CISC/CMPE452/CISC874/COGS 400
ASSIGNMENT 3

IMPLEMENT A HYBRID COMPETITIVE LEARNING NETWORK

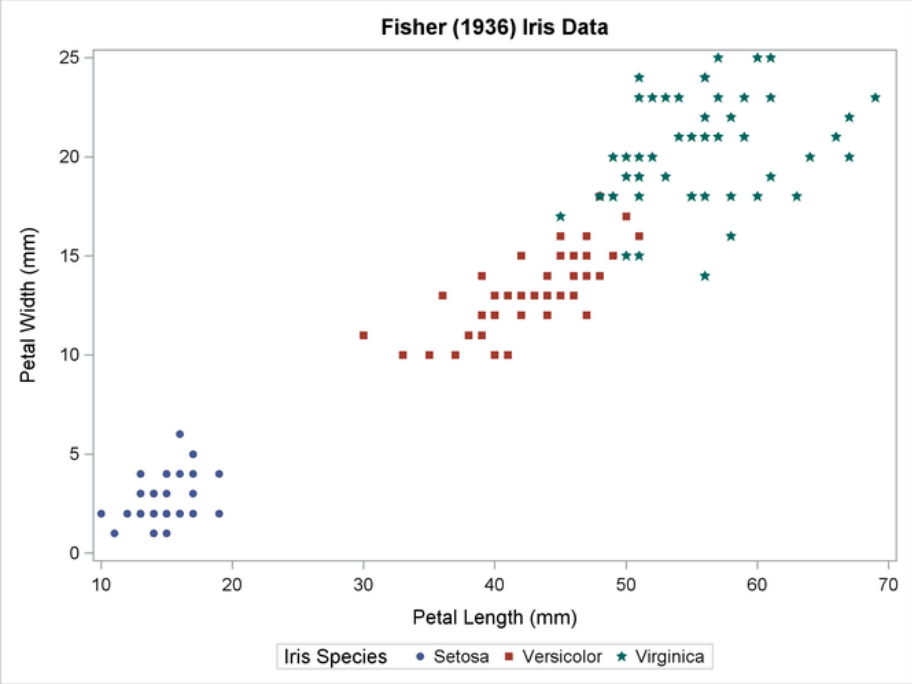
Farhana Zulkernine

10 marks

Assignment 3: The Data

- Iris data is one of the most popular datasets in machine learning tasks. Iris is a flowering plant with 260-300 different species.
- Iris data contains the measurements of 3 different species of Iris: Setosa, Versicolour and Virginia.
- The plot on the next slide shows the relation between Petal width and length.
- The data for this assignment is provided in two files: iris_train.txt and iris_test.txt.
  - *Each line in the files represents a separate data point where each data point includes the following comma separated values:*
    - Sepal length (cm), Sepal width (cm), Petal length (cm), Petal width (cm), class label
    - Example: 5.1,3.5,1.4,0.2,Iris-setosa
  - *Iris_train.txt contains 40 data points for each class. Use this to train your perceptron.*
  - *Iris_test.txt contains 10 data points for each class. Use this to test that your network puts the data point in the right cluster.*

Assignment 3 - Plot of Iris data (Petal width and length)



Assignment 3 : Tasks

1. Design a simple competitive learning clustering ANN model and train the model using the iris data set and the semi-supervised LVQ algorithm as discussed in the class.
2. Train the model weights using train.txt to find three clusters in the dataset such that the clustering error is minimum. Use appropriate terminating conditions. Note down the network parameters, structure, clustering errors for each cluster and codebook vectors. Let's refer to this as "LVQ 1".
3. Test how well the network performs by testing LVQ 1 with test.txt data to see if the data points are put in the right cluster. Calculate the accuracy and create a confusion matrix for the test data.

Assignment 3 : Tasks (cont…)

1. Next implement a PCA network (ANN approach) to reduce the number of parameters in the iris dataset from 4 to 3. PCA should be applied on the input parameters only and should not include the labels. Note down the network weights and structure to add to the report file.
2. Now direct the output from the PCA to a LVQ (Let's refer to this integrated network with fewer inputs as "LVQ 2") and train this integrated LVQ 2 network with train.txt to find 3 clusters. Note down cluster errors and codebook vectors.
3. Validate the network using test.txt to see if the data points are placed in the right cluster or Voronoi region. Compute the confusion matrix for the test data.

   Note that a random presentation of the data to the ANN models could be generally beneficial.

Mark Distribution – 10 Marks

- 5 marks - for fully functional executable code with comments in the code for both LVQ 1 and integrated LVQ 2 (PCA + LVQ).
- 5 marks - for the report text file containing
  - *Output text file (1 mark) - Codebook vectors and cluster errors for LVQ1 and LVQ2.*
  - *ANN structure (3 marks) – Number of nodes and layers, parameter values such as node output function, the initial and final weight vectors, learning rate, training iterations and terminating condition, and other value choices made and why you made the above choices for PCA, LVQ1, LVQ2. Compute confusion matrix for test data for LVQ1 and LVQ2.*
  - *Results and discussion (1 mark) – Overall accuracy in terms of cluster error with and without the PCA network, discuss and explain why results are good/bad and what were the challenges.*

General Instructions for Code and Submission (for all assignments)

- You can use any programming language (preferred Python, C, C++, Java – No GUI tool or built-in libraries should be used to build/train/test the ANN)
- Make one zip file named as Asgn3_studentID which should include
  - *A report text file containing the ANN design and network parameters for the PCA, LVQ 1 and LVQ 2 networks. Also report the total error for each cluster and the codebook vectors for LVQ 1 and LVQ 2.*
  - *Program code with comments in the code to explain what each of your program files and functions are for LVQ 1, PCA and LVQ 2.*
- Upload zip file to the OnQ site. One mark will be deducted for each day after the due date until the end date after which no assignment submission will be accepted.
- Submit the following (see the mark distribution)
  - *Program code with comments, any associated libraries needed to execute your code (and/or executables and instructions for execution).*
  - *A report text/WORD file describing the structure of the ANN (diagram or description) and parameter values for the PCA, LVQ 1 and LVQ 2 networks. Report errors for each cluster and the corresponding codebook vectors both for LVQ 1 and LVQ 2. Submit confusion matrix for test data for LVQ1 and LVQ2.*