**Exercise 2**

Ilana Pervoi 318271640, Omri Bar Oz 313325961

**Question 1:**

* 1. code
  2. code

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster number** | **Size** | **Common Label** | **Percentage** |
| 0 | 101 | 7 | 0.42 |
| 1 | 128 | 7 | 0.35 |
| 2 | 81 | 0 | 0.99 |
| 3 | 81 | 3 | 0.62 |
| 4 | 63 | 3 | 0.41 |
| 5 | 147 | 1 | 0.65 |
| 6 | 101 | 2 | 0.60 |
| 7 | 75 | 4 | 0.67 |
| 8 | 100 | 6 | 0.52 |
| 9 | 115 | 8 | 0.54 |

The algorithm classified correctly 568 out of 1000 samples classification error of 43% on the sample. We calculated it by comparing the true label of a sample to the common label of the cluster of the sample. We knew the true label of a sample by its index in the sample matrix, as we created the data by generating 100 samples from each digit.



|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster number** | **Size** | **Common Label** | **Percentage** |
| 0 | 291 | 1 | 0.1 |
| 1 | 1 | 0 | 1 |
| 2 | 1 | 2 | 1 |
| 3 | 1 | 0 | 1 |
| 4 | 1 | 4 | 1 |
| 5 | 1 | 5 | 1 |
| 6 | 1 | 5 | 1 |
| 7 | 1 | 5 | 1 |
| 8 | 1 | 6 | 1 |
| 9 | 1 | 6 | 1 |

The algorithm classified correctly 39 out of 300 samples classification error of 87% on the sample.

The k-means clustering algorithm worked better for this problem than single linkage.

**K-means**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster number** | **Size** | **Common Label** | **Percentage** |
| 0 | 192 | 3 | 0.39 |
| 1 | 182 | 7 | 0.38 |
| 2 | 95 | 0 | 0.80 |
| 3 | 148 | 6 | 0.36 |
| 4 | 211 | 1 | 0.47 |
| 5 | 172 | 2 | 0.19 |

The algorithm classified correctly 406 out of 1000 samples classification error of 59% on the sample.

TODO explain difference k=10 k=6.

**Single Linkage**:

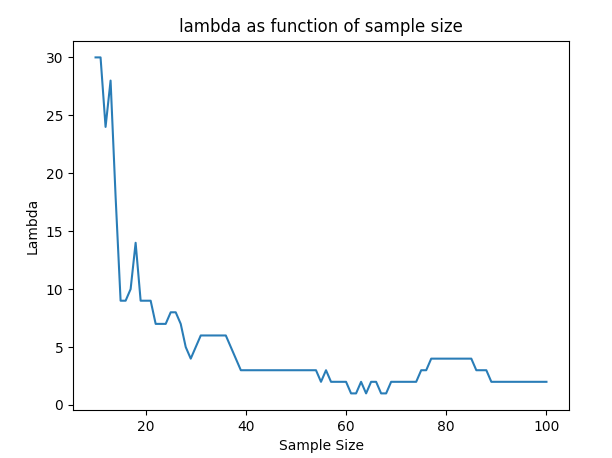
|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster number** | **Size** | **Common Label** | **Percentage** |
| 0 | 295 | 0 | 0.1 |
| 1 | 1 | 2 | 1 |
| 2 | 1 | 2 | 1 |
| 3 | 1 | 2 | 1 |
| 4 | 1 | 3 | 1 |
| 5 | 1 | 4 | 1 |

The algorithm classified correctly 35 out of 300 samples classification error of 88% on the sample.

When moving from k=10 to k=6 we don’t see a major difference in the results of the algorithm – most of the samples are assigned to one cluster, which results is about 87% error on the examples.

The reason behind it is that the distance between the closest samples of some two different (real) labeled samples, is closer than the distance between some two samples on the “edges” of the different real labels. Hence, the “real” clusters will be merged to one cluster in the algorithm, resulting in 1 big cluster and k-1 clusters with minimal number of assigned samples (as we generate the initial clusters with examples from the sample set).

**Question 2:**

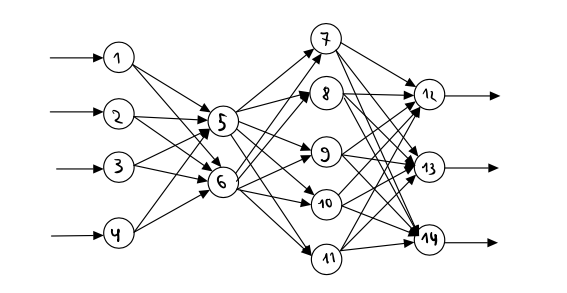


1. We expect to see the value of optimal decreases as the sample size increases. This is because when the sample size is low, the sample does not represent the distribution properly, hence we will obtain large hypothesis class which will result in overfitting. To handle that, a higher will be required as a penalty to reduce the hypothesis class size.   
   As the sample size increases, we expect to see a decrease in the optimal value of , until convergence. As the sample size increases, the hypothesis class size decreases. This results in less overfitting which means we don’t want to penalize the norm of w as much.
2. Yes, this is we got what we expected as explained 2.b, in the plot submitted in 2.a.
3. TODO

**Question 3:**

**Question 4:**

1. The graph that describes the neural network architecture:



**Question 5:**

1. Let and . Let be the hypothesis class consisting of decision trees with depth at most n and binary attribute tests of the form for .

For each tree in it has at most nodes. This is because the longest path is n. Then the tree with the largest number of nodes is a perfect binary tree with height n. Then the number of nodes is: .  
For each node in the tree, we can select an attribute , and then choose one of the possible values of to check if it is larger then. Then there are such options. A node can also be a leaf with label or . Then every node has options. Then:

1. Danny is trying to use PAC boundaries equations we learned in class. The problem in this case is that ID3 is not an ERM algorithm.  
   This means that Danny is wrong using this equation.

**Question 6:**

1. No. We will show a contradiction to the Naïve-Bayes assumption:
2. TODO

**Question 7:**

1. Since we want to reduce the dimensionality from 4 to 2, the distortion would be the sum of the lowest 2 eigen-values of .  
   Notice that in the experiment the 3rd column is linearly dependent of the 1st and 2nd columns, and the 4th column linearly dependent of the 2nd and 3rd columns linearly dependent of the 1st and 2nd columns. Hence, the rank of would be 2 and would have 2 eigen-values that are 0, which are the 2 lowest, as is SPD matrix.

In conclusion, the distortion is 0.

1. TODO

**Question 8:**

Let . Let such that . Define Trinomial distribution for as follows: . Assume that we have a sample .

Also assume that .

Denote . Then we have .  
We also have then which means that . Then: .

Also, denote the amount of in S with to be . More formally:

Now let’s continue:

we changed the likelihood function to be .

the maximum value of for this problem:

and because :

Hence the estimator is:

1. TODO