Mission 2.1

Task	Resources	Clarification / hints
Implement a neural network with PyTorch, TensorFlow or Keras to solve the classification problem from Mission 1.2 without manual feature engineering, achieving test accuracy ≥ 0.92.	Lectures 6, 7 Notes pp. 30–38 Géron chs. 10, 11 3B1B playlist	
(manually) Implement learning rate scheduling with warmup and cosine decay.	Géron pp. 388–392 d2l.ai	
Ensure that your results are reproducible.		Set all seeds to fixed values.
Visualize the network architecture and the decision boundary.	Boundary: `matplotlib.pyplot.contourf`	We don't expect anything fancy. Draw a block diagram of the model by hand or check your package's docs.
Explain the difference between SGD and Adam. Justify your choice of optimizer.	Lecture 6 Géron pp. 145–148, 384–385	You don't need to write a novel – a few sentences are enough. This applies to all the discussion questions.
Graph the learning rate, and assess the impact of learning rate scheduling on performance.		
Conduct an ablation study of each architectural and optimization choice.		The intention behind this task was not that you should conduct an ablation study in the traditional sense. Instead, we want you to show the difference (or lack thereof) in performance between SGD vs. Adam, and between constant LR and scheduling. This is largely covered by the previous two tasks. If you tested several options for other aspects of the model, such as activation function, depth and, layer width, include a comparison of those as well.

Mission 2.2

Task	Resources	Clarification / hints
Use standard clustering and dimensionality reduction techniques to assign cluster labels to each data point, achieving an accuracy of 1.0.	Lectures 7, 8 Notes pp. 38–48 Géron pp. 237–246, 260–282 scikit-learn docs (PCA, t-SNE) UMAP docs UMAP paper (for nerds only!)	
Explain the differences between PCA, t-SNE and UMAP. Justify your choice of dimensionality reduction technique(s).		
Explain the differences between k-means and DBSCAN. Justify your choice of clustering technique(s).		
Determine the geographical location of each cluster (representing a server in the storyline).		$*$
In the previous task, you identified what sort of data the file contained. Explain the relative performance of PCA, t-SNE and UMAP on data of this type.		
Optional task		Will not be assessed.

∗: Finding the server locations

If you want a fun CTF-style challenge that involves some leaps of reasoning and outside-the-box thinking, you can solve this task without these hints. However, we want to ensure that you (have the option to) spend as much time as possible learning and applying ML concepts, and as little time as possible trying to guess the thought process behind the tasks.

- Look at the column names in the first line of the CSV file by removing `skiprows=1` (make sure that you do not include the column names in the input to your dimensionality reduction or clustering algorithms, or you will become very confused).
- The task description states that the data is obfuscated. The column names are helpful in unscrambling it.
- The dataset consists of 16384 columns. $16384 = 128^2$. Locations on the Earth's surface (and thus the spatial components of data that can be mapped to geographical regions) are represented by two coordinates.
- The dataset has 4000 rows. From your clustering analysis, you know that there are far fewer than 4000 server locations, so there must be substantial redundancy in the data. This means that you do not need to look at all the rows, and you can probably get away with manually inspecting a few of them.
- This is not a cryptanalysis project.