Applied Machine Learning Assignment

Task 2 – Using a Neural Network on a fuel consumption dataset

Lovato Alessio - allo2203@student.miun.se
Squarcialupi Riccardo - risq2300@student.miun.se

Mittuniversitetet
MID SWEDEN UNIVERSITY

May 30, 2023

Introduction

The assignment required training a neural network to predict fuel consumption for bus services. This model was chosen because it was the only one that could withstand the high non-correlation between IDs assigned to drivers and buses and all other features, thus maintaining a decent degradation in accuracy.

Dataset Analysis

The proposed dataset contains a total of 9885 data points with 19 features:

- Number of bus line
- Day of the year
- Weekday
- Hour of the day
- Travel time in minutes
- Vehicle ID
- Vehicle class ID
- Driver ID
- Distance
- Average outside temperature
- Tyre diameter
- Tyre width
- Vehicle model year
- Odometer
- Fuel used (ground truth)
- Type of day
- Swedish holiday (yes/no)
- Engine power
- Days since last service

Data Processing

Following an initial analysis, it was found that the values in the 19th column of the dataset coincided row by row with the values in the 18th column (as visible in Figure 1).

18	19
2130	2130
2310	2310
1320	1320
2130	2130
2860	2860
2310	2310
2860	2860
2310	2310
2310	2310
2860	2860
2310	2310
2310	2310

Figure 1: Example of data comparison between column 18th and 19th

This was not supposed to happen since the values have different meanings, so, based on the meaning of the data, the ”Engine power” feature was retained, while the ”Days since last service” feature was removed.
Furthermore, many points in the data set (especially in the Tyre Diameter feature) were missing or inconsistent, so the MATLAB function *'fillmissing'* with the option *'knn'* was used to fix the data.
Normalization is also applied to the dataset excluding:

- Number of Bus Line, Driver ID, Vehicle ID and Vehicle class ID, as they serve as unique identifiers for their respective entities.
- Swedish Holiday since is a boolean value.
- Type of day: in this case encoding is applied, the attribute is now represented as three separate columns: ”sunny day”, ”snowy day”, and ”rainy day”, where each column will have a binary value indicating the presence or absence of that specific type of day.

Additionally, Principal Component Analysis was also applied trying to maintain the 99% variance of the dataset with fewer features. Still, the dimensional reduction was not satisfied at all removing only two attribute to the list, so all the data were kept. After the data processing the dataset was augmented to 20 features.

Ethics concerns

Since there was no correlation between drivers' personal data and their ids in the dataset, nor was there any information about the routes or companies that provided the data, anonymity is guaranteed. A main concern could be that companies, commissioning the dataset, have free access to the ids of their employees, and so this could result in discrimination or pressure on the employees to save more money on fuel.

Clustering

A way to maintain reliability when adding new identifiers in the model is to group the driver IDs into clusters. This allows the company to insert a new driver into the system and the model will place the new driver in a particular cluster, so that no training is needed again.
This approach can be expanded to the bus ID and bus lines so that the model is resilient against the insertion of new data.
With this method, we also ensure more privacy for the drivers since their real ID is replaced with a ”group ID”.

K-Means method

The method used to replace all IDs was the "K-Means" method. Quoting Wikipedia [1] "k-means clustering is a method of vector quantization, [...], that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster". Elbow analysis and later Silhoutte analysis were used to select the best k for each cluster. Figure 2 shows a visualization of the K-means process.

