Lab 1

Fundamentals of regression and classification.

David Padilla Orenga
Ignacio Pastore Benaim

October 18, 2024

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

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

1 Introduction

The complete code used for this analysis can be found in a Google Colab Notebook.

2 Introduction

Recordar que

no hice poly2 en gridsearch porque explotaba el computat SVR lo mismo en el gridsearch Ademas todos los gridsearchs dan un R2 negativo, por que? Es mejor separar todos los gridsearc para no tener que volver a correrlos todos Antes de hacer los gridsearch tengo que hacer test separados para ver si tiene sentid o no.

In this lab, we aim to solve two fundamental machine learning tasks: regression and classification.

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	Fitting T(s)	Predicting 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

Both tasks utilize different datasets and methods, and each presents unique challenges in model training and evaluation. For the regression problem, we utilize the **YearPredictionMSD** dataset, which aims to predict the release year of songs based on a variety of acoustic features. For the classification task, we use the **CIFAR-10** dataset, a widely recognized dataset for image classification, where the objective is to classify images into one of 10 possible categories.

2.1 Regression Task

The **YearPredictionMSD** dataset consists of over 500,000 examples, each representing a song with a collection of numerical features that describe its acoustic properties. The target variable is the year the song was released. This problem is inherently non-linear and affected by noise, making it an excellent case for testing various regression techniques, including linear regression, polynomial regression, and robust methods like RANSAC and Huber regression. Furthermore, due to the large number of features, dimensionality reduction techniques such as Principal Component Analysis (PCA) are explored to improve computational efficiency and model performance.

2.2 Classification Task

For the classification task, we work with the **CIFAR-10** dataset, which consists of 60,000 32x32 color images from 10 different classes (e.g., airplanes, cats, cars, etc.). Each image is represented as a feature vector of pixel values. The objective is to classify these images accurately into their respective categories using models such as k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Voting Classifiers. Dimensionality reduction is again considered through PCA to speed up training, particularly for computationally intensive models like SVM.

Both tasks involve model selection, hyperparameter tuning, and evaluation using appropriate metrics. In regression, we will use metrics such as Mean Squared Error (MSE) and the Median Absolute Error (MedAE), which is robust to outliers. For classification, accuracy, precision, recall, and F1-score will be employed to measure model performance. Through this lab, we aim to gain a deeper understanding of model behavior in different scenarios and explore techniques to address the challenges presented by high-dimensional and noisy data.