Introduction to Data Science - Assignment 5-6

Shahriyar Mahdi Robbani

April 2, 2020

Exercise 1

1. **Write explicitly the probability density function of ϵ**

$$f(\epsilon) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

2. **From it write the conditional probability distribution of y knowing a, x and b, i.e., the probability of observing y knowing a, x and b.**

$$p(y|f(x,a,b),\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(f(x,a,b)-y)^2}{2\sigma^2}\right)$$

3. **What are the parameters of the linear model?**
   The parameters of the model are a, b and σ.

4. **In a Bayesian treatment of the linear model, how many priors are needed, and over which parameters. Can you use Gaussian priors for all parameters?**
   Three priors are needed for the parameters a, b and σ. You can only use a Gaussian prior if the parameter is Gaussian, so it depends on the nature of a and b. Gaussian can not be used for σ since σ can only take positive values and Gaussian allows negative values.

Exercise 2

```
Weight for first feature:
Intercept:  [5.626]
Fixed acidity:  [0.08766808]
```

The weight of the feature is quite small, which indicates it is not a very strong predictor, but it still positively correlated with quality.

```
Weights for all features:
Intercept:  [5.626]
Fixed acidity:  [0.03409503]
Volatile acidity:  [-0.19185531]
Citric acid:  [0.00506797]
Residual sugar:  [0.06966651]
Chlorides:  [-0.13472303]
Free sulfur dioxide:  [0.05831775]
Total sulfur dioxide:  [-0.12812179]
Density:  [-0.0894315]
pH:  [-0.06780943]
Sulfates:  [0.15031245]
Alcohol:  [0.24954306]
```

Fixed acidity now has a smaller weight, as do most of the features indicating they are not useful for quality prediction. Alcohol and Sulfate content have the highest weights indicating they are the best features for quality prediction.

Exercise 3

```
RMSE for first feature: 0.7860892754162222
RMSE for all features: 0.6447172773067071
```

The RMSE is reduced after using all features so it agrees with the previous results that Fixed acidity alone is a bad predictor and alcohol and sulfates are better predictors.

Exercise 4

Normalization is useful for euclidean distance based statistical methods since it prevents a feature with a high magnitude from being weighed too highly. However, Random forests are not distance based and use absolute values to partition data. Therefore normalization will not affect the random forest result.

Exercise 5

```
Accuracy of Random Forest: 0.9668989547038328
```

Exercise 6

```
Leaning Rate: 0.1, Iteration: 10000, Value: nan
Leaning Rate: 0.01, Iteration: 116, Value: 0.9938268478110839
Leaning Rate: 0.001, Iteration: 1273, Value: 0.993826847811084
Leaning Rate: 0.0001, Iteration: 10000, Value: 0.993826847811084
```

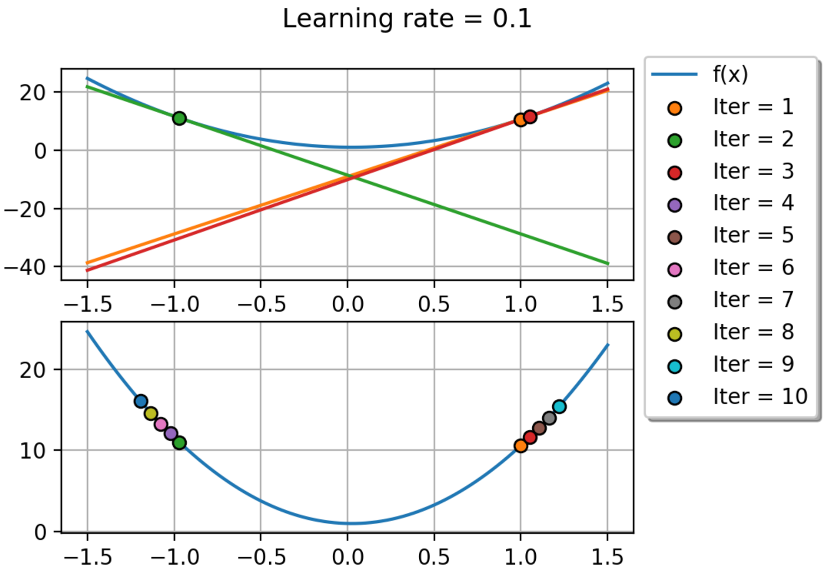


Figure 1: Gradient descent steps and tangent lines for first 3 iterations (top) and Gradient descent steps for first 10 iterations for a learning rate of 0.1