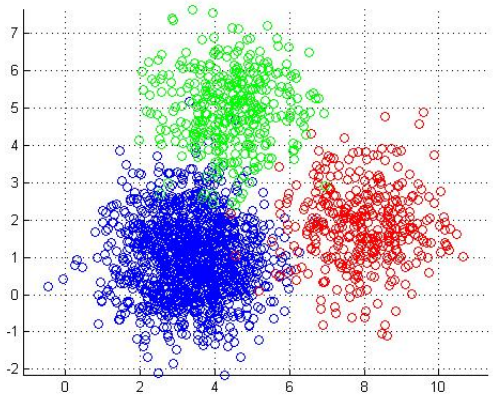


Figure 2: 2D representation of k-means clustering

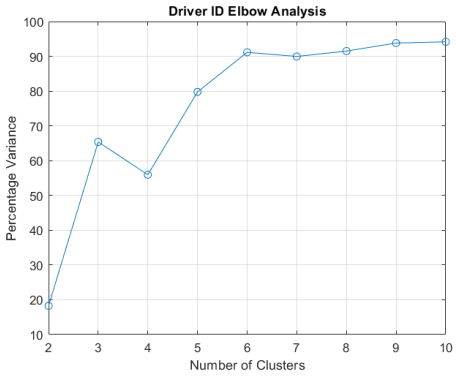
Elbow Analysis

Initially, a heuristic method named "Elbow Method" [2] was selected to obtain the most suitable number of clusters. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.
Due to the fact that the dataset had a large number of data points, the range of the cluster in which we searched for the ideal number of clusters was from two to twelve clusters.
Looking at these curves (Figures 3a, 4a and 5a), it was difficult to pick the correct number of clusters, so a further analysis has been carried out.

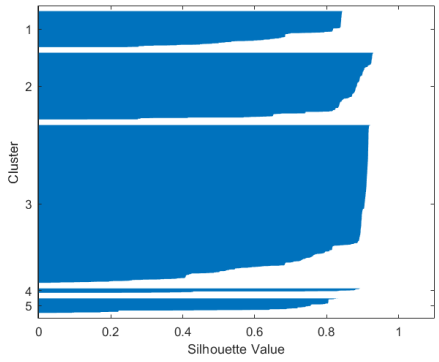
Silhouette Analysis

Silhouette analysis [3] refers to a method of interpretation and validation of consistency that provides a brief graphical representation of how well each object has been classified. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).
Based on a comparison of all the silhouettes produced by the range of cluster numbers described previously, the best results were: five clusters for driver IDs, six clusters for

vehicle IDs and five clusters for bus lines (whose graphs are shown in Figures 3b,4b and 5b).

