Machine Learning Application to

Mouse Viral Infection Study Dataset

Using Python, MLFlow, and Databricks

Dhinaz Rangasamy

Overview

This project focused on analyzing data from a controlled laboratory study of viral infections in mice, with the goal of developing a predictive model to determine infection status based on administered medication volumes. The entire pipeline was implemented on Databricks, leveraging its scalable environment and integrated tools for experiment tracking and reproducibility.

Dataset Description

The dataset consists of measurements from a laboratory study in which mice were either infected with a virus or left uninfected. The data includes:

- **Med_1_mL**: Volume (in mL) of the first medication administered to each mouse.
- **Med_2_mL**: Volume (in mL) of the second medication administered to each mouse.
- **Virus Present**: Binary target label indicating infection status (1 = Infected, 0 = Healthy).

This dataset was designed to help explore whether computational models, specifically Support Vector Machines (SVM), can accurately classify infection status based on medication-related features. It is ideal for binary classification tasks in a biomedical context.

Data Preparation

A targeted preprocessing framework was applied, including:

- **Column name standardization**: Converted all field names to lowercase and replaced spaces with underscores to ensure consistency and avoid parsing errors.
- **Missing value imputation**:
    - Numerical columns (med_1_ml and med_2_ml) were imputed using mean values.
    - No categorical columns were present.

These steps ensured the dataset was clean, consistent, and ready for model training.

Modeling Approach

A Support Vector Machine (SVM) classifier was selected as the sole modeling approach, focusing exclusively on its capabilities for binary classification.

Key steps included:

- **Train-test split**: The dataset was divided into training and test subsets to evaluate the model's generalization performance (80/20 was used)
- **Hyperparameter optimization**:
    - Hyperopt was used to perform systematic hyperparameter search for the SVM model, tuning key parameters such as kernel type and regularization strength.
    - The optimization objective was to maximize the Area Under the ROC Curve (AUC).
- **Model evaluation**:
    - The final model was evaluated using AUC and accuracy metrics.
    - The best configuration achieved an AUC score of **1**, indicating perfect discrimination between infected and non-infected mice on the test data.

Experiment Management with MLflow

MLflow was used throughout to track experiments and manage model artifacts:

- **Parameters and metrics logging**: All hyperparameter configurations and evaluation metrics (AUC, accuracy) were systematically logged.
- **Artifact storage**: Models were saved as reproducible artifacts to allow for future comparison and reuse.
- **Run tracking**: Each experiment run was versioned and documented, providing full traceability of modeling decisions.

This rigorous tracking approach ensured that results could be reproduced and shared confidently.

Outcomes

- Successfully developed and validated an SVM model that perfectly classified mice as infected or healthy (AUC = 1).
- Established a reproducible and minimal preprocessing framework tailored to the experimental data structure.
- Demonstrated the practical application of systematic hyperparameter tuning (Hyperopt) in combination with robust experiment tracking (MLflow).
- Highlighted the potential for medication-related features to predict infection status effectively.

Conclusion

This project illustrates a clear and effective approach to solving a binary classification problem in a biomedical research context using Databricks. The combination of SVM modeling, Hyperopt optimization, and MLflow experiment management showcases best practices in modern machine learning workflows. The results demonstrate the strong predictive potential of medication-based features for detecting viral infections in controlled experimental settings.