Task and Data for HW#1

By J. H. Wang

Mar. 14, 2018

Task Description

- Goal: To perform simple statistics on numeric data
- Input: Numeric data (in the column of an open dataset)
- Output: Results of simple statistics (to be detailed later)

Input Data

- Data:
  - [**Individual household electric power consumption dataset**] from UCI Machine Learning Repository
    - About 2 million instances, 20MB (compressed) in size
  - Available at: https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption
- Format:
  - One text file consisting of lines of records
  - Each record contains 9 attributes separated by semicolons:
    Date, time, global_active_power, global_reactive_power, voltage, global_intensity, sub_metering_1, sub_metering_2, sub_metering_3

Detailed Information about Data Attributes

- 1.date: Date in format dd/mm/yyyy
- 2.time: time in format hh:mm:ss
- 3.**global_active_power**: household global minute-averaged active power (in kilowatt)
- 4.**global_reactive_power**: household global minute-averaged reactive power (in kilowatt)
- 5.**voltage**: minute-averaged voltage (in volt)
- 6.**global_intensity**: household global minute-averaged current intensity (in ampere)
- 7.sub_metering_1: energy sub-metering No. 1 (in watt-hour of active energy)
  - It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave (hot plates are not electric but gas powered)
- 8.sub_metering_2: energy sub-metering No. 2 (in watt-hour of active energy)
  - It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.
- 9.sub_metering_3: energy sub-metering No. 3 (in watt-hour of active energy)
  - It corresponds to an electric water-heater and an air-conditioner.

Tasks in the Homework

- 3 subtasks:
  - (**30pt**) (1) Output the minimum, maximum, and count of the columns: ‘global active power’, ‘global reactive power’, ‘voltage’, and ‘global intensity’
  - (**30pt**) (2) Output the mean and standard deviation of these columns
  - (**40pt**) (3) Perform min-max normalization on the columns to generate normalized output

Output Format

- (1) 3 values: min, max, count
- (2) 2 values: mean, standard deviation
- (3) 1 file:
  - Each line: <normalized global active power>, <normalized global reactive power>, <normalized voltage>, and <normalized global intensity>

Implementation Issues

- Missing values
- Conversion of data types

References

- UCI ML repository:
  - Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Questions or Comments?