FTML - Exercise 5

Classification on a Given Dataset

Dieu Marc

July 5, 2025

Objective

Given a classification dataset located in `FTML/Project/data/classification/`, the goal is to achieve a test accuracy strictly greater than **0.85** using one of the five classifiers allowed in the statement:

Allowed Models (scikit-learn):

- `LogisticRegression`
- `SVC`
- `KNeighborsClassifier`
- `MLPClassifier`
- `AdaBoostClassifier`

Initial Exploration

- Dataset size: `X_train.shape = (N, D)`, `X_test.shape = (M, D)`
- Class distribution analyzed via `np.bincount(y)`
- Feature distribution: `min, max, mean, std` observed to guide preprocessing.

Step 1 – Baseline Models

Each model was evaluated using 5-fold stratified cross-validation and tested directly on the test set without preprocessing.

Model	CV Mean	CV Std	Test Accuracy
SVC	0.7815	0.0195	**0.7950**
KNeighbors	0.7710	0.0166	0.7770
Logistic Regression	0.7140	0.0160	0.7435
MLPClassifier	0.7195	0.0210	0.7400
AdaBoostClassifier	0.7030	0.0150	0.7415

Conclusion: SVC is the most promising model.

Step 2 – Preprocessing Impact

Four preprocessing techniques were evaluated with SVC:

Preprocessing	CV Mean	Test Accuracy
None	0.7815	0.7950
StandardScaler	0.7550	0.7925
RobustScaler	0.7555	**0.7970**
QuantileTransformer	0.7485	0.7935

Observation: Preprocessing did not lead to significant gains.

Step 3 – Hyperparameter Tuning (GridSearchCV)

A grid search was conducted on SVC with multiple kernels and extensive grids on `C` and `gamma`. The best models reached around **0.80–0.82** on the test set.

Step 4 – Other Models

MLP and AdaBoost variants were tested. No configuration exceeded 0.76 on the test set.

Step 5 – Bayesian Optimization (Optuna)

Using Optuna for search over a wide range:

- `C` in [0.001, 1000] (log-uniform)
- `gamma` in [0.001, 10] (log-uniform)
- `kernel`: `poly`, `rbf`, `sigmoid`
- `degree` for polynomial: [2, 3, 4, 5]

Step 6 – Best Model

Final configuration found (SVC):

- `kernel = 'poly'`
- `C = 0.0047`
- `gamma = 0.1599`
- `degree = 3`
- No preprocessing used

Results:

- Cross-validation accuracy: $\sim$0.79
- Test accuracy: **0.9070**
- Confusion Matrix, Precision, Recall, F1 all computed
- Robustness verified through variations around optimal parameters

Conclusion

- The **SVC polynomial kernel** dramatically outperformed other models.
- **No preprocessing** yielded best results, defying standard expectations.
- **Bayesian optimization (Optuna)** proved essential in finding high-performing parameters.
- The target accuracy of **0.85** was exceeded with margin: **+0.057**.

Final insight: The classification problem required careful tuning in unconventional parameter zones. A regular GridSearch would not have succeeded without very fine granularity. Exploration beyond the standard kernel (`rbf`) was critical to success.