Krishna Chaitanya Naragam (1001836274)
Using One of my late days!

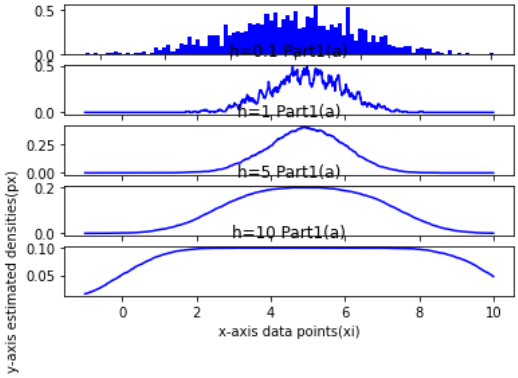
# Kernel Density Estimation

Kernel density estimation is a really useful statistical tool with an intimidating name. Often shortened to KDE, it's a technique that lets you create a smooth curve given a set of data.

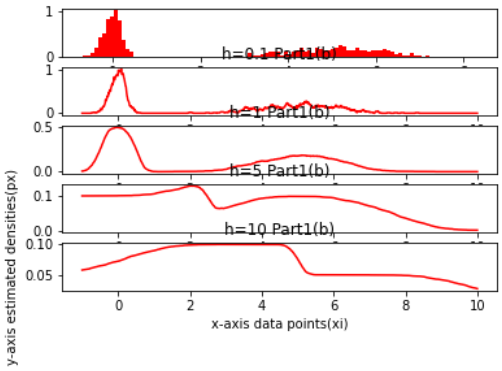
This can be useful if you want to visualize just the "shape" of some data, as a kind of continuous replacement for the discrete histogram. It can also be used to generate points that look like they came from a certain dataset - this behavior can power simple simulations, where simulated objects are modeled off of real data.

The given question asked us to calculate the following, given the input data, the following are the distributions found alongside the results for the parts of questions.

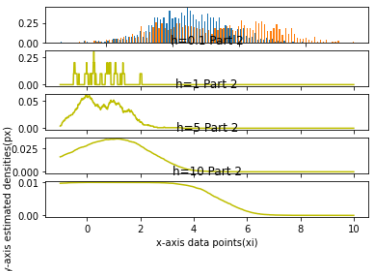
Q1 Part 1 a:




Q1 Part1 b:




Q1 Part2:

