Course Project on CIFAR-10 Dataset

Kwanit Gupta*, Hiteshi Singh†, and Zeba Karkhanawala*

*Indian Institute of Technology, Jodhpur
Pattern Recognition and Machine Learning (CSL2050)
Professor: Richa Singh
richa@iitj.ac.in

Introduction

The CIFAR-10 data set has 10 classes: aeroplane, automobile, bird,cat, deer, dog, frog, horse, ship, truck. The data set has 60,000 images to classify. The classes are completely exclusive. Training data set has 50,000 images and the testing data set has 10,000 images. The original data set being too large, it was divided into 5 training-batches and 1 test-batch having same number of images per class.

Type of Classifiers

Neural Network

Neural networks are a set of algorithms, modeled after the human brain, that are designed to recognize patterns. A neural network is composed of several layers.The layers are made of nodes. A node combines input from the data with a set of coefficients, or weights,These input-weight products are summed and then the sum is passed through a node's activation function.

K-Nearest Neighbor Classifier

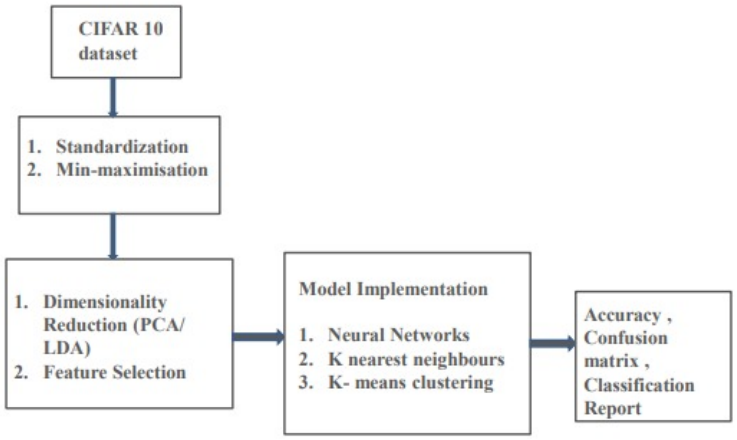
Being a Supervised model, K-Nearest Neighbor picks an example data-point and finds k nearest data-points with their corresponding labels. If the Task is Regression-based, it returns the mean of Queries, whereas for Classification-based task, it returns Mode Label of Queries.

K-Means Clustering

In K-Means clustering, the data set is partitioned into k pre-defined clusters. Each data point belongs to only one cluster. A data point is assigned to a cluster such that the sum of the squared distances between the data points and the cluster's centroid is minimum.

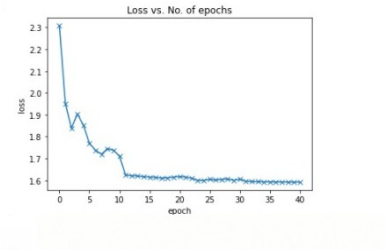
General Pipeline

A common pipeline was followed for all the three classifiers used. And it is given as follows:



Neural Network Analysis

The neural network implemented using Pytorch consists of 3 hidden layers . The optimizer used is Adam . For hidden layers , ReLu function is used as an activation function . The best accuracy is obtained for 4 hidden layers which is 44.7



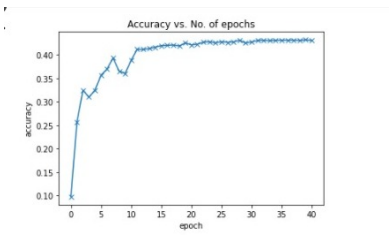


Table 1: Accuracies for Neural Network

Layers	Accuracy
4	44.7%
5	44.16%
16	42.8%

K-means Clustering Analysis

The K-means clustering model was run for different values of k such as - 10,50,100,500,1000,3000 and 4000 and the model was compared with each other. The respective accuracies are given in Table-4. And the graph to show the variation is given above. Up to certain value of k, the accuracy was increasing, but after a point the accuracy decreased.