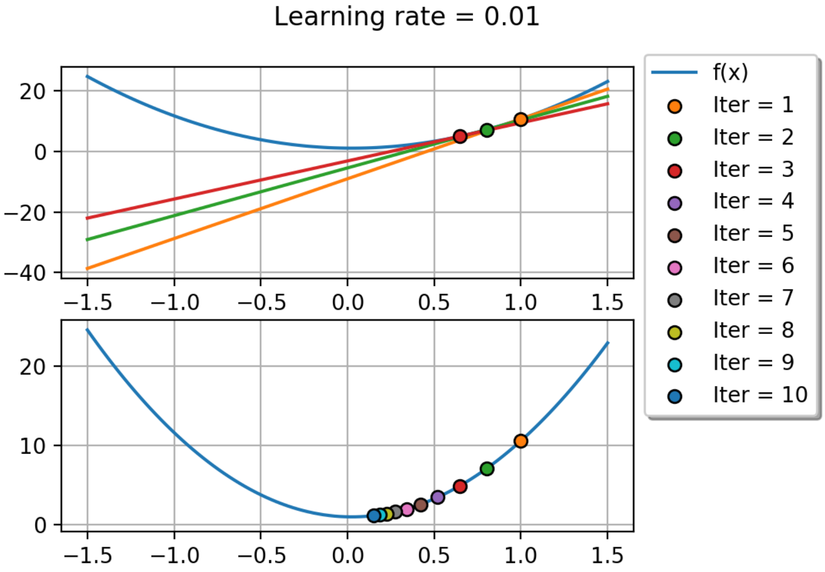


Figure 2: Gradient descent steps and tangent lines for first 3 iterations (top) and Gradient descent steps for first 10 iterations for a learning rate of 0.01

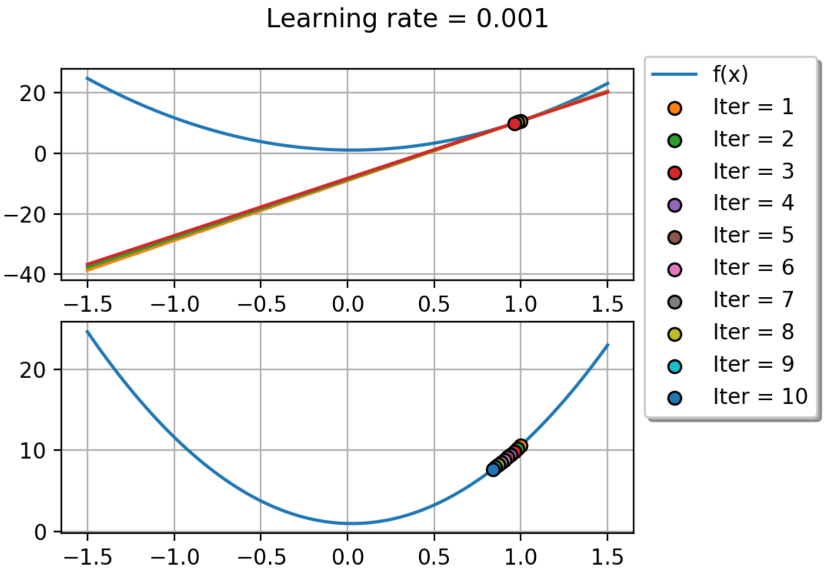


Figure 3: Gradient descent steps and tangent lines for first 3 iterations (top) and Gradient descent steps for first 10 iterations for a learning rate of 0.001

