IEMS 351 Lab 2

September 2024

Exercises

- Step 1: Import os, math, numpy, pandas, matplotlib, scipy, and sklearn packages.

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from scipy import stats
from sklearn.model_selection import train_test_split
```

- Step 2: Load CSV file into DataFrame.
  - One example is

```
spotify_song_df = pd.read_csv("Spotify_Song_Attributes/data.csv")
```

  - See syntax on https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html
- Step 3: Show the information about the data.
  - One example is

```
print(spotify_song_df.info()) # print the information about spotify_song_df
```

- Step 4: In the lecture, we investigate the relationship between loudness and energy. Here is a quick recap.
  - Draw the scatter plot. You will use

```
plt.scatter(Z,Y)
```

    Z and Y are two numpy arrays.
    Hint: To access one column of the DataFrame called "energy" and convert it into a numpy array, we do the following:

```
Z = spotify_song_df["energy"].to_numpy() # convert a column into a numpy array
```

  - Shift the data by its mean.

```
Z_shifted = Z - np.mean(Z) # shifting Z by its mean
```

  - Draw the scatter plot for shifted data.
  - Perform the linear regression.

```
res = stats.linregress(x=Z, y=Y)
```

– Print the intercept and slope, what do you see about the intercept?

```
print(res.intercept, res.slope)
```

- Step 5: Follow the instructions in Step 4 but do not shift the data by its mean. Compared with the slope of the fitted line in step 4, what do you observe on the slope of the new fitted line?
- Step 6: Follow the instructions in Step 4 to investigate the relationship between tempo and acousticness. Here, let Y be tempo and let Z be acousticness.
- Step 7: Here, we will take a look at how to split the data into a training set and a test set.

```
Z_train, Z_test, Y_train, Y_test = train_test_split(
    Z, Y, test_size=0.33, random_state=49) # Z: feature array, Y: response array
```

- Step 8: Split the (energy, loudness) data pairs into a training set and a test set. Follow Step 5 to fit a line using the training set and then compute the mean squared prediction errors for the test set.

$$\text{MSPE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} (m(z_i) - y_i)^2,$$

where $m(z) = x_{\text{intercept}} + x_{\text{slope}} \cdot z$.

Further readings: See a package for multivariate linear regression. We will discuss it when we face high-dimensional optimization problems.