6_data_normalization_viz_exercise

May 20, 2019

```python
import pandas as pd
import numpy as np

df = pd.io.parsers.read_csv(
    'https://raw.githubusercontent.com/rasbt/pattern_classification/master/\
    data/wine_data.csv',
     header=None,
     usecols=[0,1,2]
    )

df.columns=['Class label', 'Alcohol', 'Malic acid']

df.head()
```

```python
from sklearn import preprocessing

std_scale = preprocessing.StandardScaler().fit(
    df[['Alcohol', 'Malic acid']])
df_std = std_scale.transform(df[['Alcohol', 'Malic acid']])
df_std[:5]
```

```python
minmax_scale = preprocessing.MinMaxScaler().fit(df[['Alcohol', 'Malic acid']])
df_minmax = minmax_scale.transform(df[['Alcohol', 'Malic acid']])
df_minmax[:3]
```

```python
print('Mean after standardization:\nAlcohol={:.2f}, Malic acid={:.2f}'
        .format(df_std[:,0].mean(), df_std[:,1].mean()))
print('\nStandard deviation after standardization:\nAlcohol={:.2f}, \
        Malic acid={:.2f}'
        .format(df_std[:,0].std(), df_std[:,1].std()))
```

```python
print('Min-value after min-max scaling:\nAlcohol={:.2f}, Malic acid={:.2f}'
        .format(df_minmax[:,0].min(), df_minmax[:,1].min()))
print('\nMax-value after min-max scaling:\nAlcohol={:.2f}, Malic acid={:.2f}'
        .format(df_minmax[:,0].max(), df_minmax[:,1].max()))
```

```python
%matplotlib inline
```

```python
In [ ]: from matplotlib import pyplot as plt

        def plot():
            plt.figure(figsize=(8,6))

            plt.scatter(df['Alcohol'], df['Malic acid'],
                    color='green', label='input scale', alpha=0.5)

            plt.scatter(df_std[:,0], df_std[:,1], color='red',
                    label='Standardized ', alpha=0.3)

            plt.scatter(df_minmax[:,0], df_minmax[:,1],
                    color='blue', label='min-max scaled [min=0, max=1]', alpha=0.3)

            plt.title('Alcohol and Malic Acid content of the wine dataset')
            plt.xlabel('Alcohol')
            plt.ylabel('Malic Acid')
            plt.legend(loc='upper left')
            plt.grid()

            plt.tight_layout()

        plot()
        plt.show()

In [ ]: fig, ax = plt.subplots(3, figsize=(6,14))

        for a,d,l in zip(range(len(ax)),
                    (df[['Alcohol', 'Malic acid']].values, df_std, df_minmax),
                    ('Input scale',
                     'Standardized ',
                     'min-max scaled [min=0, max=1]')
                    ):
            for i,c in zip(range(1,4), ('red', 'blue', 'green')):
                ax[a].scatter(d[df['Class label'].values == i, 0],
                        d[df['Class label'].values == i, 1],
                        alpha=0.5,
                        color=c,
                        label='Class %s' %i
                        )
            ax[a].set_title(l)
            ax[a].set_xlabel('Alcohol')
            ax[a].set_ylabel('Malic Acid')
            ax[a].legend(loc='upper left')
            ax[a].grid()

        plt.tight_layout()
```

```
        plt.show()
```