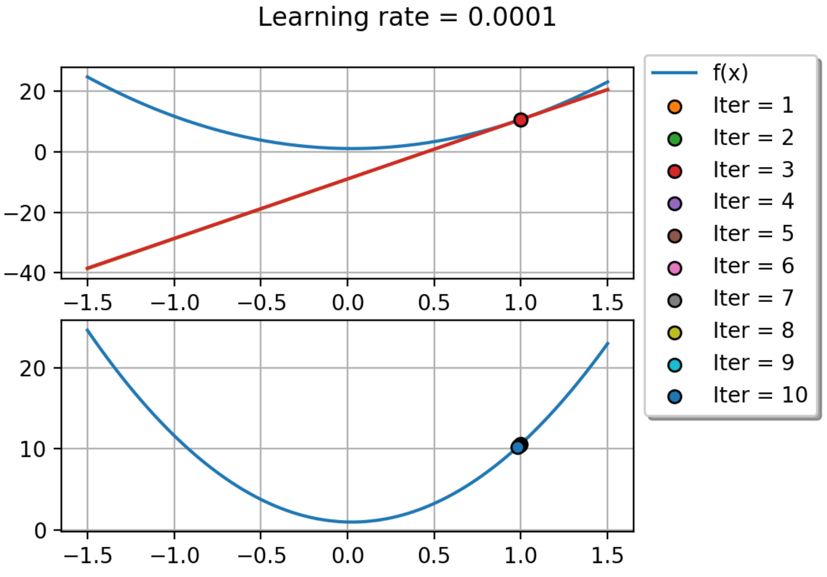


Figure 4: Gradient descent steps and tangent lines for first 3 iterations (top) and Gradient descent steps for first 10 iterations for a learning rate of 0.0001