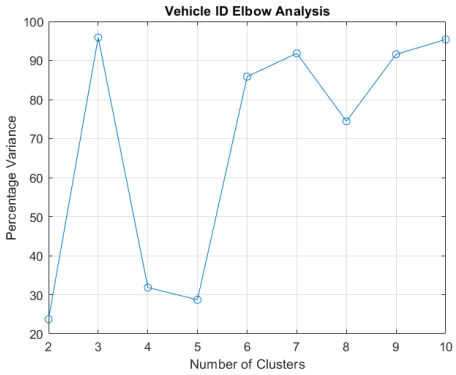


(a) Elbow analysis

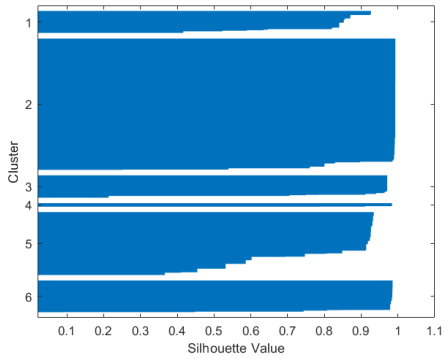


(b) Silhouette analysis

Figure 3: Driver IDs

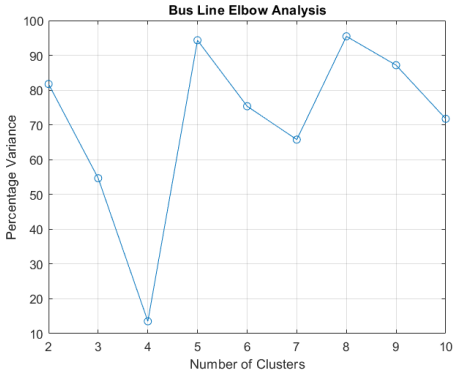


(a) Elbow analysis

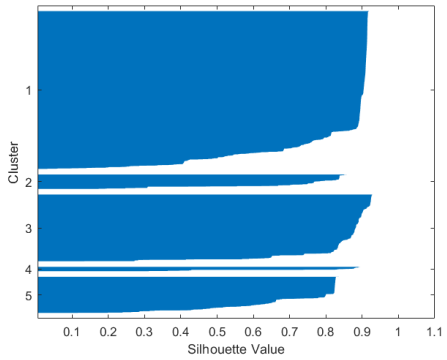


(b) Silhouette analysis

Figure 4: Vehicle IDs



(a) Elbow analysis



(b) Silhouette analysis

Figure 5: Bus Lines

Comparison with other model

To understand whether the new cluster ids assignment is indeed effective and also to check possible performance boost, it is useful to perform a comparison between a basic NN trained with the original dataset, a second NN trained on the pre-processed dataset and a NN trained on the clustered pre-processed data. Figure 6 shows the Neural Network structure shared among the solutions (in the basic one the input layer was made of 19 neurons not 20, since the original features were 19).

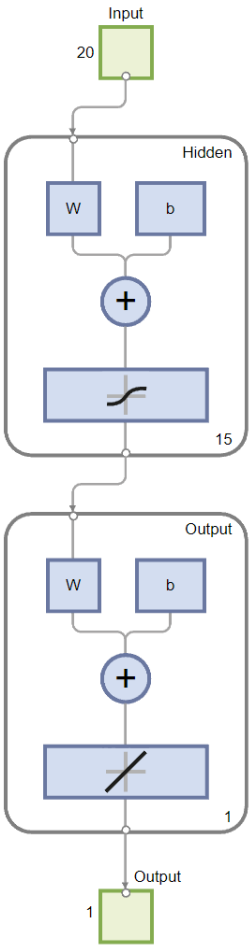


Figure 6: NN Model

Due to the high performance of the networks the data were splitted in: 25% train, 25% validation and 50% of test with a 'limit' during the train of 10 epochs.
For each Neural Network, the average values of MSE, Accuracy (% of predictions with less than 0.02 margin from the test "ground truth"), and Co-variance error were calculated based on 50 different trainings. The average performance of the NNs are reported in Table 1.

MODEL	MSE	COVARIANCE ERROR	ACCURACY
Basic	0.0537	3.7291e+04	87.47%
Pre-processed data	0.0114	3.6117e+04	91.54%
Clustered pre-processed data	1.1673e-04	3.6122e+04	93.45%

Table 1: NNs average performance

An additional effort was made on top of the clustering solution to find the optimal activation function (aside from the default "sigmoid") in the hidden layer; the end result proposed the function "satlins" with average performance:

MODEL	MSE	COVARIANCE ERROR	ACCURACY
Optimized Clustered	1.1317e-04	3.6234e+04	92.91%

However, the overall best result was made using the default activation function.

Result and Conclusion

Preprocessing and clustering data increased significantly the performance of the networks compared to the basic one while also decreasing MSE, especially with the cluster approach. However the covariance appears almost unchanged, this could be caused by the low diversity of the dataset; in fact, the data were made only from 30 days of data collection (day of the year variable starts from 122 and finishes to 152). A suggestion should be adding data from different seasons to have a larger view of the problem.
The high performance of the networks with just 10 epochs remarked the problem of the poor variety of the dataset, so further training with an augmented dataset will be indeed usefull and may result in a real application of this Neural Network.

References

[1] *K-means clustering*, May 2023. [Online]. Available: https://en.wikipedia.org/wiki/K-means_clustering.

[2] *Elbow method (clustering)*, Feb. 2023. [Online]. Available: https://en.wikipedia.org/wiki/Elbow_method_(clustering).

[3] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257.