=half_count,axis=1) # Drop any column with more than 50% missing values
loans_2007 = loans_2007.drop(['url','desc'],axis=1) # These columns are not useful for our purposes

chuyển chuỗi về số:

labelEncoder = LabelEncoder()
labelEncoder.fit(train['Sex'])
train['Sex'] = labelEncoder.transform(train['Sex'])
Hoặc sử dụng hàm map

```
df['gioitinh'] = df['gioitinh'].map({'Nam':1, 'Nữ':0})
```

Chuẩn hóa minmax

print('Min max scaling')

from sklearn import preprocessing as pp

mms = pp.MinMaxScaler()

data_mms = mms.fit_transform(data)

print(data_mms)

Chuẩn hóa #z score (standard score): $z = (x - \mu) / \sigma$

from scipy.stats import zscore

Lọc lấy trường Toán trong dữ liệu

df_toan=data['Toan']

z_score_toan=zscore(df_toan) #Tính z-score

z_score_toan #xuất ra màn hình

Bài 1: Sử dụng file dữ liệu JOB_INFO.xlsx thực hiện:

- Đọc file dữ liệu
- Hiển thị dữ liệu
- Hiển thị giá tri null cho từng trường
- Hiển thị tất cả những dòng có giá trị thiếu
- Hiển thị những trường có giá trị null và số giá trị null
- Xóa những dòng có giá trị null
- Thay thế giá trị thiếu của trường age bằng giá trị trung bình của trường age
- Ghi ra file csv hoặc excel

Thực hành theo bài mẫu sau và cho nhận xét

In [2]:

```
import pandas as pd
import numpy as np
data=pd.read_excel('d:\data\JOB_INFO.xlsx')
data
```

Out[2]:

	ENO	Age	Dept	Gender	Skill	WorkingYear	Salary	TrainedYear	OverseaProject
0	E001	29.0	HR	Female	SQL	3.0	825	3.0	No
1	E002	39.0	IT	Male	Java	7.0	1450	7.0	Yes
2	E003	NaN	IT	Male	SQL	6.0	1236	NaN	Yes
3	E004	38.0	HR	Female	C#	6.0	1324	5.0	No
4	E005	28.0	IT	Male	R	2.0	895	3.0	No

In [3]:
```
#hiển thị số giá trị null của từng trường
data.isnull().sum()
```

Out[3]:
```
ENO              0
Age              2
Dept             2
Gender           1
Skill            0
WorkingYear      1
Salary           0
TrainedYear      3
OverseaProject   0
dtype: int64
```

In [4]:
```
#Hiển thị tất cả những dòng có giá trị thiếu
mask=False
for col in data.columns: mask = mask | data[col].isnull()
datanulls = data[mask]
datanulls
```

Out[4]:

	ENO	Age	Dept	Gender	Skill	WorkingYear	Salary	TrainedYear	OverseaProject
2	E003	NaN	IT	Male	SQL	6.0	1236	NaN	Yes
6	E007	27.0	NaN	Male	SQL	2.0	1154	5.0	No
7	E008	28.0	MA	Female	R	NaN	1356	6.0	Yes

In [6]:
```
#Hiển thị những trường có giá trị null và số giá trị null
null_columns=data.columns[data.isnull().any()]
data[null_columns].isnull().sum()
```

Out[6]:
```
Age            2
Dept           2
Gender         1
WorkingYear    1
TrainedYear    3
dtype: int64
```

In [7]:
```
#xóa những dòng có giá trị null
data1=data.copy() #copy dữ liệu vào data1 rồi xóa
data1.dropna(inplace=True)
data1.isnull().sum()
```

Out[7]:
```
ENO              0
Age              0
Dept             0
Gender           0
Skill            0
WorkingYear      0
Salary           0
TrainedYear      0
OverseaProject   0
```

In [6]:
```
#Hiển thị những trường có giá trị null và số giá trị null
null_columns=data.columns[data.isnull().any()]
data[null_columns].isnull().sum()
```