Exercise 7

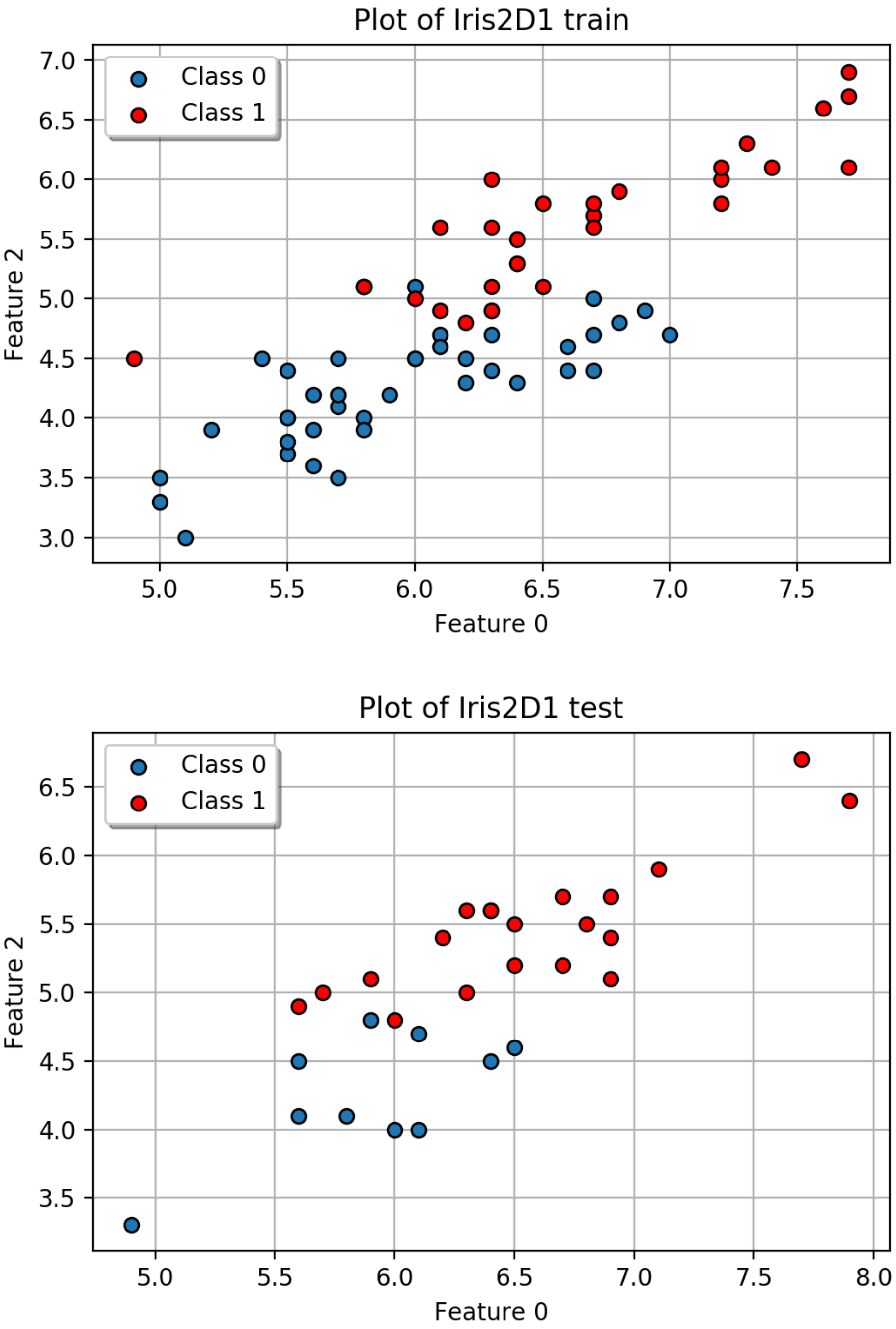


Figure 5: Plot of Iris2D1 training data (top) and testing data (bottom)

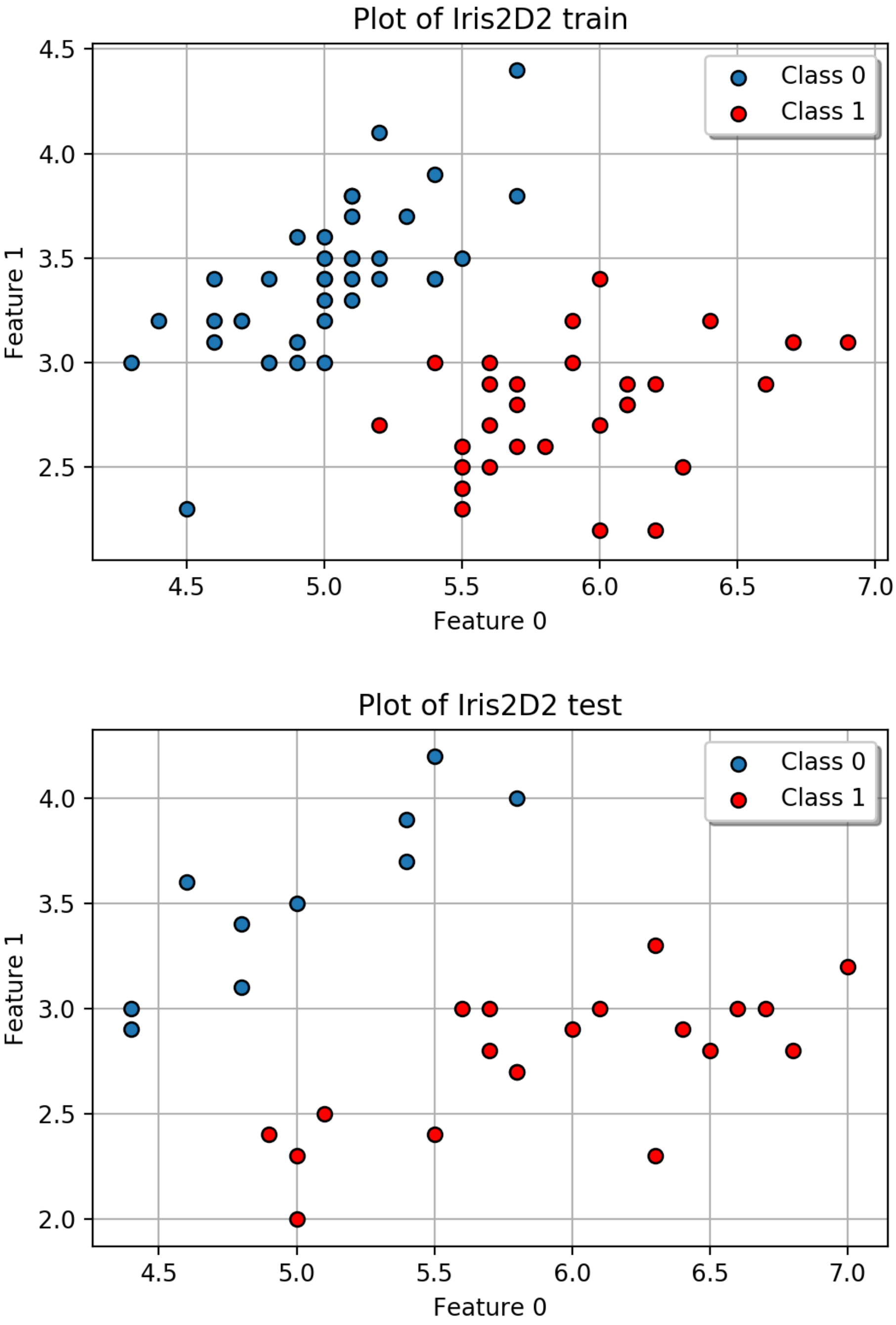


Figure 6: Plot of Iris2D2 training data (top) and testing data (bottom)

