TDPS22 - Data Science Programming Exam

Johannes Oetsch (johannes.oetsch@ju.se) &
Alexandros Tzanetos (alexandros.tzanetos@ju.se)

May 30, 2025

Please make sure to download the file `exam_2025.csv` that contains data describing if loans have been approved or not. The target variable is *loan_status*. This file is to be used in all questions. Please note that you need to hand in all your solutions similar to the assignments and you can use any programming language you want as long as it can run in a jupyter notebook. Every question will give at most 5 points and the total number of questions is eight. The grades fall within the following point ranges:

- *U*: 0 – 15
- *3*: 16 – 23
- *4*: 24 – 31
- *5*: 32 – 40

Good luck!

1. Read the data and convert categorical features to numerical. Randomly re-order the data based on your date of birth. As an example, if you were born on June 1, 2001, your random seed should be 20010601.
2. Based on the value range of each individual feature in the dataset, visualize the five features with the largest range.
3. Use a clustering algorithm based on the features to separate and plot the examples in a two-dimensional space where the true labels, *loan_status* should be used to color each example.
4. Here, you will create a machine-learning model where the purpose is to predict the target. You can do this with any type of machine-learning algorithm that you prefer and an evaluation metric of your choice. You need to carefully select the hyperparamers, (check this link for more info), for your choice of machine-learning algorithm to get the best possible model. Do a 5-fold cross validation, (check this link for more info), to find the best combination of hyperparameters. Motivate how you select the folds and visualize the distribution of the evaluation metric, that you selected, for all different hyperparameter.
5. Create a way to divide your data in to an initial training dataset of the first 500 examples in your shuffled dataset and subsequent batches of 100 examples. Subsequent training sets should consist of the previous training set and the next batch of examples. You can see a similar division of another dataset in Figures 1-3. In this case the initial training set size is 1000 examples and each subsequent batch 200 examples. We will call this a *Teaching Schedule*.
6. Use the best hyperparamters from *question 4* to create models using the Teaching Schedule from *question 5* to predict all batches of examples, one by one. Visualize your previously selected metric as a function of batch number.
7. Do the same as in *question 6* except that here you will use a machine-learning method that outputs the probability of one class or the other. Exclude all predicted examples that have a probability within the range 0.2 to 0.8 in future training. Provide a number of the total number of excluded examples.

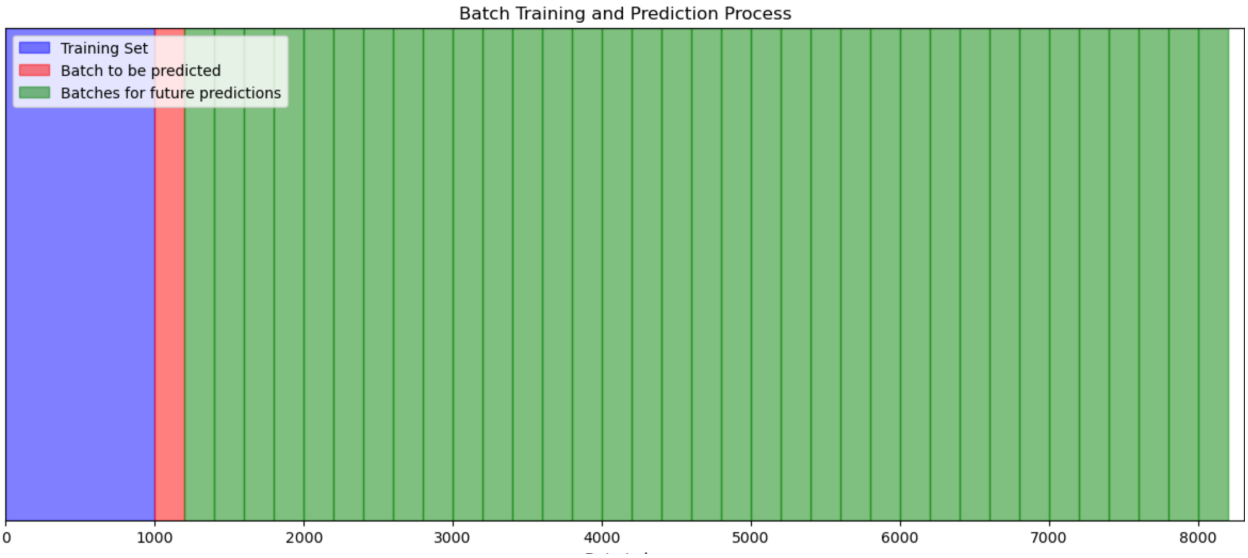

Figure 1: Initial training set.