Out[6]:
```
Age            2
Dept           2
Gender         1
WorkingYear    1
TrainedYear    3
dtype: int64
```

In [7]:
```
#xóa những dòng có giá trị null
data1=data.copy() #copy dữ liệu vào data1 rồi xóa
data1.dropna(inplace=True)
data1.isnull().sum()
```

Out[7]:
```
ENO              0
Age              0
Dept             0
Gender           0
Skill            0
WorkingYear      0
Salary           0
TrainedYear      0
OverseaProject   0
```

```
In [8]: #thay thế giá trị thiếu của trường age bằng giá trị trung bình của trường age
        data2=data.copy()
        data2= data2.fillna(data2.mean())
        #kiểm tra
        data2.isnull().sum()
```

```
Out[8]: ENO              0
        Age              0
        Dept             2
        Gender           1
        Skill            0
        WorkingYear      0
        Salary           0
        TrainedYear      0
        OverseaProject   0
        dtype: int64
```

```
In [9]: #ghi ra file csv hoặc excel
        data2.to_csv('D:/DataMissingOK.csv')
        data2.to_excel('D:/DataMissingOK.xlsx')
```

Tự làm:

- Thay thế giá trị thiếu của trường Gender và Skill bằng giá trị xuất hiện thường xuyên nhất.
- Thay giá trị thiếu của trường Dept bằng giá trị “Khongbiet”
- Ghi ra file Dlht.csv

Bài 2: Chuẩn hóa dữ liệu

Một số hàm hay sử dụng:

Đổi dữ liệu chuỗi thành 0,1

Sử dụng hàm .map()

Ví dụ chuyển trường Gender có giá trị Male và Female thành 1, 0 như sau:

df['Gender_Cat'] = df['Gender'].map({'Male':1, 'Female':0})

Chuyển cột dữ liệu nhiều giá trị về dạng 0, 1

get_dummies(data=df, columns = ['tên trường']

kết quả mỗi giá trị trong tên trường được chuyển thành 1 cột có giá trị 0, 1

Ví dụ: Chuyển trường skill về dạng 0, 1

df_dummies = pd.get_dummies(data=df, columns = ['Skill'])

Phân dữ liệu thành các mức khái niệm

Sử dụng hàm .cut(tên trường,số bin (mức),labels=[danh sách các nhãn])

Số bin ta có thể cho 1 số cụ thể hoặc gán các mức giá trị

Ví dụ: cắt trường Tuổi ra thành 4 mức khái niệm, nhi đồng, thanh niên, trung niên và người cao tuổi.

df['Tuổi_Level']=pd.cut(df['Tuổi'],4,labels=['Nhi đồng','Thanh niên','Trung niên', 'Cao tuổi'])

Ví dụ cắt trường Salary ra thành 4 mức

cut_labels = ['Low','Medium','High']

cut_bins = [0, 800, 1200, 2000] #0: min, 2000: max

df['Sal_Cat'] = pd.cut(df['Salary'], bins=cut_bins, labels=cut_labels)

Dữ liệu là file Dlht.csv bài 1

Thực hiện các yêu cầu sau: chuyển cột Gender thành dạng nhị phân

- Chuyển các giá tri của trường skill về dạng 0, 1
- Rời rạc hóa trường Salary thành 3 mức: 'Low','Medium','High'

Bài mẫu

import numpy as np

import pandas as pd

df=pd.read_csv(‘Dlht.csv’)

df

chuyển cột giới tính thành dạng nhị phân

df['Gender_Cat'] = df['Gender'].map({'Male':1, 'Female':0})

#chuyển trường skill về dạng 0, 1 (trường này có nhiều giá trị nên chuyển mỗi giá trị thành 1 cột nhận giá trị 0,1 nên sử dụng hàm get_dummies()

df_dummies = pd.get_dummies(data=df, columns = ['Skill'])

df_dummies.head()