The classes are quite distinct and form separate clusters in the Iris2D2 dataset, whereas they are not so easily separable for the Iris2D1 dataset.

```
Iris2D1 Error: 0.06666666666666667
Iris2D1 Weights:
[[-13.04908134]
 [ -4.57254234]
 [  8.43444076]]
```

```
Iris2D2 Error: 0.0
Iris2D2 Weights:
[[-28.75235773]
 [ 12.61049236]
 [-12.64102124]]
```

Exercise 8

1. The insample error function is defined as:

$$E_{in} = \frac{1}{N}\sum_{n=1}^{N} ln\left(\frac{1}{\theta(y_n\mathbf{w}^T\mathbf{x}_n)}\right)$$

The logistic function θ is defined as:

$$\theta(s) = \frac{e^s}{1+e^s}$$

Therefore:

$$\frac{1}{\theta(s)} = \frac{1+e^s}{e^s}$$

Rearranging this gives us:

$$\frac{1}{\theta(s)} = \frac{1}{e^s} + \frac{e^s}{e^s}$$

$$\frac{1}{\theta(s)} = e^{-s} + 1$$

$$\frac{1}{\theta(s)} = 1 + e^{-s}$$

Therefore the insample error function is equivalent to:

$$E_{in} = \frac{1}{N}\sum_{n=1}^{N} ln(1 + e^{-y_n\mathbf{w}^T\mathbf{x}_n})$$

According to the chain rule, the derivative of E_{in} is:

$$\nabla_{\mathbf{w}}E_{in} = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{1+e^{-y_n\mathbf{w}^T\mathbf{x}_n}}(-y_n\mathbf{x}_n e^{-y_n\mathbf{w}^T\mathbf{x}_n})$$

Rearranging this gives us:

$$\nabla_{\mathbf{w}} E_{in} = \frac{1}{N} \sum_{n=1}^{N} -y_n \mathbf{x}_n \frac{e^{-y_n \mathbf{w}^T \mathbf{x}_n}}{1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}}$$

Since the logistic function θ is defined as:

$$\theta(s) = \frac{e^s}{1 + e^s}$$

We can therefore rewrite the derivative as:

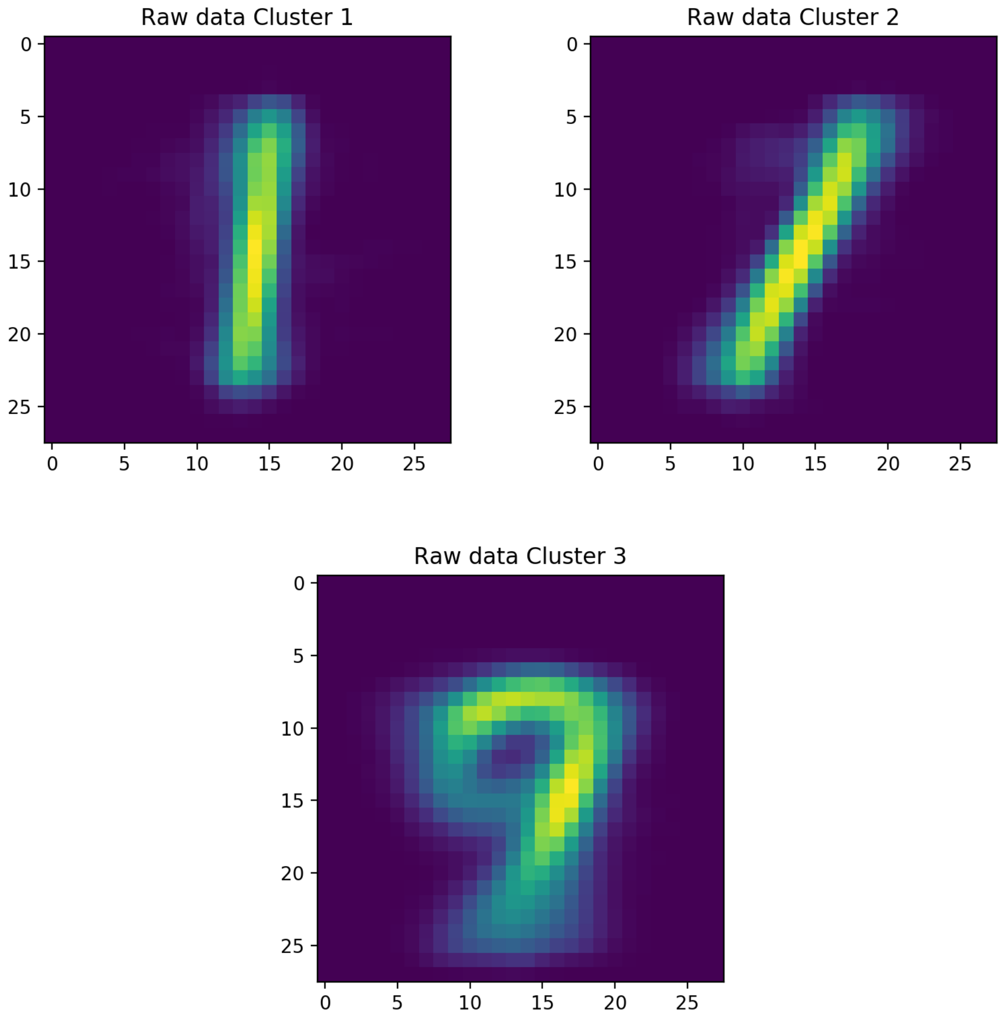
$$\nabla_{\mathbf{w}} E_{in} = \frac{1}{N} \sum_{n=1}^{N} -y_n \mathbf{x}_n \theta(-y_n \mathbf{w}^T \mathbf{x}_n)$$

2. **Argue that a misclassified' example contributes more to the gradient than a correctly classified one.**
   For a correctly classified example, the $-y_n \mathbf{w}^T \mathbf{x}_n$ term will always be negative and for an incorrectly classified example, the $-y_n \mathbf{w}^T \mathbf{x}_n$ term will always be positive. Since the logistic function converts all values to a range between 0 and 1, all negative inputs will have an output lower than 0.5 and all positive inputs will have an output greater than 0.5. Since a misclassified example will always be positive, $\theta(-y_n \mathbf{w}^T \mathbf{x}_n)$ will always be larger than for a correctly classified example, therefore it will contribute more to gradient.

Exercise 9

I used the KMeans function I created in the previous assignment to cluster the data. That function calculates a list of distances between each data point and the centroid for k centroids, and converts these lists into a matrix. It then finds which data point has the smallest distance to a centroid, and then assigns that data point to a cluster. This results in k clusters. The new centroids are then calculated by finding the mean position of each cluster. The loss is defined as the sum of squared distances between each point and the corresponding centroids and is used as a measure of how good the centroids are. The algorithm continues looping until the loss converges, indicating no change in the centroids.

```
Cluster: 1
Proportion of 1: 83.33%
Proportion of 7: 8.94%
Proportion of 9: 7.72%
Cluster: 2
Proportion of 1: 85.35%
Proportion of 7: 11.62%
Proportion of 9: 3.03%
Cluster: 3
Proportion of 1: 0.15%
Proportion of 7: 48.46%
Proportion of 9: 51.4%
```



Raw data Cluster 1

Raw data Cluster 2

Raw data Cluster 3

Figure 7: Plot of the cluster centers of the raw data as images

While the image of each mean is quite distinct, the different numbers are not clustered properly. Cluster 1 and 2 both mainly contain the number 1 while cluster 3 contains 7 and 9 in almost equal proportions.

```
Best k: 1
Accuracy: 0.9442222222222222
```

Exercise 10

The eigh function is used to calculate the eigenvectors and eigenvalues of the dataset, and then these values are sorted in descending order. The cumulative sum of eigenvalues was plotted against the number of principal components

