Report

Hang Yao

1. Overview

This report is to demonstrate the experiment of kernel density estimation with mixture of Gaussians on datasets of MNIST and CIFAR100.

2. Implementation

The project is split into two parts: (1) preprocessing of the image data (*preprocess.py*), and (2) computing the mean of the log-probability (*computeMLP.py*).

Preprocessing includes loading data from downloaded files, shuffling data, splitting data into training and validation datasets, and visualization.

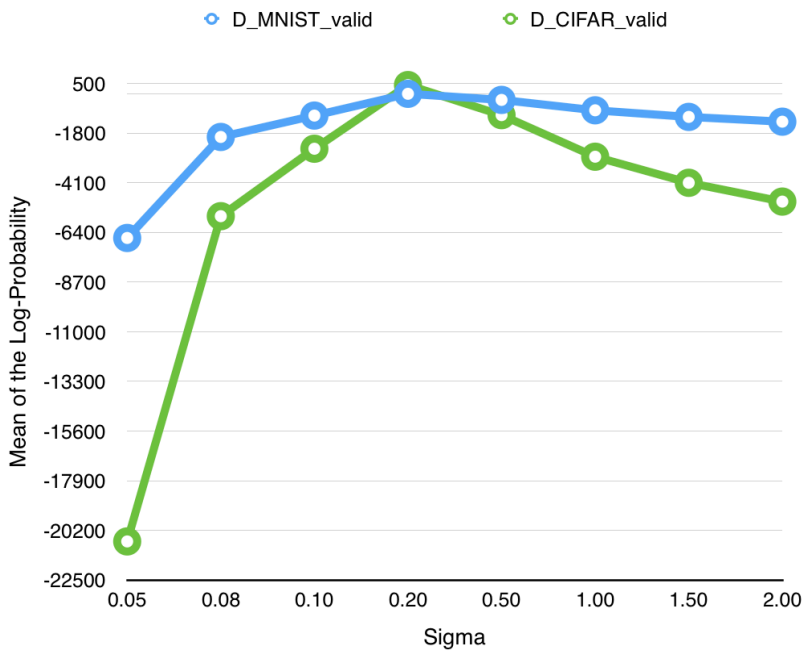
Computing the mean of the log-probability includes grid-search for the optimal value of σ no the validation dataset, and computing the mean of the log-probability on the testing dataset.

3. Results

The results of grid-search for the optimal value of σ on the validation dataset is shown in the following table:

sigma	D_MNIST_valid	D_CIFAR_valid
0.05	-6658.770	-20711.290
0.08	-1980.605	-5649.670
0.10	-990.937	-2526.212
0.20	18.457	423.048
0.50	-265.072	-969.758
1.00	-746.195	-2895.857
1.50	-1052.173	-4103.653
2.00	-1273.055	-4973.988

A trend can be observed while increasing σ. At a specific σ, the log-likelihood reaches at a global maxima, which means that the Gaussian distributions with this specific standard deviation



can best fit the data. The physical meaning of it is probably that the bandwidth of this standard deviation matches the size of features from the image patterns.

The mean of the log-probability and the running time on testing dataset is shown in the following table:

	D_MNIST_test	D_CIFAR_test
Mean of the Log-Probability	233.701	856.558
Running Time	421.584 s	5088.729 s

4. Insights

The pro of this method is that it doesn't require the number of clusters in prior to run.

The con of this method is that it is not scalable. When the data size is larger, the running time is dramatically longer. This algorithm is $O(n^2)$.