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

Exercise 2

Due on 25th March 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 3 and one from Lecture 4) 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 Data Science for the Internet of Things Sensing Pipeline (4 pts.)

Choose ONE of the following Data Science for the Internet of Things applications

- Richardson, Sharon. *"Predicting Presence in Urban Outdoor Spaces."* IEEE Pervasive Computing 18, no. 3 (2019): 21-30.. Link here
- Bhattacharya, Sourav, Petteri Nurmi, Nils Hammerla, and Thomas Plötz. *"Using unlabeled data in a sparse-coding framework for human activity recognition."* Pervasive and Mobile Computing 15 (2014): 242-262. Link here
- Motlagh, N.H., Irjala, M., Zuniga, A., Lagerspetz, E., Rantala, V., Flores, H., Nurmi, P. and Tarkoma, S., 2022. Toward Blue Skies: City-Scale Air Pollution Monitoring using UAVs. IEEE Consumer Electronics Magazine, 12(1), pp.21-31. Link here

1. What is the sensing pipeline used in the paper? Describe which components of the application are involved on each phase.

Note: You can access the papers through OpenVPN of the University of Helsinki. More info here

2 Programming Tasks

2.1 Signal Synchronisation and Interpolation (4 pts.)

Load the dataset of the simultaneously collected sensors' measurements (*sensors_sample.csv*). You can access to the dataset through Moodle.

1. Synchronise the samples of the dataset (consider a sampling frequency of 50Hz).
2. Plot the measurements after and before synchronisation and describe their difference.
3. Interpolate sensor measurements using uniform sampling intervals: 25Hz, 40 Hz, 50Hz, 100Hz. Plot your results and briefly discuss the results: what is good and what is bad in each interval? which sampling interval would you use in practice?

2.2 Ground Truth Evaluation: Annotation Consistency (4 pts.)

Consider the questionnaire answers given in Table 1.

	Q1	**Q2**	**Q3**	**Q4**	**Q5**
Participant 1	5	2	3	5	4
Participant 2	4	3	5	4	3
Participant 3	4	2	4	1	1
Participant 4	4	3	5	4	2
Participant 5	3	1	1	2	3
Participant 6	1	2	3	4	5
Participant 7	2	2	3	2	2

Table 1

1. What is the reliability of the questionnaire?
2. Compare and discuss Cronbach's alpha for different combination of users.
3. Which participant results in the largest decrease in internal consistency? What could be possible reasons for this? What is the internal consistency of the questionnaire when the responses of this participant are removed?
4. Assume the last item (Q5) would be reverse scored instead. What the responses for the final column would be? What is the resulting internal consistency after reverse scoring the item?