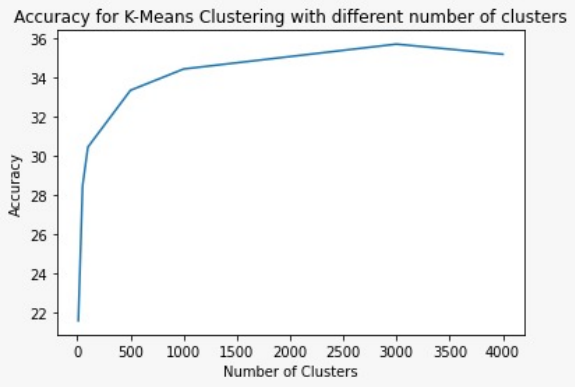


Table 2: Accuracies for K-means Clustering

Number of Clusters	Accuracy
10	21.606060606060606
50	28.484848484848484
100	30.454545454545457
500	33.34848484848485
1000	34.43434343434343
3000	35.7020202020202
4000	35.18686868686869

K-Nearest Neighbor Classifier Analysis

K-Nearest Neighbor Model was then applied for Different Data sets formed by applying Different Standardization Techniques and Dimensionality Reduction through Principal Component Analysis. Firstly, 2-D Data Scatter Plot was built and then Classification Report was generated as following :-

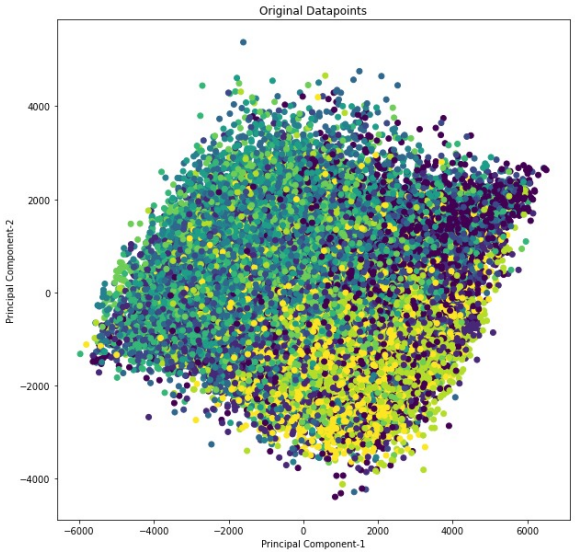


Table 3: Class-wise classification report

Classes	Precision	Recall	F1 Score
0	0.45	0.51	0.48
1	0.78	0.15	0.25
2	0.23	0.45	0.30
3	0.34	0.15	0.21
4	0.22	0.56	0.31
5	0.48	0.21	0.29
6	0.34	0.28	0.31
7	0.67	0.17	0.27
8	0.37	0.72	0.49
9	0.69	0.13	0.22

Table 4: Comparison between Original, Min-Max and Standardized data

Classes	Precision (Orig)	Precision(Std)	Recall(Orig)	Recall(Std)	F1 Score(Orig)	F1 Score(Std)
0	0.46	0.46	0.56	0.54	0.50	0.50
1	0.70	0.74	0.24	0.25	0.35	0.37
2	0.26	0.25	0.43	0.42	0.32	0.32
3	0.34	0.34	0.15	0.15	0.21	0.20
4	0.25	0.25	0.52	0.52	0.34	0.34
5	0.51	0.51	0.22	0.23	0.31	0.31
6	0.32	0.33	0.48	0.49	0.39	0.40
7	0.64	0.66	0.25	0.25	0.36	0.36
8	0.42	0.42	0.69	0.70	0.52	0.53
9	0.71	0.70	0.24	0.23	0.36	0.34

Conclusion

After implementing 3 different models with a general pipeline, it was observed that Neural Networks with 4 Layers performed best with 44.8% accuracy, K-Nearest Neighbor with Standardized PCA Transformed Data giving 38% accuracy and K-Means with 3000 clusters performed worst with 35.7% accuracy.They lack the adaptive ability to handle performance according to the nature of Data-set, like KNN and KMeans only relied on Distance Metrics. But nowadays, there are more advanced and optimizable models like Support Vector Machines, State-of-the-Art Deep Neural Networks, etc. with more control on Hyper-Parameters.

References

- `https://pytorch.org/tutorials/beginner/blitz/neural_networks_tutorial.html`
- `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html`

- https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
- https://scikit-learn.org/stable/
- https://www.cs.toronto.edu/~kriz/cifar.html

Contribution

Kwanit Gupta (B19EE046):- K-Nearest Neighbor and Data Pre-processing Techniques and Report
Zeba Karkhanawala (B19EE093):- K-Means Clustering as a Classifier and Report
Hiteshi Singh(B19EE039):- Neural Network Analysis and Report