CMInDS
Indian Institute of Technology Bombay

Programming for Machine Learning and Data Science

==========

Course Project Description Document

General Instructions

1. This project should be preferably done in groups. Groups may comprise between 2 to 4 members each. ***Single member groups are not barred, but discouraged.***
2. Learners are required to form the groups themselves and identify one of the group members as the **Group Coordinator**
3. **The Group Coordinator should register the group details using the following link**: **https://forms.gle/VWmxdgWa5ify9Wkg9**

Submission Guidelines

4. Each group is required to submit the following as part of their final project submission:
5. A **presentation** that documents all steps taken and decisions made.
    a. It should have the names of all the group members mentioned on the first slide.
    b. It should contain maximum 3 slides of 'Executive Summary'
    c. The presentation should be well laid-out, contain adequate plots, tables, **relevant metrics** accompanied with precise explanations, analysis and conclusions. Alternate approaches – if taken – hurdles encountered, and major learnings should also be captured in the presentation.
    d. The presentation should capture all the important steps and conclusions. **Only the presentation will be used to assess the project. All other artifacts like code, data, etc. will be treated as proof of performance – to be consulted only if required**. **Therefore, if something is not mentioned in the presentation, it will be deemed not done.**
6. All Python Notebooks / Python code files, additional data files generated as part of the project should also be uploaded.

Due Date

7. **The due date / time for submitting the project is 23:55 Hrs on May 7, 2025**

Marks and Evaluation Criteria

8. The project will carry **20 marks**
9. Evaluation scheme will be as follows:
    a. **05 marks:** Data science steps followed and completeness of the project
    b. **10 marks**: Innovation and correctness in feature creation, correct selection and application of algorithms, correctness of results (including comparison against the heldout set). Presentation and discussion of the steps, analysis of plots / charts / tables / metrics / results, and conclusions.
    c. **05 marks**: The overall documentation completeness and quality of the presentation will be assessed in this part.

==========

The following data sets have been referred to in this problem statement

Text dataset:
https://drive.google.com/file/d/1IjYJuH5bjq4ZStxOMEbYV5nxrxS8M5QS/view?usp=sharing

Image dataset:
https://drive.google.com/file/d/1eL6EwTjTlDYskKK3OlqwqQCl3JhgxPwg/view?usp=sharing

---

Course Project: Learning to Label – From Clustering to Classification

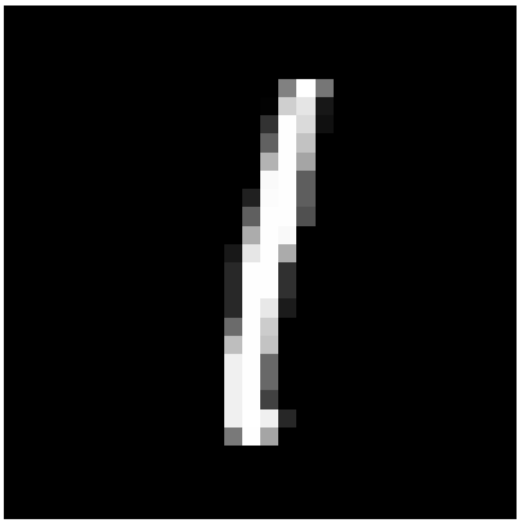
In the real world, data rarely comes with clean labels. As a machine learning engineer at a data-centric company, you're handed fairly large, unlabelled datasets across two domains—images and text. Your team is tasked with building a functional classification system, but without labeled training data, the first challenge is to *create your own labels*.

Welcome to your mission: learning to label intelligently and building robust classifiers from scratch.

Datasets

Image dataset

This dataset contains select examples of handwritten digits from MNIST but are unlabelled. Here is an example image:



Text dataset

The text dataset is a corpus of documents. Each document belongs to one out of the five topics (n_clusters = 5). The documents are unlabelled raw text.

We have given boilerplate code for looking at the datasets in this colab. Make a copy of this colab and start preparing your solution based on the following problem statement.

Project Overview

This project simulates the workflow of a modern unsupervised-to-supervised machine learning pipeline. It is structured in **two stages**, each of which builds upon the other:

Stage 1: Clustering and Self-Supervised Labeling

As mentioned above, you are given an image and a text dataset. The image dataset has **10 true classes**. The text dataset has **5 true classes**. Your job is twofold:

- **First,** create labels for the unlabelled datasets by creating appropriate embedding / encoding for the entries in the dataset. Then, use a clustering algorithm of your choice to create clusters. Once the clusters are created, you can use the clusters to assign labels to each example in the dataset. We will call the dataset given to you as X_train and the labels that you have generated as Y_train.

- **Second,** use {X_train, Y_train} to train a classifier of your choice.

It is advisable to look at different featurization/clustering methods and train multiple classifiers and check performance of each. Choose the best combination. You can also explore model validation and early stopping approaches.

Stage 2: Evaluation Against a Hidden Test Set

We have a held out test set. Your approach will be evaluated against this set by us and a portion of your project grade depends on your performance on this set.
