University of Helsinki
Data Science for the Internet of Things (DS4IoT) - Spring 2024

Exercise 3

Due on 8th April 2024 by 23:59, Helsinki time.

Instruction: All course participants are requested to submit their solutions (in English) through Moodle by the due date. Please take into account the following guidelines about including AI-generated text: https://studies.helsinki.fi/instructions/article/using-ai-support-learning

Submission: You can submit your homework using one of the two following options:

- *Option 1:* submit both the report of non-programming tasks and the solution of the programming tasks in Jupyter Notebook format (*.ipynb*).
- *Option 2:*
  - Submit the report of non-programming tasks in PDF file format.
  - Submit the solutions and the source code of the programming tasks in Python using Jupyter Notebook format (*.ipynb*).

Use the following format for naming the files: *[last name_first name]_[your file name], (i.e., Nguyen_Ngoc_Exercise 1 solution.ipynb).*

Assessment: Participants are encouraged to review course materials to answer the problems and in some cases write computer programs to derive solutions. In all the exercises, do not just give the answers, but also justify, derive and contextualise the answers in Data Science for the Internet of Things contexts. The assessment parameters for each of the tasks are: solution (25%), justification and derivation (50%), contextualisation (25%).

1 Non-programming Tasks

1.1 Learning diary (Compulsory, 4 pts.)

Choose two concepts from this week's lectures (one from Lecture 5 and one from Lecture 6) and describe **what** they mean and **why** they are relevant for the Data Science for the Internet of Things. For full points give an example (or examples) of the concept in a Data Science for the Internet of Things context or otherwise illustrate their importance. Do not copy the definitions from Wikipedia or other sources, but explain the concepts in your own words and relate them with Data Science for the Internet of Things. Use about 100–150 words to describe each of the concepts.

1.2 Similarity (4 pts)

Consider the applications described below. Which type of similarity measure(s) would be best suited for the applications and why?

1. Bicycle route tracking application that attempts to find similar routes for a given GPS route.
2. Fingerprint authentication system on a smartphone.
3. Wireless spectrum analyzer that compares different sources of interference (see Figure 1).
4. Traffic sign recognition in an autonomous vehicle.

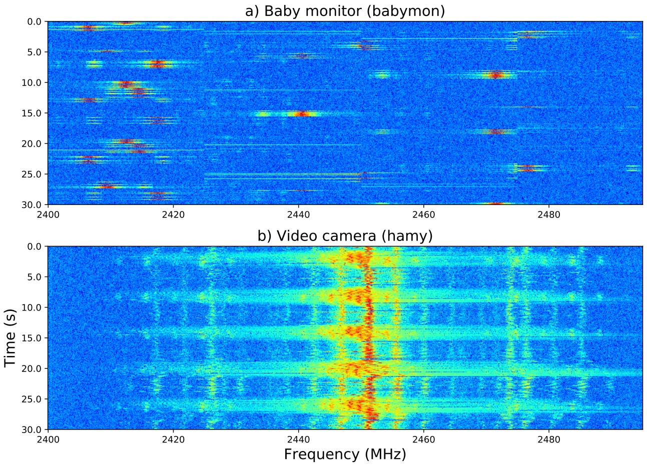


Figure 1: Output of WiFi Spectrum analyzer

2 Programming Tasks

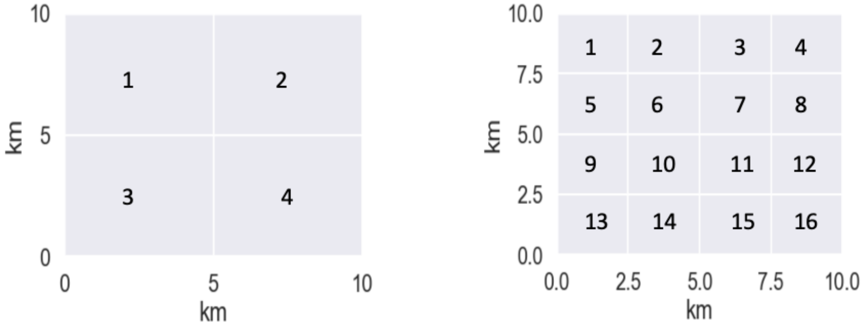
2.1 Measurements and Sampling (4 pts)

Load the dataset *measurements.csv*. The dataset corresponds to simultaneously collected measurements of 76 temperature sensors at certain 10×10 km area, A. The dataset contains the features detailed in Table 1. You can access the dataset in Moodle.

1. For each sensor, replace the missing values using median case imputation. What is the average number of missing values (*nan*) before and after imputation?
2. Use boxplots to highlight the outliers in the data after imputation. How many outliers are there in average?
3. For each sensor, remove the measurements that are not within the interquartile range. Then, remove the sensors having less than 61 samples. How many sensors remain after the preprocessing? Elaborate your answer using the longitude and latitude of the sensors to make a scatterplot that shows sensors' location before and after preprocessing. *Hint: After the preprocessing there should be at least one sensor on each of the* subarea16 *ids.*

Table 1

Feature	Description
timestamp	Time (in minutes) when a temperature measurement was collected by a sensor
id	Sensor identifier
subarea4	Sub-area identifier (from 1 to 4) when deploying a 2x2 grid over A (see Figure 2a).
subarea16	Sub-area identifier (from 1 to 16) when deploying a 4x4 grid over A (see Figure 2b).
temperature	Temperature (in Celsius degrees) measured by the sensor.
longitude	Sensor location on the x-axis.
latitude	Sensor location on the y-axis.



(a) 2x2 grid

(b) 4x4 grid

Figure 2: Grid deployment over A with subareas ids

4. Use the sensors that remain after the preprocessing. Describe the differences on the average temperature and standard deviation at A obtained from applying two sampling strategies (simple random sampling and stratified random sampling) on the two grids configurations (2×2 and 4×4) grids. The number of sensors will depend on the grid size (4 sensors for the 2×2 grid and 16 for the 4×4 grid) and the selection on the subarea identifier (subarea4 for the 2×2 grid and subarea16 for the 4×4 grid). *Hint: Here is an example: simple random sampling using the 2×2 grid should randomly select the 4 sensors from the whole area, A, while stratified random sampling should randomly select one sensor on each subarea (4 sensors in total).* To ensure the reproducibility of the results, set the random_state (seed) of the random sampling function to 1234.

2.2 Preprocessing and Framing (4 pts)

The file *rainforest.wav* contains a 33 second audio recording taken in a tropical forest. Different animal species can be observed from the recording, specially crickets and birds. You are requested to set the specifications of two sensors, each of them focused on the sounds made by specific animals: (i) crickets and (ii) birds.

a) Plot spectrograms of the raw audio signal.

b) Which should be the working frequency range of each sensor?

c) Which should be the duty cycle?

d) Plot spectrograms of the audio signal corresponding to crickets and the audio signal corresponding to birds.

Hints:

- A band-bass Butterworth filter can be used to separate the working frequencies, you just need to provide the lowest frequency value, the highest frequency value and the order of the filter. Python function: *scipy.signal.butter*.

- The duty cycle can be approximated observing the repetitive patterns in the signal. Refer to slides 24-25, Lecture 5 of the course.

- The frequency range corresponding to the audio signal of crickets is expected to be higher than the one corresponding to birds.