Matwej Novgorodtsev,
ID: 244580IV,
Home assignment 3

1 Exercise 1

The aim of this task was to analyze a chosen text dataset. The classification process utilized a dataset containing emails labeled as spam or regular. Preprocessing involved removing stop words, converting text to lowercase, and removing punctuation and numbers. Subsequently, an SVM model was trained on the prepared data. The results demonstrated an accuracy of 84%. Next, centroids and the decision boundary were visualized. For this purpose, PCA dimensionality reduction was performed, and the model was retrained on the reduced dimensions. The trained model was then used to predict the decision boundary, as illustrated in the figure below.

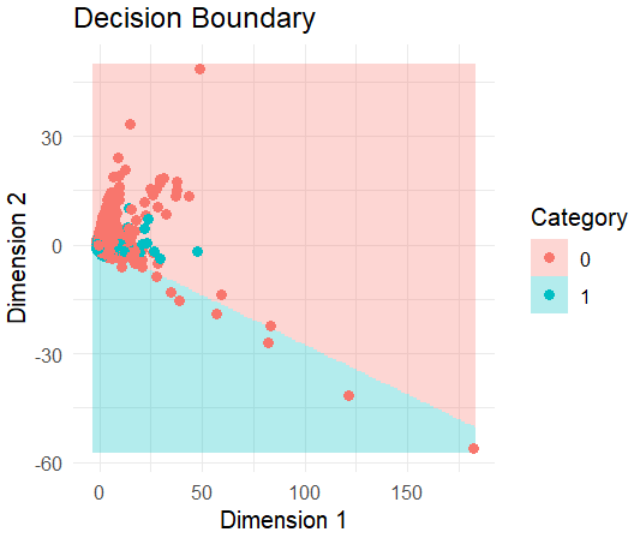

Figure 1: Boundaries; 0 - spam, 1 - regular

Next, the centroids were calculated and displayed.

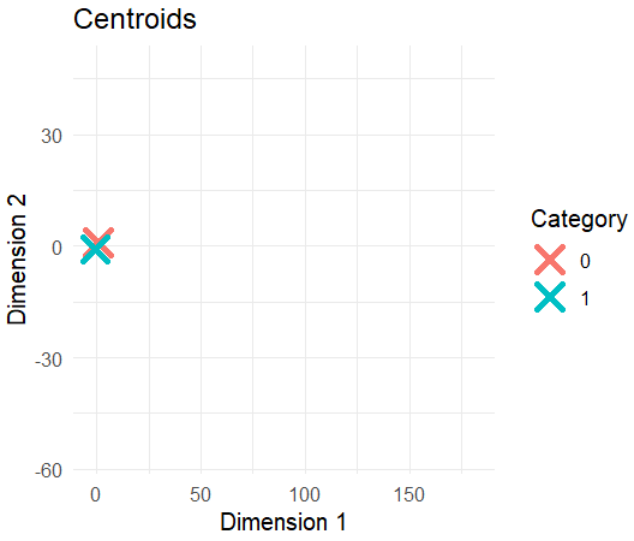

Figure 2: Centroids; 0 - spam, 1 - regular

We can observe that the boundary between spam and regular emails is very narrow, as indicated by the distance between the two centroids—they are very close to each other.

2 Exercise 2

This task involves implementing foundational network measures to provide insights into the roles and importance of individual nodes and their connections within the network. Where possible, the custom implementation was compared to built-in functions from iGraph. The plot below visualizes the local clustering coefficient. The red nodes are considered to have the highest values, while the white nodes have the lowest.

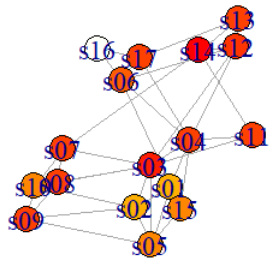

Figure 3: Local clustering coefficient visualization

The gregariousness measure was successfully implemented, and the results indicate that the highest value of 1.0 was achieved for node s16.

The prestige measure, representing a node's influence based on its connections, was successfully implemented and validated against iGraph's built-in function, yielding identical results. The highest prestige value was observed for node s03 with a value of 0.5625.

The common neighbor measure function calculates the number of common neighbors between two nodes i and j in a network defined by the edges dataset. For example, when applied to nodes 's05' and 's06', the function returns 1, as they have one common neighbor.

Jaccard distance was implemented, for nodes s01 and s02, a value of 0.25 was achieved, which, after brief calculations, was confirmed to be correct.

Degree Centrality algorithm measures the number of direct connections a node has, indicating its immediate influence within the network, Closeness Centrality and Proximity Prestige algorithms assess how quickly a node can reach all other nodes in the network, with proximity prestige reflecting its closeness to others. The Betweenness Centrality algorithm identifies nodes that act as bridges, measuring how often a node appears on the shortest paths between other nodes. These algorithms have been also implemented in the code to compute

these centrality and clustering measures for the nodes in the network.

3 Exercise 3

Reservoir sampling is an algorithm used to randomly select a representative sample from a data stream of unknown or infinite size. This exercise involves implementing the algorithm, simulating a streaming dataset with concept drift, and analyzing how the reservoir adapts under different scenarios. The aim is to demonstrate streaming behavior, illustrate concept drift, and compare reservoirs.

First of all we had to generate the data. It was done, splitting the data to 10 different "periods". The plot below illustrates the data distribution in the first scenario, where k=3000. As observed, this represents a complete depiction of the data stream.

