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

Exercise 5

Due on 22th 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 9 and one from Lecture 10) 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 Contextual Security (4 pts.)

Read one of the following articles and explain (a) the sensing pipeline and (b) the attack that is described in the paper. (c) Which sensor data is used and how? (d) Do you think the attack is realistic in practice and why / why not?

- Roy, N., Shen, S., Hassanieh, H., & Choudhury, R. R. (2018). Inaudible Voice Commands: The Long-Range Attack and Defense. In 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18) (pp. 547-560). Link here.
- Ning, R., Wang, C., Xin, C., Li, J., & Wu, H. (2018, March). Deepmag: Sniffing mobile apps in magnetic field through deep convolutional neural networks. In 2018 IEEE International Conference on Pervasive Computing and Communications (PerCom) (pp. 1-10). IEEE. Link here.
- Jin, W., Murali, S., Zhu, H., & Li, M. (2021, November). Periscope: A Keystroke Inference Attack Using Human Coupled Electromagnetic Emanations. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (pp. 700-714). Link here.

2 Programming Tasks

2.1 DS4IoT Evaluation (4 pts.)

The file Task5.csv in Moodle contains measurements of temperature (in Celsius degrees) and $PM_{2.5}$ (in $\mu g/m^3$) collected by 40 low-cost weather stations at certain 10×10 km area, A. Consider the following values of temperature= 21.33°C±3.11°C and $PM_{2.5} = 39.45\mu g/m^3$ $\pm 5.22\mu g/m^3$ as the ground truth collected by the local meteorological institute at A.

(a) Aggregate the measurements of temperature and $PM_{2.5}$ by weather station *id* using the median. Perform K-Means and Agglomerative Clustering using *n_clusters*=6 (from 0 to 5). Plot and discuss the results obtained from both algorithms (one plot for K-Means and one plot for Agglomerative Clustering) using different colours for each cluster. Aggregate the measurements of temperature and $PM_{2.5}$ by cluster id (aka *area* in the csv file) using the mean. Compare between the descriptive statistics of the aggregated measurements and the ground truth. Which of the two clustering methods has the lowest error? Why? (**Hint:** Use *KMeans* and *AgglomerativeClustering* functions from *sklearn.cluster* library to implement the classification model. Use a $random\_state = 1234$ for KMeans).

(b) Consider the column *Area* as the actual form that the clusters should be arranged. Compare the classification performance using precision, recall and F1-score, confusion matrix of both algorithms. How significant is the difference between the clustering prediction of the two methods? (**Hint:** Use *precision_recall_fscore_support* and *confusion_matrix* functions from *sklearn.metrics* library to evaluate the performance, and *checkerboard_plot*, *mcnemar_table* and *mcnemar* functions from *mlxtend* library to evaluate the significance.)

2.2 Differential Privacy (4 pts.)

Consider the file Task5.csv. The provider of the weather stations wants to use differential privacy to report in a safe way the average of the median values of $PM_{2.5}$ at A. Aggregate the $PM_{2.5}$ measurements for each weather station using the median.

(a) Aggregate the median values of $PM_{2.5}$ at A using the mean. Report the mean value and standard deviation value.

(b) Implement a function that calculates the sensitivity and report the sensitivity of the aggregated query (a).

(c) Apply differential privacy using the Laplace mechanism on the aggregated query (a) for ϵ values from 0.01 to 1, step size = 0.005. Plot the noisy average value for each value of ϵ. Which ϵ value should be used to ensure the noisy average value kep between $\pm 0.1 \mu g/m^3$ of the actual mean value?

(d) Repeat the previous analysis to safely report the average of the median values of temperature. Determine the appropriate value of ϵ to ensure the noisy average value kept within ± 0.15 Celsius degrees of the actual mean value.