

Practical Machine Learning – NumPy Exercises



Question 1 - Numerical Analysis Exercises using NumPy – Rainfall Dataset:

The objective of the exercises below is to familiarize yourself with the use of NumPy. These exercises are mainly based on rainfall data in Cork for each month over the past half century. In the dataset folder you will find a file called `CorkRainfall.txt` and a file called `DublinRainfall.txt`. This is a **space** delimited file (note this is the default delimiter when using `np.genfromtxt ...` by default, any consecutive whitespaces act as delimiter.).

Each line of the file contains the following precipitation information pertaining to a specific month and year:

- Year
- Month (1 = Jan, 2 = Feb, 3 = March, etc.)
- Total Rainfall (Millimetres)
- Most Rainfall in a Day (Millimetres)
- Number of Rain days (A day is classified as a rain day if it has ≥ 0.2 mm rain) (Number)

Please use NumPy to answer the following questions. The objective of this task is to familiarize yourself with the operation of NumPy (there is no need to incorporate error checking).

- (i) Print out the max ‘Most Rainfall in a Day’ value and the average ‘Most Rainfall in a Day’ value for the Cork data (that is, obtain the maximum value contained in this column of data and the average value in this column of data).
- (ii) Display all unique years for which there is data in the dataset (you can use [np.unique](#)) Ask the user to select a specific year. You should then output the sum of the Rain Days column for that year (you do this by adding up the "Number of rain days" for all 12 rows pertaining to the selected year).
- (iii) Calculate the wettest month of the year in Cork based using the “Total Rainfall” value. The month that has the highest cumulative “Total Rainfall” value across all years should be classified as the wettest (please note you can use a for loop in answering this question).
- (iv) This question focuses on the Number of Rain days column. The user is asked to enter a maximum threshold value for the number of rain days. Your code should then output the percentage of the time (percentage of rows in the dataset) where the number of rain days is less than or equal to the threshold value.

For example, if a user enters a maximum threshold value of 6, then your code should output the percentage of rows where the number of rain days fell below the threshold value of 6.

- (v) Calculate the average 'total rainfall' value for the summer months (June, July and August) and the Autumn months (Sept, Oct, Nov).
- (vi) Read in the contents of the file DublinRainfall.txt into a NumPy array. Append the all rows from the Dublin array to the Cork NumPy array. Calculate the average number of raindays for the new array and write the new NumPy array to a CSV file.

Question 2 - Numerical Analysis Exercises using NumPy Bike Dataset:

For each of the following questions you will use the bike rental dataset called bike.csv.

Where possible use NumPy to answer the questions below.

The following are the details of the various fields in this dataset.

- 0. instant: record index
- 1. season : season (1:springer, 2:summer, 3:fall, 4:winter)
- 2. yr : year (0: 2011, 1:2012)
- 3. mnth : month (1 to 12)
- 4. hr : hour (0 to 23)
- 5. holiday : weather day is holiday or not (extracted from [Web Link])
- 6. weekday : day of the week
- 7. workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
- 8. + weathersit :
 - i. 1: Clear, Few clouds, Partly cloudy, Partly cloudy
 - ii. 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
 - iii. 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
 - iv. 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- 9. temp : Normalized temperature in Celsius. The values are divided to 41 (max)
- 10. atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- 11. hum: Normalized humidity. The values are divided to 100 (max)
- 12. windspeed: Normalized wind speed. The values are divided to 67 (max)
- 13. casual: count of casual users
- 14. registered: count of registered users
- 15. cnt: count of total rental bikes including both casual and registered

(i)

Calculate the average temperature value (column index 9) for the entire dataset. Note the temperature values in this column have been already normalized by dividing by 41.

(ii)

Print out the average number of casual users for all days classified as holidays as well as the average for all days classified as non-holidays. (Note holidays =1 and non-holidays = 0). Holidays attribute is stored at index 5.

(iii)

Write NumPy code that will print out the total number of casual users for each month of the year. You would expect to see an increase in the number of casual users over the summer months and a decline for the winter months.

(iv)

We will now look at the relationship between temperature and the number of rental users. Your code should work out the average number of rental users for the following temperature ranges.

- 1, 5
- 6, 10
- 11, 15
- 16, 20
- 21, 25
- 26, 30
- 31, 35
- 36, 40

Remember the temperature values specified in the file have been normalised by dividing by 41.

For temp in range 1 to 5 the mean number of total rental users was 49.2954545455
For temp in range 6 to 10 the mean number of total rental users was 73.6670630202
For temp in range 11 to 15 the mean number of total rental users was 130.681770652
For temp in range 16 to 20 the mean number of total rental users was 169.066772655
For temp in range 21 to 25 the mean number of total rental users was 211.700074516
For temp in range 26 to 30 the mean number of total rental users was 242.172678691
For temp in range 31 to 35 the mean number of total rental users was 337.473005641
For temp in range 36 to 40 the mean number of total rental users was 314.991111111

