Data Analytics 1

Assignment 4

Release: 31 October 2025
Deadline: 14 November 2025 (11:55 pm)

Section 1: Classification

In this section, you are required to create a multiclass classifier to classify Customers, based on some given attributes, a mix of numerical and categorical attributes.

Use the data provided in `train.csv` to train your classifier for the column 'Segmentation' (target variable).

Note: You are allowed to use sklearn or other libraries for implementing the metrics and individual classifiers like SVM (task 1) and decision trees (task 2) but you are not allowed to use direct functions like `sklearn.multiclass.OneVsOneClassifier` or `sklearn.multiclass.OneVsRestClassifier` (for task 1) and `sklearn.ensemble.RandomForestClassifier` (for task 2).

Task 1 [50 marks]

Build two multi-class classifiers one-vs-one, one-vs-all using SVM classifier and train it on the given data (`Customer_train.csv`). [correctness: 30 marks, accuracy score on `test.csv`: 20 marks] - Consider factors such as data cleaning, data skew, handling numerical vs categorical variables.

a. Running `teamId_classifier_ovo.py <path to test file>` (will be in same format) should output a `ovo.csv` file with the predicted labels by one-vs-one classifier with column names as "predicted" (all in lower case), which will then be checked with the actual labels to determine your model's accuracy score.

b. Running `teamId_classifier_ova.py <path to test file>` (will be in same format) should output a `ova.csv` file with the predicted labels by one-vs-all classifier with column names as "predicted" (all in lower case), which will then be checked with the actual labels to determine your model's accuracy score.

Task 2 [30 marks]

Build a random forest classifier using n Decision trees for the above given dataset (try different values of n).

a. For this task just print the output for the predictions of test dataset in the notebook itself and print the accuracy and other metrics also in the notebook file.

Task 3: Report and Analysis [20 marks]

a. How does class imbalance affect multiclass classification, and what strategies can be employed to mitigate its impact, especially with small datasets? (5 marks)

b. How can the choice of hyperparameters make the random forest classifier and SVM classifiers more prone to over or under fitting? (5 marks)

c. Plot the confusion matrix and include the precision, recall, f1-score metrics in the report. (5 marks)

d. Compare the results obtained for one-vs-one and one-vs-all (which according to you performs better for the above dataset). (5 marks)

Include all the above plots, analysis and answer for theoretical question in `team_id_report_1.pdf`.

Section 2: Clustering

Introduction

This section is on clustering and data visualization. The goal is to gain experience with common clustering techniques and data preprocessing/visualization tools. There are three parts in this section; the first part involves data visualization and preprocessing using PowerBI, the second part requires the implementation of K-Means from scratch, and the third part requires different clustering algorithms. In the third part, sklearn can be used for implementations, including for K-Means.

Task 1: Data Visualization and Preprocessing Using PowerBI [20 marks]

Use `Dataset1/` for this part. In this part, you are required to use PowerBI to perform data visualization and preprocessing on the provided dataset. This dataset encompasses data for approximately 2,240 customers. Each row within the dataset includes details about an individual customer, including their personal information, purchasing behavior, and interactions with marketing campaigns.

- **Data Exploration and Visualization:**
- Import the dataset into PowerBI.
- Create various visualizations to explore the dataset, such as:
  - Bar charts to show the distribution of categorical variables (e.g., Education levels, Marital Status).
  - Histograms to show the distribution of numerical variables (e.g., Age, Income).
  - Scatter plots to identify relationships between variables (e.g., Income vs. Total Spending).
  - Heatmaps or correlation matrices to identify correlations between variables.
- Provide insights based on the visualizations. For example, identify any trends, patterns, or anomalies in the data.
- Handle the basics of data preprocessing using the tool.

Task 2: K-Means Clustering Implementation [35 marks]

Use the preprocessed dataset from Part 1 for this part. This part requires you to implement the K-Means clustering algorithm from scratch.

- Implement the K-Means clustering algorithm without using any built-in clustering functions.
- Choose k = 2,5,7,9.
- Use the elbow method and silhouette score to find the optimal number of clusters (from 2, 5, 7, and 9).
- Analyze the clusters formed and interpret the results in the context of customer segmentation.

Task 3: Clustering with Different Algorithms [35 marks]

Use datasets in `Dataset2/` folder for this part.

- For the given 4 datasets in Part 3, use the following clustering algorithms:
  - K-Means
  - Agglomerative Clustering (you can try any linkage strategies: single, complete, etc.)
  - DBSCAN
- Perform the following tasks:
  1. Apply each clustering algorithm to each dataset and visualize the clusters using a scatter plot (different color for each cluster). Since there are only 2 features, the data points can be plotted directly.

2. Analyze which algorithm performs the best for each dataset and mention the reason if an algorithm fails for a particular dataset.
3. Tabulate silhouette scores for each dataset and each algorithm.

Feel free to visualize the data before applying the algorithms and set the parameters accordingly (or you can use the elbow method). For DBSCAN, try eps values in the range of 1 to 5 and adjust other parameters (min_samples from 1 to 50) based on data visualization.

4. Try using another metric apart from silhouette score (e.g., sum of all intra-cluster distances) and analyze if there is any difference in the results.
5. Ignore the noise points predicted by DBSCAN while computing the silhouette score (mention the number of noise points in your table).

- Plot the visualizations in your notebook and write the reasoning/inference. Tabulate the scores in the notebook markdown.

Task 4: Report and Analysis (10 marks)

Write a comprehensive report detailing the results and conclusions derived from all parts of this section. Your report should include the various plots and insights as well as reflection on the use of PowerBI for data visualization and preprocessing. You should take into account both insights from visualization and from clustering to reach conclusion.

Submission Instructions

Submit a `teamId.assignment4.zip` file containing the following:

- `teamId_classifier_ovo.py`
- `teamId_classifier_ova.py`
- `teamId_random_forest.ipynb`
- `teamId_report_1.pdf`
- `teamId_clutering_Part1.pbix` (PowerBI file for Section-2 Part 1)
- `teamId_clutering_Part2.ipynb` (Jupyter Notebook for Section-2 Part 2)
- `teamId_clutering_Part3.ipynb` (Jupyter Notebook for Section-2 Part 3)
- `teamId_report_2.pdf`
- Any intermediate files or datasets used