Lab 1

Fundamentals of regression and classification.

David Padilla Orenga
Ignacio Pastore Benaim

October 18, 2024

1 Introduction

The goal of this practical session is to train and evaluate various machine learning models for both regression and classification tasks using two datasets:

- **YearPredictionMSD dataset:** Used for the regression task. This dataset contains audio features extracted from the Million Song Dataset, with the aim of predicting the release year of a song.
- **CIFAR-10 dataset:** Used for the classification task. The CIFAR-10 dataset contains 60,000 32x32 color images across 10 different classes, such as airplanes, cars, birds, and ships.

The objectives are:

- Train and evaluate different models to predict the year of a song using audio features.
- Train models on the CIFAR-10 dataset to classify images, with a focus on using both raw pixel data and GIST descriptors for feature extraction.

All the experiments, including model training, hyperparameter tuning, and evaluations, were run locally on a MacOS machine with an M1 processor. The reported fitting and prediction times correspond to this specific hardware configuration. The local notebook containing the code is attached to this report.

In this report, we will detail the preprocessing steps, model selection, hyperparameter tuning, and evaluation metrics. Results for both tasks are analyzed, and conclusions are drawn based on model performance.

2 Methodology

2.1 Regression Task: Year Prediction (Million Song Dataset)

The YearPredictionMSD dataset consists of over 500,000 songs represented by 90 audio features extracted from each song. The target variable is the release year of each song.

2.1.1 Data Splitting

The dataset was split into training, validation, and test sets:

- 81% for training.
- 9% for validation.
- 10% for the final test.

Cross-validation was performed on the training set to select the best model.

2.1.2 Exploratory Data Analysis

We first explored the distribution of the target variable (year), noting the imbalance with more songs in recent years, as shown in Figure 1.

Additionally, we analyzed the correlations between features and the target variable to select promising

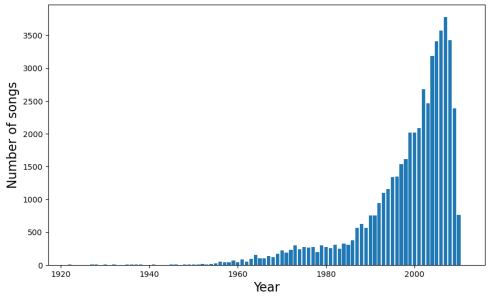

Figure 1: Distribution of songs by year in the Million Song Dataset.

features for modeling but we realized that it was better to apply PCA. Figure 2 shows correlations between the top features and the year of the song.

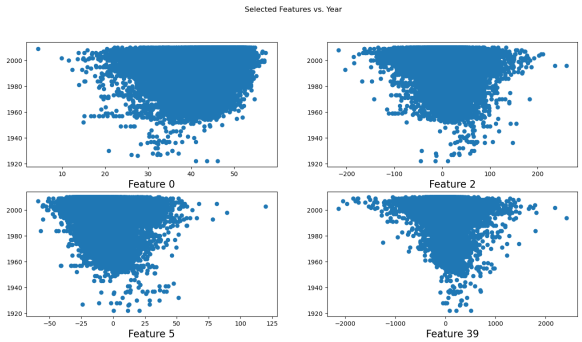

Figure 2: Selected features vs. year in the Million Song Dataset. Features where selected having correlation with output > 0.12.

2.1.3 Model Training and Hyperparameter Tuning

We trained several regression models including:

- Linear Regression
- Ridge Regression
- Lasso Regression
- ElasticNet
- Support Vector Regressor (SVR)
- Random Forest Regressor (RF)
- Gradient Boosting Regressor (GB)

Hyperparameters were tuned using GridSearchCV on the training and validation set, and models were evaluated using Mean Absolute Error (MAE), and Median Absolute Error (MedAE). The results are detailed in the following tables (Section **??**).

2.2 Classification Task: CIFAR-10

The CIFAR-10 dataset contains 60,000 images, evenly split across 10 classes. We tested different models for image classification, using both the raw pixel data and GIST descriptors.

2.2.1 Data Preprocessing

The CIFAR-10 dataset contains 60,000 images, evenly split across 10 classes. However, due to computational constraints, we limited our experiments to the first batch of 10,000 images from the dataset.