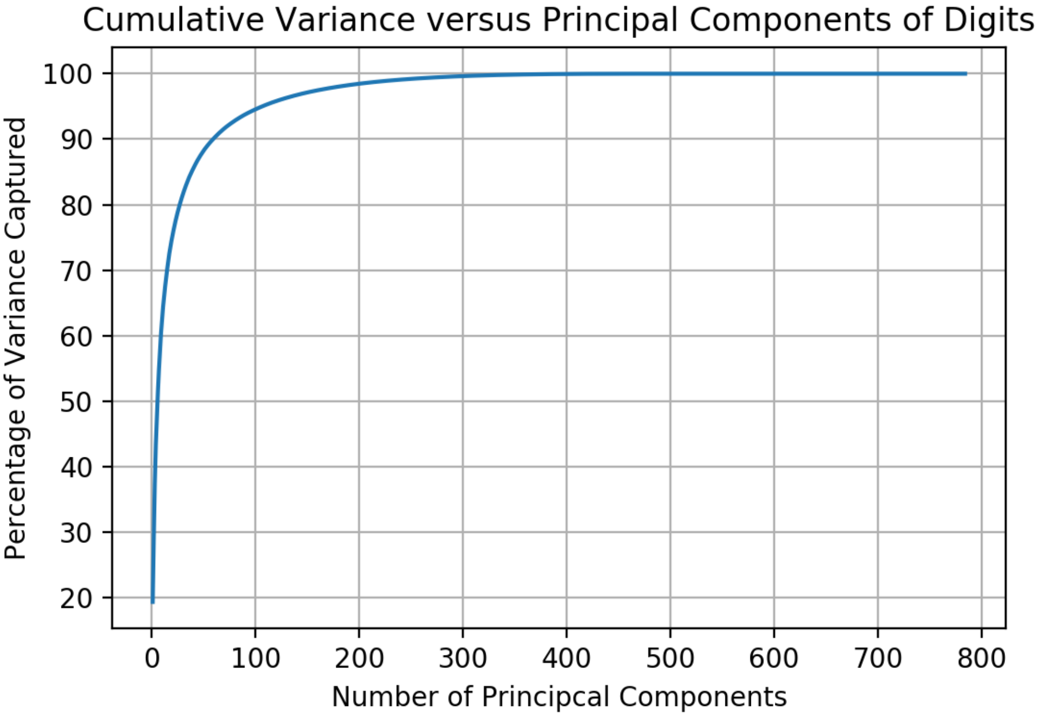


Figure 8: Percentage of Cumulative Variance with respect to principal components for the digits dataset

The first few principal components account for the majority of the variation. Over 90% of the variation is accounted for by less than 100 components and almost 100% is accounted for by 300 components.

```
===============================================================================
1 dimension:               20 dimensions:              200 dimensions:
===============================================================================
Cluster: 1
Proportion of 1: 5.48%     Proportion of 1: 83.33%     Proportion of 1: 83.33%
Proportion of 7: 46.97%    Proportion of 7: 8.94%      Proportion of 7: 8.94%
Proportion of 9: 47.55%    Proportion of 9: 7.72%      Proportion of 9: 7.72%
Cluster: 2
Proportion of 1: 95.7%     Proportion of 1: 86.22%     Proportion of 1: 85.35%
Proportion of 7: 3.49%     Proportion of 7: 11.22%     Proportion of 7: 11.62%
Proportion of 9: 0.81%     Proportion of 9: 2.55%      Proportion of 9: 3.03%
Cluster: 3
Proportion of 1: 0.0%      Proportion of 1: 0.15%      Proportion of 1: 0.15%
Proportion of 7: 49.01%    Proportion of 7: 48.46%     Proportion of 7: 48.46%
Proportion of 9: 50.99%    Proportion of 9: 51.39%     Proportion of 9: 51.4%
```

