Lab 1

Fundamentals of regression and classification.

David Padilla Orenga
Ignacio Pastore Benaim

October 18, 2024

1 Introduction

- En este laboratorio se realizan tareas de regresion sobre el dataset year-song y clasificacion sobre el dataset CIFAR-10. En las dos secciones se sigue la misma metodología. Primero un screening de modelos con los parametros por defecto y PCA=0.95 (explicar como se eligio el PCA) . Luego gridsearch, luego cross-validation con 5 folds y Finalmente se se eligen los mejores modelos para probar sobre el data test. - AClarar que en el CIFAR-10 solo se utilizó el primer batch de los 5 (usando 10000 imagenes) por cuestiones de tiempo de computo.

- Finalmente para el dataset CIFAR-10 se realiza una busqueda de descriptores con GIST. Gracias a esto se ve un ahorro de tiempo de computo considerable pero bajan las meetricas por la mitad. Esto podria solucionarse alimentando el modelo con todo el dataset.

- En el year son dataset: no se utilizaron features polinomailes por que era costoso computacionlmente. Tampoco se pudo explorar en profundidad los parametros para el SVR y el MLP por la misma razon

- En cuanto a metricas: explicar que se utilizo MAE y MedAE (dejando afuera a R2 por que es un dataset completo y a RMSE o MSE porque hay datos espureos) para regresion y Acc, F1, Prec y Rec para clasificacion. Aunque no se muestra en el reporte por falta de espacio, en el notebook también se utilizo confusion matrix.

- Explicar el dataset de year song - Explicar el dataset de CIFAR-10

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($n\_neighbors = 7, p = 2, \text{weights} = \text{'uniform'}$), SVR($C = 10.0, \gamma = \text{'scale'}, \text{kernel} = \text{'rbf'}$). 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	Fitting T(s)	Predicting T
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

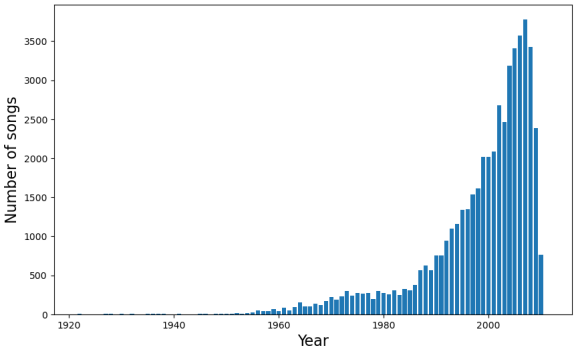

Figure 1: Quantity of songs per year in dataset.

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

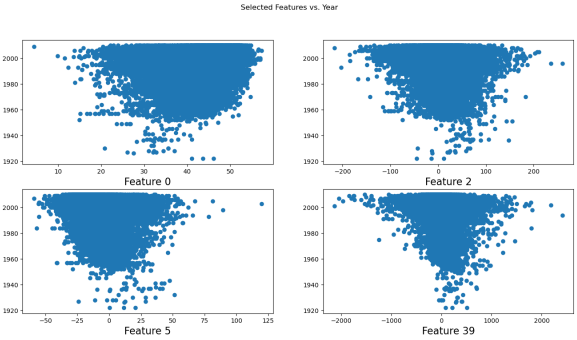

Figure 2: Distribution and relation with output of features which correlation with output ¿ 0.12.

Table 5: Cross-validation performance of classification models wit PCA=0.95. SVC($C = 1$), MLP($hidden_layer_sizes = (100,)$). 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

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

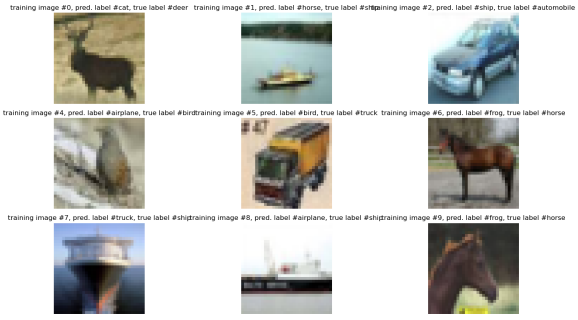

Figure 3: Missclasification of 9 images with SVC($C = 1$) and PCA=0.95 over the CIFAR-10 dataset.

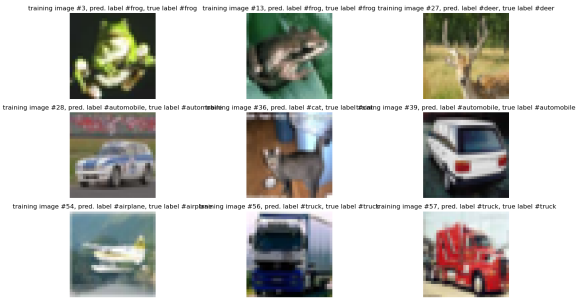

Figure 4: Correct classification of 9 imagaes with SVC($C = 1$) and PCA=0.95 over the CIFAR-10 dataset.