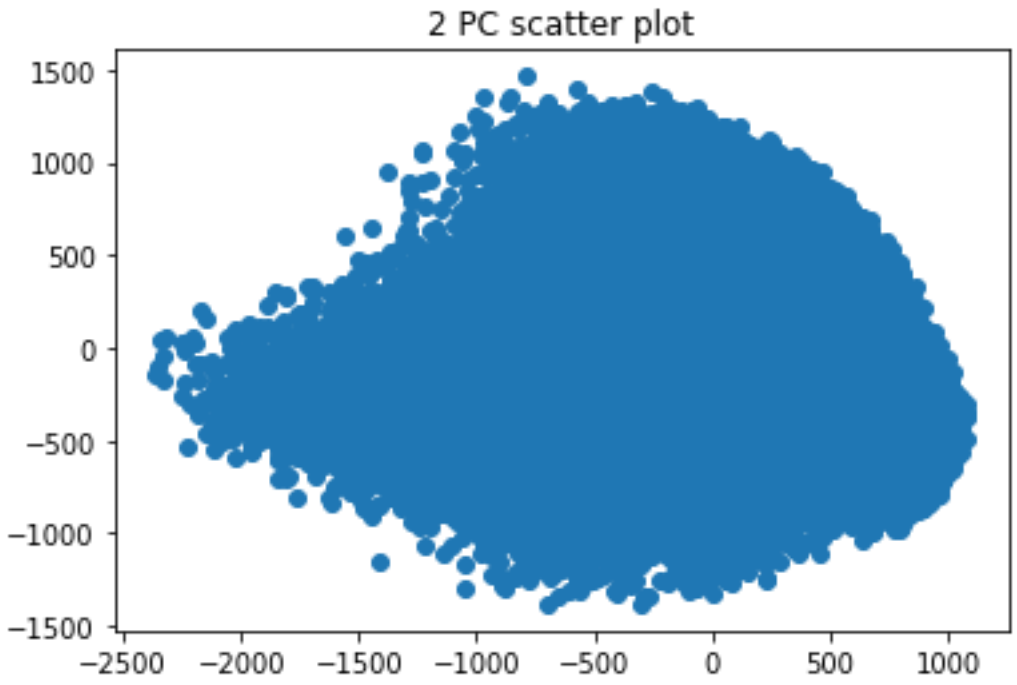


Krishna Chaitanya Naragam (1001836274)
Using One of my late days!

# Principal Component Analysis (PCA)

Large datasets are increasingly common and are often difficult to interpret. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. Finding such new variables, the principal components, reduces to solving an eigenvalue/eigenvector problem, and the new variables are defined by the dataset at hand, not a priori, hence making PCA an adaptive data analysis technique. It is adaptive in another sense too, since variants of the technique have been developed that are tailored to various different data types and structures. This article will begin by introducing the basic ideas of PCA, discussing what it can and cannot do. It will then describe some variants of PCA and their application.

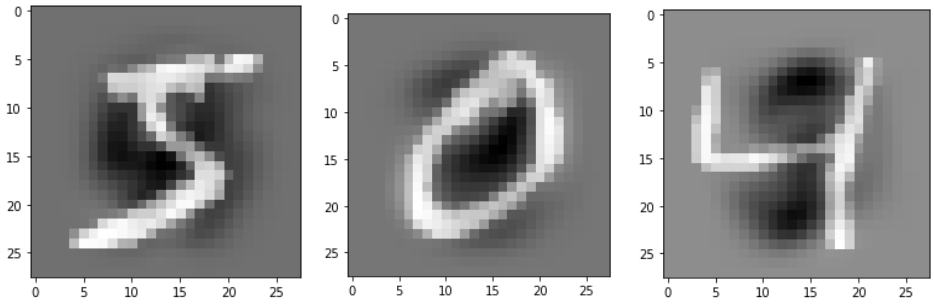
The problem 2 part 3,




Scatter plot of k = 2, values of the MNIST data.

They look scattered all over and overlapping, which is not the best way to have such scatter plots. The data needs more accuracy to have anything more, the data needs to be more dimensional to get the original value or to reach near the original value of 768, the k value of 2 is very low and not enough to do training on this kind of data.

The next part was to use k=10 and re draw the original data, that resulted something like this which is pretty much readable and considerable to train the classifier on.



The next part was to use k=30 and use a logistic regression model to train and predict the numbers.

This turned out to be a very good method to speedup the training process by reducing the dimensionality of the data and decreasing the training time. The decrease in accuracy was found to be around, 1% which is considerable as the data was reduced from 768 dimensions to a 30. With this kind of reduce in computation required to calculate the same data, the amount of accuracy lost is considerable.

The next part of the assignment was to print the weights that the network used while during the logistic regression model was predicting on the unmodified dataset. This turned to show the shape of numbers that the classifier was trying to predict. The weights nearly resembled the numbers or the shape of numbers that they were predicting. This shows that the model learnt good and the classifier would recognize handwritten digits with the accuracies given.

The features in part 4 were clearer images of shapes as the original numbers were reconstructed form the eigen images. In part 6, the weights were only printed, the weights roughly show all the images of the class combined, whereas the ones in the part 4, were reconstruction of some of the datapoints and that would more clearly show the number than the model learnt by itself. The model in part 4 was meant to reconstruct the original datapoint and few such points were chosen and reconstructed with the reduced dimensionality.

# References

Principal component analysis: a review and recent developments | Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences (royalsocietypublishing.org)

Kernel Density Estimation (mathisonian.github.io)

Variable Kernel Density Estimation on JSTOR

Machine Learning with Python: Training and Testing the Neural Network with MNIST data set (python-course.eu)

Logistic Regression in Python – Real Python