The dataset was split into training, validation, and test sets as follows:

- 8,100 images (81%) for training.
- 900 images (9%) for validation.
- 1,000 images (10%) for the final test.

We applied feature scaling to the raw pixel data and reduced its dimensionality using Principal Component Analysis (PCA) with 95% variance retained. Afterward, GIST descriptors were computed for the same subset of images, allowing us to compare the performance of models trained on raw pixel data versus GIST-based features.

Although we performed comparisons on both the training and validation sets in the notebook, the reported comparison between raw pixel data and GIST descriptors is focused on the test set. The models were trained and evaluated using these two distinct feature sets to assess the effectiveness of each representation.

2.2.2 Model Training and Hyperparameter Tuning

Several classifiers were trained, including:

- Support Vector Machines (SVM)
- Random Forest (RF)
- Multi-Layer Perceptron (MLP)
- K-Nearest Neighbors (KNN)

The models were tuned using GridSearchCV to find the best combination of hyperparameters.

Figures 3 and 4 show successful and misclassified examples from the CIFAR-10 dataset.

Figure 3: Correctly classified images from CIFAR-10 dataset.

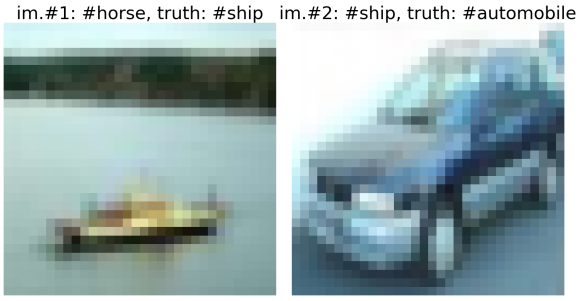

Figure 4: Misclassified images from CIFAR-10 dataset.

Table 1: Screening of regression models with default parameters and PCA=0.95 for the year-song dataset.

Model	MAE	MedAE	Time (s)
Linear	6.97	5.32	0.16
Ridge	6.97	5.32	0.08
Lasso	7.59	6.34	0.10
ElasticNet	7.43	6.11	0.09
Huber	6.73	4.68	0.60
RANSAC	25.34	18.20	0.43
SVR	6.15	3.97	125.17
RF	7.01	5.46	226.12
GB	6.77	5.09	84.34
KNN	6.68	4.80	0.07

3 Results

3.1 Regression Task: Year Prediction (Million Song Dataset)

The table 1 provides a screening of all the models.

The models were evaluated based on MAE (Mean Absolute Error) and MedAE (Median Absolute Error), as these metrics are more robust to outliers compared to RMSE. The best-performing models were the Support Vector Regressor (SVR) and Huber Regressor.

Due to time and computational constraints, we did not experiment with Huber Regressor using polynomial features, which might have further improved its performance. Additionally, K-Nearest Neighbors (KNN) was included in the comparison due to its simplicity and ease of implementation.

The following table summarizes the results of cross-validation for the trained models.

All the models in the cross validation step were selected for the final test, shown in Table 3.

In summary, SVR and Huber stood out as the top performers, with KNN serving as reference.

3.2 Classification Task: CIFAR-10

For the classification task, the models were evaluated based on several metrics, including accuracy, precision, recall, and F1-score. The performance

Table 2: Cross validation of best models. Huber($\alpha = 0.001, \epsilon = 1.35$), KNN($n\_neighbors = 7, p = 2$, weights = 'uniform'), SVR($C = 10.0, \gamma =$ 'scale', kernel = 'rbf'). Time = time elapsed for all folds. Metrics are the mean over the 5 folds.

Model	MAE	MedAE	Time (s)
Huber	6.80	4.69	1.2
SVR	6.01	3.89	960.0
KNN	6.79	4.91	10.9

Table 3: Test performance of best models. Huber($\alpha = 0.001, \epsilon = 1.35$), KNN($n\_neighbors = 7, p = 2$, weights = 'uniform'), SVR($C = 10.0, \gamma =$ 'scale', kernel = 'rbf').

Model	MAE	MedAE	Fit. T.(s)	Pred. T.(s)
SVR	5.91	3.89	638.19	26.90
Huber	6.76	4.67	0.91	0.00
KNN	6.70	5.00	0.08	0.44

of the models was first assessed using a screening approach to identify potential candidates, followed by hyperparameter tuning through cross-validation. Finally, the best models were tested on the holdout test set to evaluate generalization.

