K-means clustering and Autoencoder

Asheeque Chericham Veettil Maliyakkal
Department of Computer Science
University at Buffalo
Buffalo, NY 14260
asheeque@buffalo.edu

Abstract

K-means is used to make 10 clusters using the cifar-10 dataset.The aim of the algorithm is to make k partition for n sample which is done by finding the euclidean disitance from the centroids.

An autoencoder neural network is created to generate sparse image representation of cifar-10 data.Encoded data is used to do k-means clustering and then ASC (Average Silhouette Coefficient) and DI (Dunn's Index) is calculated.

1 Introduction

1.1 Unsupervised learning

The main difference that differentiate supervised learning from unsupervised learning is that,in unsupervised learning the data is unlabelled.Machine learning algorithms are used to detect hidden patterns which can be further analysed by humans.Unsupervised learning is mainly used for clustering,dimensionality reduction and association.They are as follows:

1.1.1 Clustering

In this data is grouped based on their differences and similarities.They can be further divided to exclusive, overlapping, hierarchical, and probabilistic clustering.

- Exclusive and overlapping cluster : The speciality of the cluster is that a point will remain in only one cluster at a time.K-means is an example of this algorithm
- Hierarchical clustering : This is further divided into two.First is agglomerative and second one is divisive.Agglomerative clustering does a bottoms up approach of grouping similar items.All the points are initially separate.Then iteratively they are clustered depending upon the similarity.Divisive can be considered as a top down approach,where data is divided in iterations based on the differences.
- Probabilistic clustering:This is used to solve density estimation or soft clustering problems

1.2 K-Means clustering

It is a method of vector quantization.Our aim is to divide the entire data to k meaningful partitions.If there are n observations $(x_1, x_2...x_n)$ we divide this into k sets $\mathbf{S} = \{S_1, S_2, ..., S_k\}$ which reduces sum of squares within the cluster.This can be defined as:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} || \ x - \mu_i \ ||^2 = \arg\min_{S} \sum_{i=1}^{k} | \ S_i \ | \ VarS_i$$

where μ_i is the mean of points in S_i.

1.2.1 Algorithm

It is an iterative algorithm.It works as follows:

1. Fix the number of clusters **K**, to be formed
2. Take K random points initially and assign them as centroids
3. Compute sum of squared distance between the points and the centroid.
4. Assign the points into different groups such a way that their distance is least from the selected centroid.
5. Compute the new centroid by taking the mean of all samples in the cluster.
6. Do step 3 to 5 till they converge
7. return the final cluster

1.3 Autoencoder

An autoencoder is an unsupervised artificial neural network that is used to do encoding of unlabelled data.It consists of two parts.

- Encoder : Encoders are used to compress the data and the inputs are mapped to code.
- Decoder : Decoders are to generate the orginal input from encoded data.

Autoencoders are data specific and lossy.Autoencoders can compress only data similar to the training data and the the final outputs are usually degraded. compressed representation of data are also known

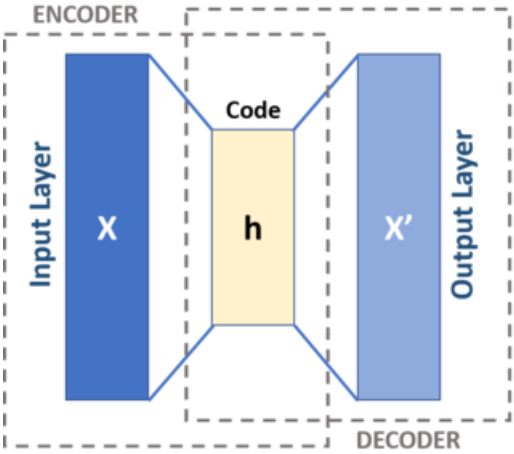

Figure 1: A basic Autoencoder.

as latent variables.If its represeneted by *h*, it can be written as :

$$h = \sigma(Wx + b)$$

where W is a weight matrix,b is bias and x is the input.σ is the activation function.Initially weights and bias are random values which is then updated by backpropagation iteratively.Then the reconstructed data is generated by decoder from the output of encoder.It can be written as:

$$x' = \sigma'(W'h + b')$$

where x' is the reconstructed data ,w' is the weight matrix and b' is the bias.If x is the input and x' is the final output then during training,we define a loss function such as mean squared error and loss is minimised.It can be written as:

$$L(x, x') = || \ x - x' \ ||^2$$

2 Dataset

2.1 CIFAR-10 Dataset

CIFAR-10 dataset contains around 50000 training images and 10000 test images.They can be broadly classified into 10 categories.They are airplane,automobile,bird,cat,deer,dog,frog,horse,ship and truck.

3 Preprocessing

Preprocessing the data, converts the data to a form that can be understood.

3.1 K-Means clustering

According to question,the model is trained on test dataset of CIFAR-10 dataset.There are around 10000 images. First images are converted from RGB to greyscale.Then the images are normalized by dividing all the images by 255.Then each sample is converted to a 1d vector containing 1024 features.Then K-means clustering is performed on the same.

3.2 Autoencoder

For autoencoder,training dataset of cifar-10 is used.It contains around 50000 images.Images are normalized by dividing with 255.

4 Experimental setup

4.1 K-Means clustering

The number of iterations is taken as 200.Since we know CIFAR-10 contains 10 classes,we use k as 10.We do iterations till we get the centroids same as that of previous iteration.We stop the iterations and return the current cluster groups.Finally ASC (Average Silhouette Coefficient) and DI (Dunn's Index) were calculated to check the quality of clusters

4.2 Autoencoder

For autoencoder mean square error is chosen as the loss function.Adam is used as the optimiser.Number of epoch is taken as 5 since the loss after epoch 5 was minimal.For encoder part,2 convolutional layer and maxpooling for downsampling is used.For decoder,2 deconvolutional layer is used,then the output is passed through sigmoid activation function to get final output.

5 Result

5.1 K-Means clustering

K-Means clustering was done on test dataset of CIFAR-10 dataset.The model achieved Average Silhouette Coefficient of 0.06 and a dunn score of 0.091.

5.2 Autoencoder

A convolutional autoencoder was created.The model was trained on training dataset of cifar-10.The encoded data that is the output from encoder was used to do K-Means clustering with number of clusters as 10. Then a positive value for ASC (Average Silhouette Coefficient) and DI (Dunn's Index) was calculated.

6 Conclusion

CIFAR-10 dataset was divided to 10 cluster using k-means algorithm.An autoencoder model was created and trained on cifar-10 training dataset.Then K-means clustering was done and ASC and DI was calculated and found to be 0.04 and 0.1 respectively.

References

[1] K-Means-wikipedia

[2] IBM-unsupervised-learning

[3] k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks

[4] Autoencoder-wikipedia

[5] Autoencoder-keras-blog