Unit 6: Assignment

CIS625 – Machine Learning for Business

Neural Network Replication Using the Wine Quality Dataset

Najse Foster

Professor JC Martel

M.S. in Information Systems and Business Analytics

Park University

April 26, 2025

Introduction

The objective of this project was to replicate a neural network example from the course lecture material, using an independently selected dataset. The chosen dataset for this assignment is the Wine Quality Dataset from the UCI Machine Learning Repository. This dataset was selected due to its real-world applicability, moderate complexity, and suitability for binary classification tasks. The following sections describe the rationale behind the dataset selection, the preprocessing steps, the model building process, the evaluation methods, and the outcomes generated by the program.

Dataset Selection

The Wine Quality Dataset was selected because it offers a structured yet nuanced classification challenge. It contains various physicochemical measurements such as pH, alcohol content, and citric acid levels for red wine samples, each associated with a quality score assigned by sensory assessors. This dataset is advantageous for machine learning replication because it is clean, publicly available, and represents a classification problem based on continuous features.

Additionally, the dataset’s balance between complexity and manageability makes it ideal for building a neural network without requiring advanced computational resources. The classification task predicting whether a wine is of "good" quality provides an interpretable and measurable outcome, aligning well with the original intent of the assignment to replicate a neural network model and discuss the results in a comprehensive manner.

Process and Methodology

The project workflow mirrored the general architecture introduced in the lecture. The process included loading and preprocessing the data, building a neural network model, training it, evaluating its performance, and visualizing the results.

1. Data Preprocessing

The dataset was imported using pandas. Initially, the target variable, quality, was transformed into a binary label where wines rated 7 or higher were classified as "good" (1), and the rest as "not good" (0). This binarization simplified the task into a binary classification problem.

The feature set (X) and target labels (y) were separated. A standard 80/20 train-test split was applied to allow the model to learn from most of the data while reserving unseen data for evaluation. StandardScaler from scikit-learn was used to normalize the feature values. Normalization ensured that features with larger scales did not disproportionately influence the model during training, leading to faster convergence and more balanced learning across features.

1. Model Building

Rather than using TensorFlow, a neural network model was constructed using scikit-learn’s MLPClassifier, which provides a fully connected multi-layer perceptron implementation. The model architecture included:

* An input layer accepting the standardized features
* Two hidden layers with 64 and 32 neurons respectively, both using the ReLU activation function
* An output layer for binary classification using the logistic sigmoid function

The model was configured to train for a maximum of 500 iterations with the Adam optimizer to ensure sufficient convergence without overfitting.

1. Model Training

The model was trained using the training dataset. Internally, MLPClassifier utilized backpropagation to update weights and biases across the layers. Training progressed until either convergence was achieved or the maximum number of iterations was reached.

1. Model Evaluation

After training, predictions were made on the test dataset. Model evaluation included several components:

* Training and testing accuracy were printed to assess the model’s generalization.
* A classification report was generated, presenting precision, recall, f1-score, and support for each class.
* A confusion matrix was plotted to visually analyze true positives, false positives, true negatives, and false negatives.
* The classification report was also saved to a text file (classification\_report.txt) for documentation purposes.

The classification report provides a detailed view of the model’s predictive capability, while the confusion matrix offers an immediate, intuitive visualization of correct versus incorrect predictions.

Results

The model achieved a training accuracy of approximately 91 percent and a testing accuracy of approximately 86 percent. These results indicate that the model was able to generalize well from the training data to unseen data, with a moderate drop-off between training and test performance—a healthy sign that overfitting was minimal.

The classification report showed:

* Precision and recall values above 0.70 for the positive class which we would consider "good wine"
* An f1-score balanced across both classes
* A weighted average f1-score of around 0.85

The confusion matrix further confirmed the findings, showing a higher true positive rate for the majority class and a reasonable performance in identifying positive cases. The model did exhibit some bias toward the majority class ("not good" wines), reflecting the dataset’s slight class imbalance, but overall performance was acceptable given the limited data volume.

Discussion

The results demonstrate that a relatively simple feed-forward neural network is capable of performing binary classification tasks effectively when preprocessing and model architecture are thoughtfully managed. The normalization of input features played a crucial role in ensuring that the optimization algorithm converged efficiently.

Using two hidden layers allowed the model to capture more complex relationships among the chemical properties without introducing significant overfitting. The evaluation metrics suggest that while improvements could be made, such as balancing the dataset or fine-tuning hyperparameters, the basic architecture and process were sound.

Furthermore, visualizing the confusion matrix added depth to the evaluation, allowing a straightforward interpretation of the types of errors the model was making. Saving the classification report also ensured that documentation of results was preserved for future reference or review.

Conclusion

This project successfully replicated the structure and methodology outlined in the course lecture, using a different dataset to reinforce the learning objectives. The Wine Quality dataset proved to be an excellent choice, offering a challenging yet manageable classification task.

The preprocessing, modeling, training, and evaluation steps were executed following standard machine learning best practices. The added components of the confusion matrix visualization and results saving further enhanced the clarity and professionalism of the output. This replication confirmed the versatility of neural networks in handling structured, tabular data and highlighted important considerations in model building, such as feature scaling, overfitting management, and performance evaluation.

References

* UCI Machine Learning Repository. (2024). Wine Quality Data Set. <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>
* Scikit-learn Documentation. (2024). Neural network models (supervised). <https://scikit-learn.org/stable/modules/neural_networks_supervised.html>