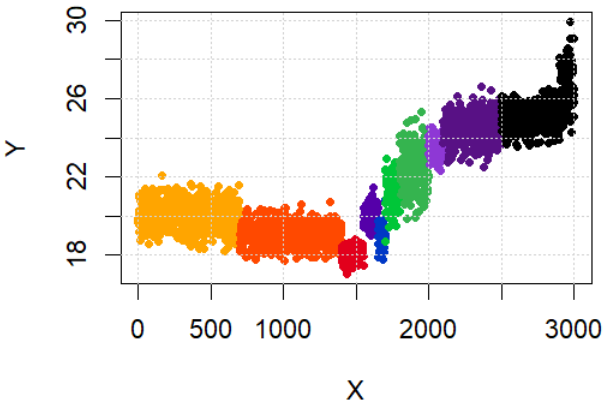

Figure 4: Presentation of data

Then, the reservoir sampling function was implemented to test it without the drift concept. The plot illustrates the reseivour sampling in the first scenario.

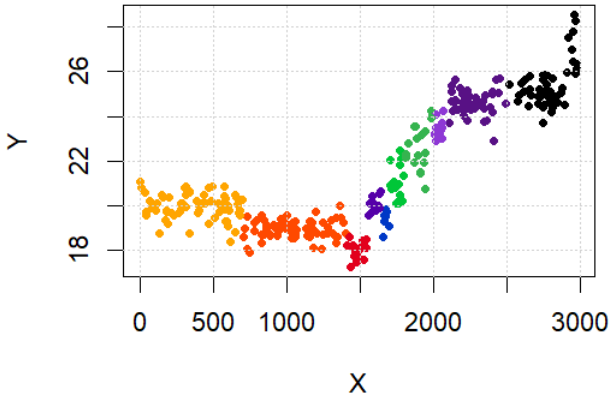

Figure 5: Reseivour sampling without drift adjustments

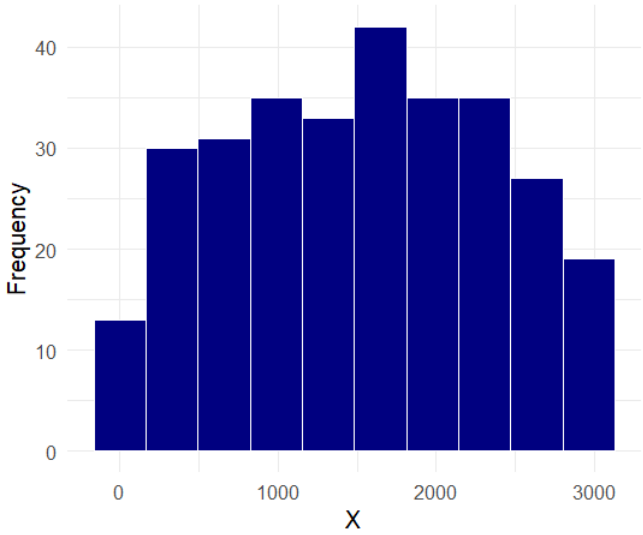

Figure 6: Distribution of reseivour sampling without drift adjustments

Concept drift refers to the situation where the data changes over time. As older data may no longer accurately represent the current trends, it is essential to update the model to reflect these changes. Reservoir sampling can be used to handle concept drift by continually selecting a representative sample that prioritizes the most recent data, ensuring that the model adapts to the evolving data stream.

The plot below demonstrates the use of reservoir sampling with adjustments to concept drift, as applied to 300 samples from the original dataset.

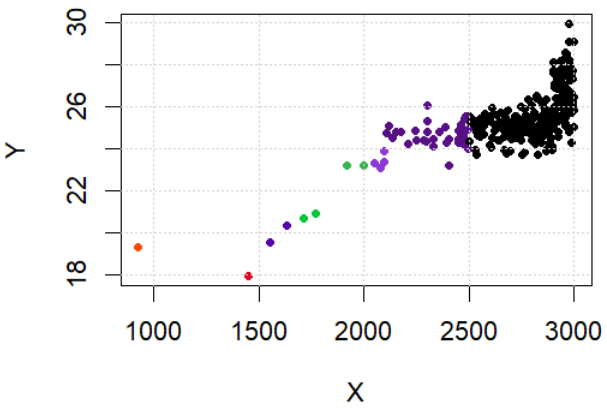

Figure 7: Reseivour sampling with adjusting to drift

In our task, reservoir sampling with adjustments for drift was implemented. The plots represent the distribution of points and the points that were ultimately selected.

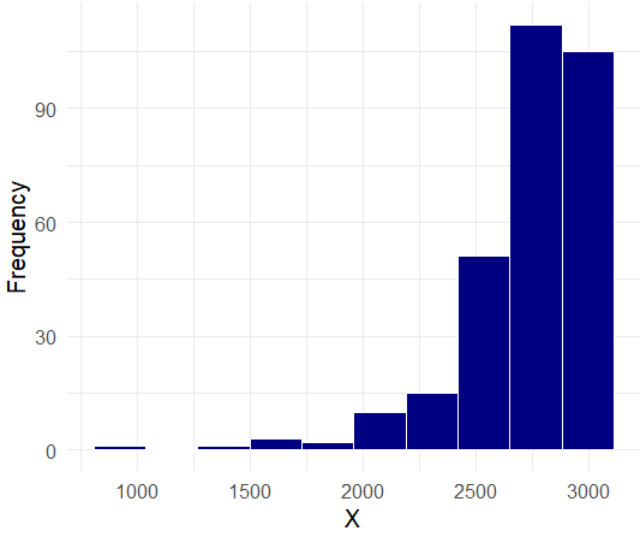

Figure 8: Distribution of reseivour sampling with adjusting to drift