# Data mining HW3: Report

Malik DAHMANI

June 21, 2024

# Contents

# 1 Question 1: Explain your implementation which get the best performance in detail.

First, I use the function normalize using MinMaxScaler from sklearn to scale the features of the dataset. This ensures that all features have values in the range [0, 1], which is important for distance-based algorithms like k-nearest neighbors to function correctly.

Then, the function calculate_distances computes the k-nearest neighbors for the test data using the specified metric. The function NearestNeighbors from sklearn.neighbors is used with metrics such as 'manhattan', 'euclidean', and 'minkowski'. For each test point, it finds the k-nearest points in the training data and returns the distances to these neighbors.

To continue, the evaluate_metrics function evaluates different distance metrics by calculating the average distance to the k-nearest neighbors for each test point. The silhouette score is used to measure the quality of clustering (how similar a point is to its own cluster compared to other clusters). The metric with the highest silhouette score is selected as the best metric.

To conclude, I use a function to export my result to a CSV file.

# 2 Question 2: Explain the rationale for using auc score instead of F1 score for binary classification in this homework.

First, the Area Under the ROC Curve (AUC-ROC) score measures the ability of a classifier to distinguish between classes. It plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The F1 score is the harmonic mean of precision and recall. It balances the trade-off between the precision (positive predictive value) and recall (sensitivity).

To continue, in the context of this homework assignment, where anomaly detection is performed on the Letter Image Data features, and the goal is to detect outliers (anomalous letters) in the testing set, the AUC (Area Under the ROC Curve) score is more appropriate than the F1 score for several reasons:

- Class Imbalance: Anomaly detection problems typically involve a significant imbalance between the normal and anomalous classes. The

number of anomalies (outliers) is usually much smaller compared to the number of normal instances. AUC is more robust to class imbalance than the F1 score because it evaluates the performance across all classification thresholds, rather than focusing on a single threshold

- Threshold Independence: The AUC score measures the ability of the model to discriminate between the two classes across all possible classification thresholds. This is particularly useful in anomaly detection, where the optimal threshold for distinguishing normal from anomalous instances may not be known in advance. The F1 score, on the other hand, depends on a specific threshold, which might not be optimal for the given problem.

- Receiver Operating Characteristic (ROC) Curve: The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The AUC score, derived from this curve, provides a single scalar value that summarizes the model's performance. It gives insight into the trade-off between the TPR and FPR, which is crucial for anomaly detection where the cost of false positives and false negatives can vary significantly.

- Overall Performance Measurement: The AUC score considers the entire range of decision thresholds, providing a comprehensive measure of the model's performance. This is important in anomaly detection as it helps to ensure that the model is effective in distinguishing between normal and anomalous instances across different scenarios.

## 3 Question 3: Discuss the difference between semi-supervised learning and unsupervised learning

First, Supervised learning is a machine learning technique where an algorithm is trained on a labeled dataset. Each training example consists of an input and a corresponding output, and the goal of the algorithm is to learn a function that maps inputs to their respective outputs. Unsupervised learning is a machine learning technique where the algorithm is used to draw inferences from datasets without labeled responses. The goal is to discover hidden structures or patterns in the

data. The difference between semi-supervidsed learning and unsupervised learning are:

- Labeling: Semi-supervised learning uses both labeled and unlabeled data, while unsupervised learning uses only unlabeled data.
- Goal: Semi-supervised learning aims to improve model performance using the available labeled data, while unsupervised learning aims to uncover hidden patterns and structures within the data.
- Application: Semi-supervised learning is useful when labeling data is costly, and there's a need to enhance performance with limited labeled data. Unsupervised learning is useful for exploratory data analysis and discovering hidden patterns.