Module 11 - Bonus Challenges

§M11: Unsupervised Learning, Clustering

- Below are open-ended bonus challenges; solving them is not required but can help you better understand ML/AI in the context of engineering, and how to use them in practical cases.
- Bonus points earned in all homework assignments will be averaged (6 bonus points for each assignment) and then directly added to your final score to calculate your final letter grade.

Challenge 1.1. One important application of clustering is in semi-supervised learning, where clustering can help predict labels for unlabeled data points. In this task, you will explore how clustering can improve predictions when some data labels are missing. It is useful in situations where obtaining labels for all data points is challenging.

Now, you are given a dataset (`m11_bonus.pkl`) for multi-class classification, where 50% of the labels in the training set are masked (unlabeled). The dataset `m11_bonus.pkl` contains:

- `X_train`: the input features for the training set (2D points).
- `y_train_missing`: the class labels for the training set, with some values set to `-1` indicating missing labels.
- `X_test`: the input features for the test set.
- `y_test`: the true class labels for the test set.

Your task is to: (*6pts*)

1. Load the dataset and explore the variables. You can use the `pickle` library to load the dataset. Refer to this tutorial if you are unfamiliar with how to read pickle files. Below is the code for data loading:

```
import pickle
with open(’m11_bonus.pkl’, ’rb’) as f:
    X_train, y_train_missing, X_test, y_test = pickle.load(f)
```

2. Using **K-means** to assign pseudo-labels to the unlabeled data.

   (a) Choose the number of clusters: Set the number of clusters `k` equal to the number of unique classes in the dataset.

   (b) Cluster the data: Apply the K-Means algorithm to the training data, including both the labeled and unlabeled data.

   (c) Assign pseudo-labels: For each unlabeled data point, assign it the class label that is most frequent among the labeled points in the same cluster.

3. Train two logistic regression models:

   (a) One model using only the originally labeled data.

   (b) One model using both the originally labeled data and the pseudo-labeled data generated from clustering.

4. Compare the performance of the two models on the test set by:

   (a) Visualizing and comparing the confusion matrices.

   (b) Computing and comparing the accuracy.

5. Submit your Jupyter Notebook file with appropriate comments.