* 1. code
  2. code

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster number | Size | Common Label | Percentage |
| 0 | 74 | 3 | 0.53 |
| 1 | 130 | 7 | 0.45 |
| 2 | 103 | 9 | 0.29 |
| 3 | 79 | 2 | 0.67 |
| 4 | 117 | 6 | 0.29 |
| 5 | 108 | 1 | 0.46 |
| 6 | 100 | 3 | 0.4 |
| 7 | 77 | 4 | 0.49 |
| 8 | 138 | 1 | 0.36 |
| 9 | 74 | 0 | 0.93 |

correct 460 out of 1000. error of 0.54



|  |  |  |  |
| --- | --- | --- | --- |
| 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 |

correct 39 out of 1000. error of 0.87  
 The k-means clustering algorithm worked better for this problem.

* 1. K-means:

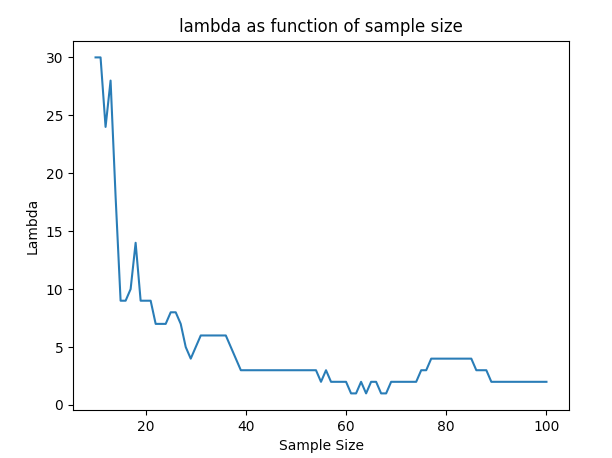
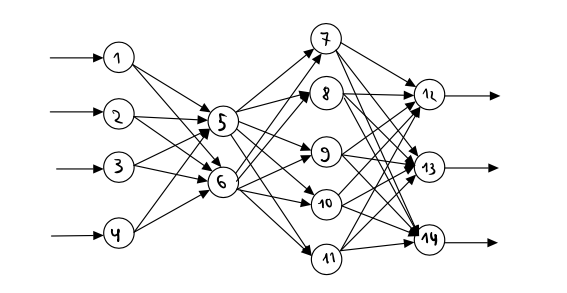
|  |  |  |  |
| --- | --- | --- | --- |
| Cluster number | Size | Common Label | Percentage |
| 0 | 228 | 3 | 0.38 |
| 1 | 96 | 0 | 0.89 |
| 2 | 223 | 1 | 0.43 |
| 3 | 114 | 6 | 0.7 |
| 4 | 270 | 4 | 0.32 |
| 5 | 69 | 2 | 0.87 |

correct 493 out of 1000. error of 0.51

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 |

correct 35 out of 1000. error of 0.88

* 1. 
  2. 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.
  3. Yes, this is we got what we expected as explained 2.b, in the plot submitted in 2.a.
  4. …
  5. …
  6. …
  7. The graph that describes the neural network architecture:  
     

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 .  
   1. For each tree in it has at most nodes. This is because the longest path is n. Then the tree with the largest amount of nodes is a perfect binary tree with height n. Then the amount of nodes is: .  
      For each node in the tree, we can select an attribute , and then choose one of the possible value 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:
   2. 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.
   3. No. We will show a contradiction to the Naïve-Bayes assumption:
   4. …
   5. …
   6. …
2. Let . Let such that . Define Trinomial distribution for as follows: . Assume that we have a sample .
   1. Let . 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:  
      Now the likelihood function is . Now let’s find the maximum value of for this problem:

and because :  
Then the estimator is:

* 1. …