Assignment 1

Due Date: Feb 5, 2022
Author: Tommy Lay, V00855688
Submission: includes this pdf and a code file

1 Introduction

In this assignment we were tasked to implement and train 3 different machine learning classification algorithms that being the k-nearest neighbors(KNN) algorithm, k-means clustering(K-mean) algorithm and Softmax classifier algorithm. The code for this assignment can also be found at the following github link.

1.1 CIFAR-10 Dataset

For this assignment the CIFAR-10 data set was used for each of the following classifiers below, this dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.

1.2 CIFAR-10 Normalization

All of the values for each images was where converted into numpy arrays. Afterwords, I normalized the datasets to ensure that the data is consistent. First I divide the training and testing data by 255 to get them into decimal values. Second I merged the width, height and RGB values of each image into one.

2 KNN Classifier

The K-Nearest Neighbors algorithms works by determining the values of the k closest neighbours and then determines which class appears the most frequently among them. For my implementation I decided to use an euclidean distance to determine and selected the k-values of 3, 5, 7 and 11.

$$\sqrt{\sum_{i=1}^{k}(x_1 + y_1)^2} = \sqrt{\sum_{i=1}^{k}(x_1 + y_1 - 2xy)}$$

Figure 1: Formula for the Euclidean Distance

2.1 Results

Overall I found that the general accuracy of my model sat around 30%. The results of running cross validation on the training set and results from the testing set can be found in table 1. It shows that best overall accuracy was when k-value the is equal to 7, followed closely by k-value of 5. When tested on the test set, it was found that the best k-value was 5.

k-value	Accuracy
3	0.3246
5	0.3321
7	0.3327
11	0.3294

(a) 5-Fold Cross Validation Accuracy

k-value	Accuracy
3	0.3303
5	0.3398
7	0.3358
11	0.3414

(b) Test Set Accuracy

Table 1: KNN Cross Validation and Testing Set Accuracy

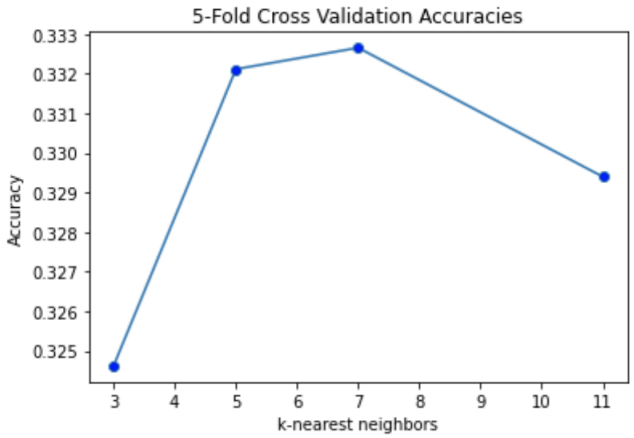

Figure 2: KNN Cross Validation Accuracy overtime

3 K-Means Classifier

K-means clustering Algorithm works by forming groups of similar clusters, the amount of clusters set are set by k. Since the dataset has 10 different classes, my implementation chooses the most few class in each cluster. Some classes will not be represented and there is a possibility of numerous of the same class being selected for clusters. But as the k-value increases so should the accuracy, since after a certain value all of the 10 classes will be represented.

3.1 Results

The results from cross validating the training set seemed to improve as the k-value increase, I believe this is because as the k-value increases past 11 it allows for all of the classes to be represented. The results for cross-validation and testing set can be found on table 2. Overall for both the cross validation and test set accuracy the k-value of 11 was the best.

k-value	Accuracy
3	0.1769
5	0.1873
7	0.1979
11	0.2342

(a) 5-Fold Cross Validation Accuracy

k-value	Accuracy
3	0.1753
5	0.1973
7	0.1998
11	0.2374

(b) Test Set Accuracy

Table 2: K-means Cross Validation and Testing Set Accuracy

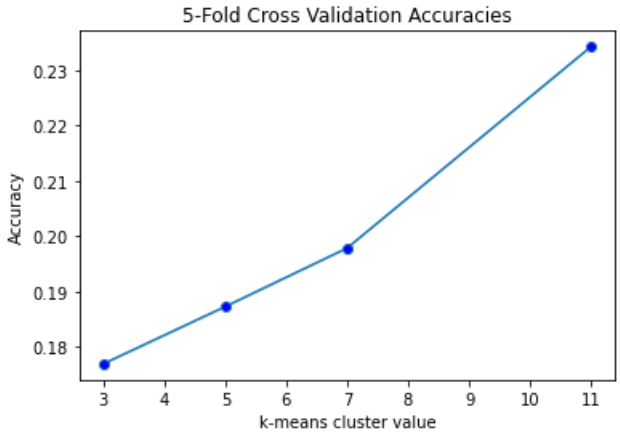

Figure 3: Kmean Cross Validation Accuracy overtime