Practice 2: Supervised Learning in Music Datasets

Machine Learning

Grado en Ingeniería Informática y Tecnologías Virtuales

Ingeniería de Datos I

Grado en Ingeniería del Software

Academic year 2025/2026

Antonio M. Durán Rosal

Practice 2: Supervised Learning in Music Datasets

The main goal of this practice is not only to apply Python libraries for Machine Learning, but also to learn how to **interpret the results** obtained and to **justify the decisions** taken throughout the process.

For this reason, the practical evaluation will take into account both the quality of the code and the explanations or comments included in each required section.

For the development of this practice, we will use the dataset available at `https://www.kaggle.com/datasets/purumalgi/music-genre-classification`. Specifically, we will work with the `train.csv` subset, which will be treated as our complete database.

The dataset contains a collection of songs, each described by a set of numerical features such as *danceability*, *energy*, *loudness*, *speechiness*, and other audio characteristics that represent the style and mood of the music.

In this practice we will address two types of **supervised learning problems** using the same dataset:

- **Regression:** predicting the numerical variable `popularity`, which represents how popular a song is on Spotify.
- **Classification:** predicting the categorical variable `genre`, which indicates the musical genre of each song (e.g., pop, rock, classical, etc.).

Both problems will be approached using three basic but fundamental algorithms: **ZeroR**, **OneR**, and **K-Nearest Neighbors (KNN)**. Through these models, you will learn to:

- Build simple baselines and interpret their significance.
- Evaluate model performance using appropriate metrics.
- Compare predictive results between regression and classification tasks.

Feature Selection and Dataset Preparation (0.5)

Before training any supervised learning models, it is essential to clearly define which features are relevant for our task.

In this section, you must select the subset of features that you consider appropriate for the analysis. Remember that not all variables necessarily contribute valuable information to the prediction, and part of your work is to reason about your selection.

At the end of this section, you should have created two separate datasets:

- `music_regression.csv`: containing the variables used to predict the numerical feature `popularity`.
- `music_classification.csv`: containing the variables used to predict the categorical feature `genre`.

Regression (4)

For the regression task, we will work with the `music_regression.csv` dataset using three different training configurations:

a) **Single Holdout 75-25:** randomly split the data into 75 % for training and 25 % for testing.

b) **10 Repeated Holdouts 75-25:** repeat the random 75-25 split ten times and report the mean and std performance.

c) **5-Fold Cross-Validation:** divide the dataset into five folds, training on four and testing on the remaining one in rotation.

For each of the validation configurations above, you must train and evaluate the following algorithms.

a) **ZeroR** (baseline model - predicts the mean value of the target variable).

b) **OneR** (rule-based model using a single attribute).

c) **KNN regression** with $k = 3$, 5, and 10.

Note that **the input variables must be standardized** using `StandardScaler` fitted on the training set only.

For each model, compute the four regression metrics discussed in theory on the test sets (e.g., `MAE`, `MSE`, `RMSE`, and R^2).

Then, answer the following questions:

a) Compare the results obtained using each validation technique.

b) Are there noticeable differences depending on the validation method used?

c) Which validation technique would you choose to report your final results, and why?

d) According to that validation method, which of the tested algorithms performs best?

e) If you trained OneR using the entire dataset (without splitting), what rule would it produce?

Classification (4)

In this section, we will use the dataset `music_classification.csv` to build and evaluate models that predict the `genre` of a song based on its audio characteristics.

As in the regression task, we will experiment with three validation configurations:

a) **Single Holdout 75-25:** use 75 % of the data for training and 25 % for testing.

b) **10 Repeated Holdouts 75-25:** repeat the random 75-25 split ten times and report the mean and std performance.

c) **5-Fold Cross-Validation:** divide the dataset into five folds, using each one as a test set once.

For each of the above validation schemes, train and evaluate the following models. Remember that the features must be **standardized** using `StandardScaler` fitted on the training data only.

a) **ZeroR:** baseline classifier that always predicts the majority class.

b) **OneR:** rule-based classifier that uses a single feature to predict the class.

c) **K-Nearest Neighbors (KNN):** with $k = 3, 5,$ and 10.

For each model and validation technique, you should compute and report the following **classification metrics** on the test sets:

- `Accuracy, Kappa` - overall proportion of correctly classified instances.
- `Precision`, `Recall`, and `F1-Score` - for each class.
- `Aggregated Confusion Matrix` - to visualize which genres are most frequently confused.

Finally, answer the following questions and include your reflections:

a) Compare the results obtained by each validation technique. Are the conclusions consistent?

b) Which validation approach seems more reliable for this type of data?

c) Which algorithm performs best overall, and for which classes does it fail more often?

d) Discuss how the value of k affects KNN performance in this dataset.

e) What kind of rule does OneR generate when trained on the entire dataset? Does it make intuitive sense musically?

You are encouraged to include plots that support your conclusions, such as confusion matrices or bar charts comparing accuracies across validation methods. Clear visual analysis will strengthen your report and help you better interpret model performance.

Extra (1.5)

For a classification problem with three features, the following training patterns have been collected:

Pattern	x_1	x_2	x_3	**Class**
1	4.6	3.2	1.4	1
2	5.3	3.7	1.5	3
3	5.7	4.4	1.5	1
4	5	3.5	1.6	2
5	5.5	2.5	4	1
6	5.7	3	4.2	2
7	5.7	2.8	4.1	2
8	5.8	2.7	5.1	1
9	6.3	2.5	5	2
10	5.9	3	5.1	3

It is requested:

a) Find the model's predictions for the following test set, with the value of $K = 2$ and $K = 5$ neighbours and the Euclidean distance.

b) Evaluate the classifier's performance on this set for both values of K.

Pattern	x_1	x_2	x_3	**Class**
1	5	3.5	1.7	1
2	4.3	2.8	1.5	1
3	2.7	4.5	1.2	3
4	5	4.2	1.3	3
5	6.3	2.5	4.1	1
6	5.2	3	4.5	2
7	4.5	3	4.2	2
8	5.9	2.9	5.2	2
9	5	2.4	5.1	1
10	4.5	3.2	5	2