#rời rạc hóa trường Salary thành 3 mức: 'Low','Medium','High'

cut_labels = ['Low','Medium','High']

cut_bins = [0, 800, 1200, 2000] #0: min, 2000: max

df['Sal_Cat'] = pd.cut(df['Salary'], bins=cut_bins, labels=cut_labels)

Tự làm:

- Chuyển trường OverseaProject về dạng 0, 1 với {'Yes':1, 'No':0}
- Xóa trường Gender, OverseaProject
- Rời rạc hóa trường TrainedYear thành các khái niệm ‘College', 'University', 'Master_PhD'

  Bài 3: chuẩn hóa dữ liệu:

  Thực hiện trên tệp dữ liệu EmployeeSalary.csv

  Bài mẫu:

  ```
  import pandas as pd
  data=pd.read_csv('EmployeeSalary.csv')
  data
  df=data.drop(['ID'], axis=1)
  #chuẩn hóa minmax các trường dữ liệu
  mms=MinMaxScaler()
  mms.fit(df)
  data_mms=mms.transform(df)
  ```

#đưa vào dataframe

data_mms=pd.DataFrame(data_mms, columns=['WorkingYears','Salary'])

data_mms

làm thêm rời rạc 1 trường Salary

Phần tự làm

Bài 1. Xử lý các trường hợp thiếu giá trị cho các trường toán, văn và anh văn trên tệp dữ liệu thiTHPT2018.csv theo các yêu cầu sau:

Toán thay thế bằng giá trị trung bình

Văn thay thế bằng giá trị xuất hiện thường xuyên

Anh văn thay thế bằng giá trị thường xuyên theo trường tên tỉnh.

Các trường còn lại thay tthế bằng giá trị "Khong thi"

- Xóa những cột có số gá trị thiếu >=50%
- Điền gá trị 0 cho những ô bị thiếu
- Điền giá trị NaN cho giá trị thếu
- Điền giá trị trung bình hay phổ biến nhất cho những cột có giá trị thiếu <20%

Chuẩn hóa z-score cho các môn: Hóa, lý, Địa (xóa các giá trị thiếu trước khi chuẩn hóa)

Lưu lại với tên TotnghiepOK.csv

Có thể xem file mẫu "HandlingMissingValues052020",

Bài 2. Chuẩn hóa minmaxtrên file dữ liệu maketing. Xem bài mẫu Data Normalization_Standardization_092020

Bài 3. Rời rạc hóa dữ liệu trên file dữ liệu TotnghiepOK theo các yêu cầu sau.

Chuyển trường điểm toán thành 4 mức yếu, trung bình, khá, giỏi.

Chuyển trường điểm anh văn về 3 mức: trung bình, khá, giỏi

Bài mẫu: Data_Preprocessing_Discretization_Transformation_092020

Bài 4. Rời rạc hóa các trường trong tệp dữ liệu điểm 105.csv thành 0,1 (0 cho những giá trị điểm >=6, 0 cho những giá trị còn lại) Lưu lại thành file 105moi.csv

Bai 5: Xem bài mẫu chuyendoi. Thực hiện chuyển đổi trên dữ liệu sau:

dataset = [['Milk', 'Onion', 'Apple', 'Kidney Beans', 'Eggs', 'Yogurt'],

['Banana', 'Onion', 'Apple', 'Kidney Beans', 'Eggs', 'Yogurt'],

['Milk', 'Apple', 'Kidney Beans', 'Eggs'],

['Milk', 'Ice cream', 'Corn', 'Kidney Beans', 'Yogurt'],

['Corn', 'Apple', 'Onion', 'Kidney Beans', 'Ice cream', 'Eggs']]

- Thực hiện chuyển dổi cho file dữ liệu 105.csv về dạng true, false. (true cho những cột có điểm, false cho những cột không có điểm)
- Lưu lại thành file 105OK.csv