SMAI

Mini project-2

Report

Prakyath. M

20161236

- The zip folder consists of ‘main.py’ and report. To run the ‘main.py’ file data folder should consist of the CIFAR-10 data files. This data folder should be placed within this project folder.
- Run the code as follows:

  python3 main.py <arg1> <arg2>
- It consists of two arguments arg1 and arg2. ‘arg1’ specifies the type of dimensionality reduction and ‘arg2’ specifies the type of classifier to be used.
- ‘arg1’ can be ‘nan’, ‘pca’, ‘lda’
- ‘arg2’ can be ‘SVM_linear_soft’, ‘SVM_rbf’, ‘mlp’, ‘decision_tree’, ‘logistic regression’.
- The hyperparameter are set through ‘kargs’ in function calls.

Table of accuracies and F1 scores

Type of data	Classifier	Accuracy	F1 score
Raw	Logistic Regression	0.4021	0.4020
Raw	SVM_linear	0.3521	0.3520
Raw	Decision Tree	0.2605	0.2605
Raw	MLP	0.4565	0.4565
PCA	Logistic Regression	0.3985	0.3985
PCA	SVM_linear	0.3471	0.3471
PCA	Decision Tree	0.2542	0.2542
PCA	MLP	0.4665	0.4665
LDA	Logistic Regression	0.3801	0.3801
LDA	SVM_linear	0.3864	0.3864
LDA	Decision Tree	0.3056	0.3056
LDA	MLP	0.3710	0.3709

- A general trend observed was that the accuracy of Raw > PCA > LDA.
- Also the run time order : Raw > PCA > LDA
- Run time order for classifiers :

  LR < Decision tree < MLP < SVM_linear < SVM_rbf

Hyperparameter Tuning

Logistic Regression

Varied the C-value, which is the inverse of regularization strength, using the lbfgs optimizer. Greater value of C implied greater regularization strength, also the accuracy decreased when compared to lower C values. Optimal near C=0.0001 for LDA, C=0.0005 for PCA and C=0.001 for Raw data.

Soft margin SVM linear

Varied the C parameter, which is used to set the allowed penalty for the soft margin. Greater values of C reduced the accuracy of the classifier. Optimal near C=1 for LDA, C=0.001 for PCA and C=0.001 for Raw data.

Decision Tree

Varied the max depth from 3 to 15 for LDA, 50 to 70 for PCA and 2990 3100 for Raw data. Accuracy increased for some time and then decreased beyond the optimal point. Optimal max depth around 10 for LDA, 58 for PCA and 2999 for Raw data.

MLP

Varied the type of activation function, Relu and Tanh. Accuracies for Relu are better than Accuracies of Tanh for the same hyperparameters. Also varied the learning rate initializations and alpha values. Optimal accuracy was obtained at learning rate = 0.001 for LDA and PCA learning rate = 0.0001 for Raw data. Lower values of alpha() gave better accuracies than higher values of alpha.

Problems

OverFitting and Early Stopping

Comparing the accuracies of the test data by varying number of maximum iterations while training the classifier helped finding the early stopping point for all the classifiers. It can be observed that beyond this point the accuracy of the classifier decreases or remains constant.

Run time

CIfAR-10 data has a size of 50000 x 3012 for the training dataset and 10000x3072 for the testing dataset. Processing of this data takes a long period of time (hours to train and test a single classifier)

Also the accuracies are pretty low to consider real life applications.