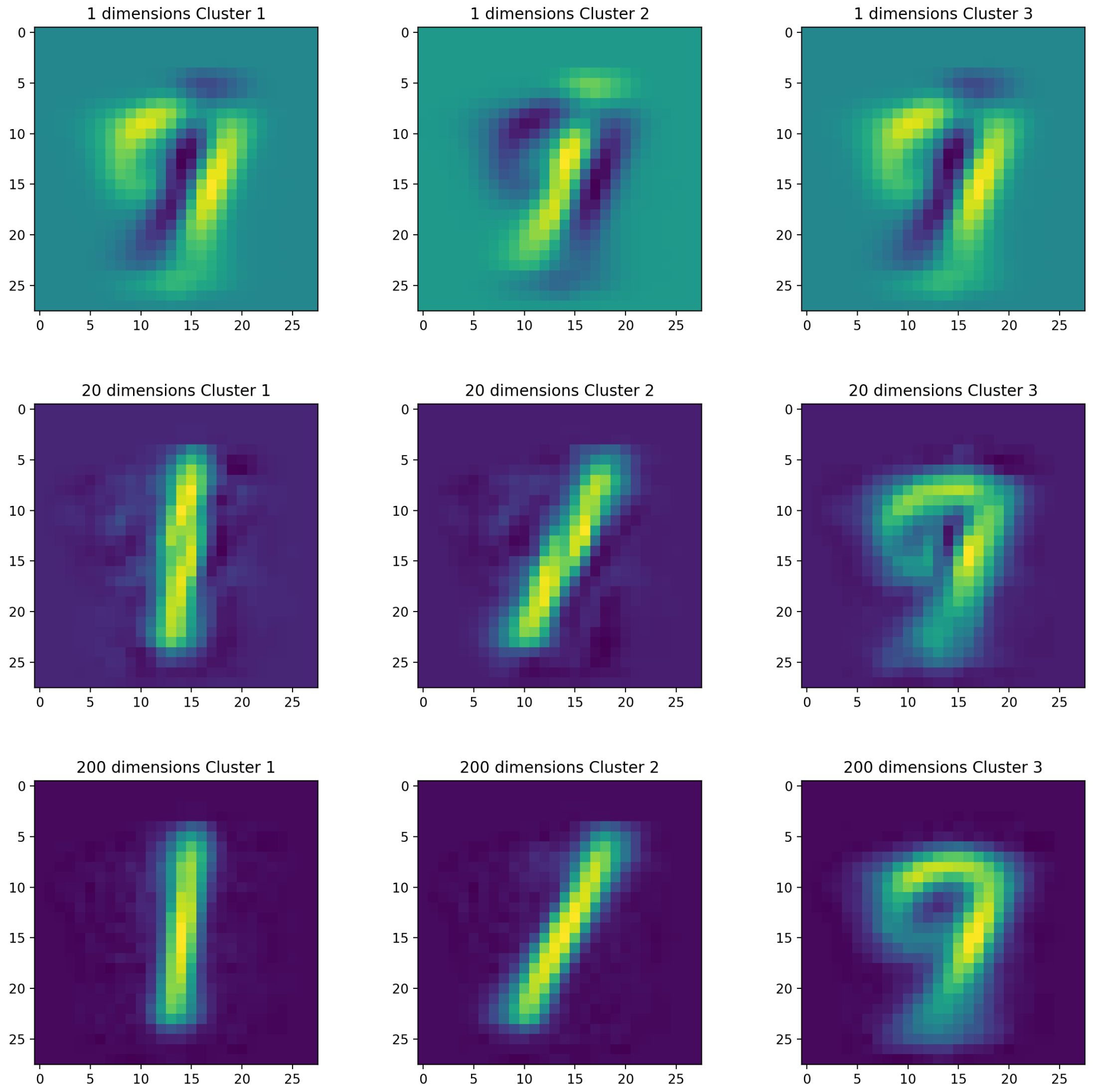


Figure 9: Plot of the cluster centers as images for the digits dataset after reduction to 1 dimensions (top), 20 dimensions (middle) and 200 dimensions(bottom)

The means produced by reduction to 1 dimension is very poor the three numbers cannot be distinguished. The means produced by 20 dimensions are better and the ones produced by 200 dimensions is very similar to the means produced by the whole raw dataset.

```
===============
20 dimensions:
===============
Best k: 1
Accuracy: 0.9502222222222222
===============
200 dimensions:
===============
Best k: 1
Accuracy: 0.9442222222222222
```

The best k was 1 regardless of using the raw data or the data with reduced dimensions. Reducing the data to 20 dimensions improved accuracy slightly while reducing to 200 dimensions kept the accuracy the same.