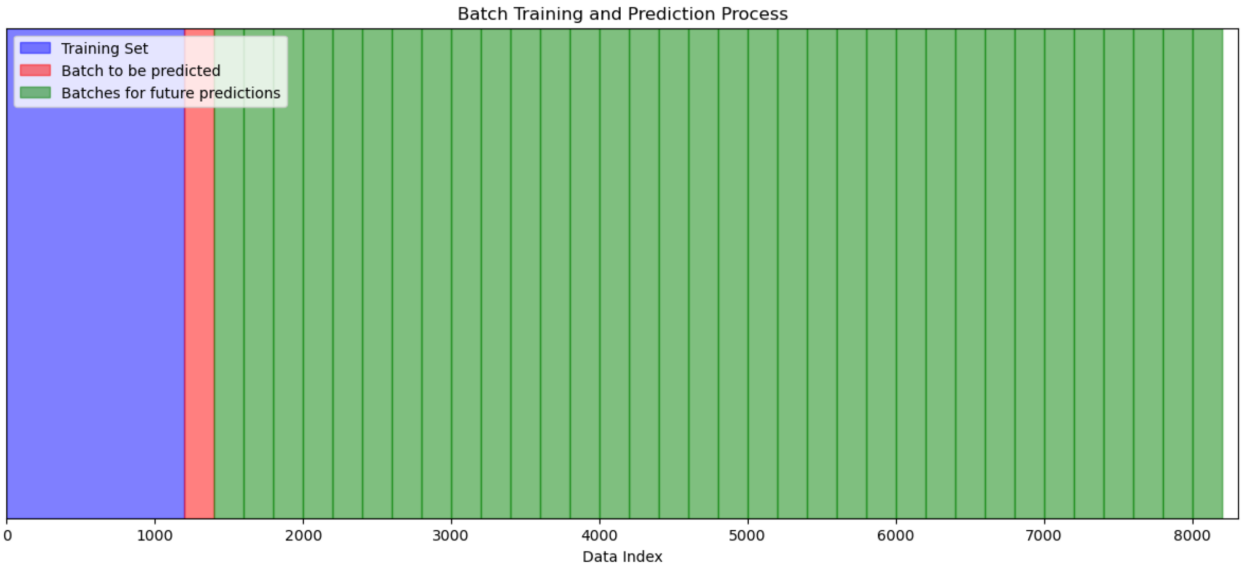

Figure 2: Second training set.

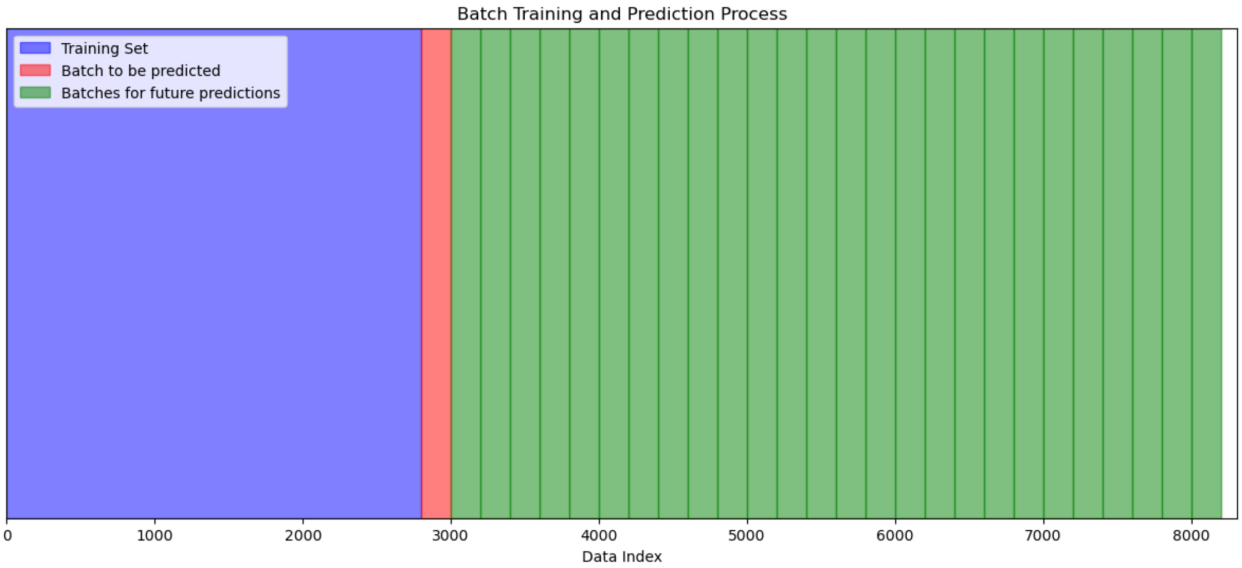

Figure 3: A later training set.

8. Assume that a bank will be using your model to determine whether loans should be approved or not. Use the procedure in *question 7* to determine how much revenue they will gain by assuming that the exclusion in future training is 100 and that a misclasification is -1000.