2-Day Project

You are given 3 questions in this area. Work on 1 question of your choice.

1. Anomaly Detection System

Problem Description. Anomaly detection is typically characterized as the process of identifying events or observations that significantly deviate from the majority of standard data and do not conform to a well-established notion of normal behavior.

In this project, our objective is to construct a monitoring system based on anomaly detection capable of issuing alerts for the maintenance of factory components. We assume that the system gathers observations, each containing up to 51 feature values, from individual components and must determine whether they require maintenance. To facilitate this, we have a dataset collected from various components within the factory, which we will use to train a model. This dataset not only includes the observations but also labels indicating the maintenance requirements for each component. A label of $y = 1$ indicates that the component requires maintenance, while a label of $y = 0$ indicates normal conditions. This information can be used to train a model in a supervised manner.

By leveraging this dataset, we aim to build a robust monitoring system that can accurately identify anomalous behavior in factory components and alert maintenance accordingly.

- Please refer to the **dataset** in the link below:

  `https://tinyurl.com/2a55tr96`

  The dataset includes feature values and labels.

(a) **Data Preprocess.** Describe how you preprocess the data into a format to build your model and how you split the dataset to properly build your model.

(b) **Exploratory Data Analysis.** Explore the dataset with descriptive statistics and visualization to understand the problem.

(c) **Learning methods.** Build a model to achieve the goal explained above using available data. Describe your learning algorithm (or statistical method) to build your model, and explain how it is implemented.

(d) **Evaluation Metric.** Define your evaluation metric, and explain why it is appropriate for the problem.

(e) **Result and Discussion.** Present your work at the oral exam, including (but not limited to) problem setting, hypotheses, how you preprocess data, description and justification of your learning algorithm, prediction results, interesting observations or insights learned, and potential future work.

2. Kaggle Dataset Classification and Prediction

Problem Description. Kaggle is an online platform that provides a community and infrastructure for data scientists to work on various data-related projects, such as data analysis, predictive modeling, and deep learning. It offers a diverse collection of datasets. Kaggle has become a popular platform for data scientists due to a wide variety of datasets and its easy-to-use interface.

You have been given access to two files—an Excel spreadsheet containing the 10,000 datasets with the highest number of votes on Kaggle, and a compressed zip archive of the original JSON files. The Excel file was generated by extracting data from the JSON files, and provides a more user-friendly view of the dataset information. However, if you require more detailed or raw information about any of the datasets, the compressed JSON archive can be used to access this data.

- `kaggle_datasets.xlsx`: `https://bit.ly/hp-kaggle-datasets-xlsx`
- `kaggle_datasets_json.zip`: `https://bit.ly/hp-kaggle-datasets-json`

The Excel file contains the following fields.

- `rank`: Rank
- `datasetUrl`: URL for the dataset
- `ownerUrl`: The dataset owner's URL
- `creatorUrl`: The dataset creator's URL
- `scriptCount`: Number of notebooks (scripts) created for the dataset
- `viewCount`: Number of views
- `downloadCount`: Number of downloads
- `dateCreated`: Date created
- `dateUpdated`: Date updated
- `datasetSize`: Size of the dataset in bytes
- `totalVotes`: Number of votes
- `datasetId`: Dataset ID
- `categories`: Category IDs delimited by — (pipe character)
- `tags`: Tags delimited by — (pipe character)
- `usabilityRating`: Usability Rating
- `medal`: Gold, Silver, Bronze, or empty

You have three tasks. (1) You are tasked to find important factors for getting votes and/or attaining a Gold, Silver, or Bronze medal. **(2)** You then need to build a prediction model for the number of votes and a classification model for which medal a dataset would attain. **(3)** Make predictions for the datasets created in 2023 how much votes they would eventually receive and which medal they would attain in a year.

To complete the tasks above in the following steps.

(a) **Exploratory Data Analysis.** Explore the data with descriptive statistics and visualization. (ex. finding important features or factors having a high positive or negative impact on the target variable)

(b) **Model Development.** Build models to achieve the goal explained above using the dataset. Describe your learning algorithm (including statistical methods if used) to build your model, and explain how it is implemented.

(c) **Results and Discussion.** Present your work at the oral exam, including (but not limited to) problem setting, hypotheses, how you collect and preprocess data, description and justification of your learning algorithm, prediction/classification results, interesting observations or insights learned, and potential future work.

3. Sum of two digits

Problem Description. You will work with the modified MNIST image datasets in this project. MNIST [1] is a commonly used dataset for training image recognition systems. Standard MNIST consists of handwritten digits expressed as 28 by 28 pixels with binary (0/1) pixels; and it contains 60k training, 10k testing instances where each image is accompanied by the digit labels corresponding to the image.

In this project, we will provide a `Paired MNIST` dataset where each training instance consists of two handwritten-digit images and a sum of those two digits as a label instead of the traditional setup of one label per image. Example Figure and the link is provided below. The first goal of this project is to examine whether machine learning models can learn to classify single digits given a 28 by 28 pixels (the standard MNIST task) even when we provide aggregate information (sum) for a pair of digits. For that purpose, the test set will remain the same as the standard MNIST dataset, i.e., one label per image.

Please refer to the **datasets** in the link below; the dataset description is included in the zip file.

- `Pair MNIST dataset link`: `https://tinyurl.com/4fzefxwj` is the link to the dataset zipfile.
- The zip file has both training, test set as well as `instructions.pdf`. The `instructions.pdf` describes data formats, file types and how to load data.
- Inside the training set, there are 60k pair of images in one binary tensor and `summation` 60k-dimension label (sum of two digits) vector.

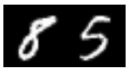
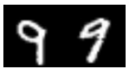
Image	Interpretation	Label
	$5 + 0$	5
	$8 + 5$	13
	$9 + 9$	18

Figure 4: The first column shows two digits side by side, and the last column shows the pair's label. Each pair's label is the summation of the two digits as described in the interpretation; however, labels for individual digits are not given to you.

[Warning] Strict rules for this project:

- **Do not use pre-trained model.**
- **Do not label any data yourself**
- **No external data downloading**

(a) **Exploratory Data Analysis.** Explore the data with descriptive statistics and visualization. Which labels occur most frequently? Is this data biased? If so, why do you think it became biased?

(b) **Baseline model:** Describe your model architecture and how you would train a single-digit classifier given the sum-of-pair information.

(c) **Learning methods:** Describe your learning approach and explain your logical reasoning behind it.

(d) **Evaluation Metric:** Define your evaluation metric, and explain why it is appropriate for the problem.

(e) **Low-resource model:** If you were given a standard MNIST dataset (one label per single-digit image), but with a very small amount of training data (e.g. 1% dataset). Do you think you can utilize similar techniques you used on the pair dataset to improve the low-resource model? Why do you think it will be helpful? [**Note: Actual experiment is not required for this question. But if you wish to experiment, follow the below instructions.**]

In case one wants to carry out the (optional) experiments on a low-resource model, feel free to download official MNIST data (from anywhere) for solving this question only. Note that the official data should only be utilized for question (e).

(f) **Result and Discussion:** Present your work at the oral exam. Your presentation should include a description of the problem setting, hypotheses, how you analyzed your data, description, and justification of your learning approaches, prediction results, interesting observations or insights learned, and potential future work.

References

[1] The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.