3.2.1 Model Screening

The initial screening of models was performed using both raw pixel data and GIST descriptors to compare their effectiveness in representing image features for classification. The results from this screening process are summarized in Table 4, where only the metrics for the raw pixel data are shown.

From the screening, we observed that models trained on raw pixel data showed significantly better performance, with the Support Vector Classifier (SVC) and MLP among the top-performing models. Models trained on GIST descriptors struggled to achieve comparable accuracy due to the lower dimensionality and feature representation limitations of the descriptors.

Table 4: Screening of classification models with default parameters and PCA=0.95 for the CIFAR-10 dataset.

Model	Acc.	F1	Prec.	Rec.	Fitting T. (s)
LR	0.37	0.37	0.37	0.37	21.98
SVM	0.46	0.46	0.46	0.46	29.40
KNN	0.31	0.30	0.39	0.31	28.11
RF	0.35	0.34	0.35	0.34	40.01
MLP	0.39	0.39	0.39	0.39	56.40

Table 5: Cross-validation performance of classification models wit PCA=0.95. SVC($C = 1$), MLP($hiddenlayers = (100,)$). Time = cross-validation time over 5 folds.

Model	Acc.	F1	Prec.	Rec.	T.(s)
SVC	0.4581	0.4561	0.4587	0.4580	87.13
MLP	0.3786	0.3784	0.3795	0.3782	82.19

3.2.2 Cross-Validation Results

After the screening, we performed hyperparameter tuning using GridSearchCV on the validation and training set to find the best hyperparameter combinations for the top-performing models. The results of this cross-validation process are shown in Table 5, where models were evaluated across 5 folds for stability and generalization.

The SVC with a radial basis function (RBF) kernel consistently performed the best across different folds, achieving the highest average accuracy. Also MLP performed similarly, but SVC was elected for testing due to computational and time constraints. GIST-based models generally showed lower cross-validation scores.

3.2.3 Test Set Results

Finally, SVC was evaluated on the test set. Table **??** presents the final performance over raw pixel data and GIST descriptors. There is a clear improvement in performance when using raw pixel data. This could be due to the fact that we are using a subset of the dataset, and the GIST descriptors may not be able to

Table 6: Test performance of SVC models over the CIFAR-10 dataset. SVC + G = SVC + GIST. Both SVC($C = 1$). $SVC + G()$ without PCA.

Model	Acc.	F1	Prec.	Rec.	Fit. T.(s)
SVC	0.47	0.46	0.46	0.46	43.91
$SVC + G$	0.25	0.24	0.24	0.25	7.63

capture the necessary information for classification.

4 Conclusion and Discussion

In this report, we evaluated several machine learning models for both regression and classification tasks.

4.1 Regression: Year Prediction Task

For the YearPredictionMSD dataset, we found that the Support Vector Regressor (SVR) performed the best, achieving a Mean Absolute Error (MAE) of 5.91 years. Despite the strong performance of SVR, it was computationally expensive, making it less practical for large-scale applications.

K-Nearest Neighbors (KNN) was also evaluated and selected for its simplicity and ease of implementation. While it did not outperform SVR, KNN provided a practical baseline with acceptable performance for comparison.

RANSAC, on the other hand, struggled significantly with the complexity and outliers in the dataset, resulting in poor overall performance.

4.2 Classification: CIFAR-10 Task

In the CIFAR-10 classification task, models trained on raw pixel data consistently outperformed those trained on GIST descriptors. The best-performing model, SVC with an RBF kernel, achieved an accuracy of 47%. However, it was observed that GIST descriptors, while faster to train on, r esulted in a significant loss of classification accuracy.

A possible explanation for this drop in performance when using GIST descriptors is the smaller subset of images used in this experiment. With only 10,000 images (instead of the full 60,000), the GIST descriptors may not have captured enough variability in the dataset, limiting their effectiveness as feature representations.

Future work could explore deep learning models such as Convolutional Neural Networks (CNNs) for image classification, as they are known to perform better on image datasets like CIFAR-10 by automatically extracting more robust features. Additionally, training with the full CIFAR-10 dataset may help improve the performance of models based on GIST descriptors, as the larger sample size would allow for better feature extraction and learning.