CSC 4740/6740 Data Mining

Assignment 5

Due Date: 11:59pm, 11/20/2022

Only the electronic version will be accepted. Please submit it through iCollege.

David obtained two datasets. He wants to practice the Neural Network and Support Vector machines classifiers on these two datasets. Please note that the two datasets should not be mixed. That is, you need to train a model using the dataset 1 training dataset and test the model using the dataset 1 testing dataset, and train another model using the dataset 2 training dataset and test the model using the dataset 2 testing dataset.

Dataset 1 is linearly separable. The raw dataset is in the file "dataset1_training.txt" and "dataset1_testing.txt". These two datasets are visualized as follows. Each row in the dataset represents one data object. Each dataset contains 3 columns: the first column represents x axis coordinates, the second column represents the y axis coordinates, the third column represents the class label of the objects.

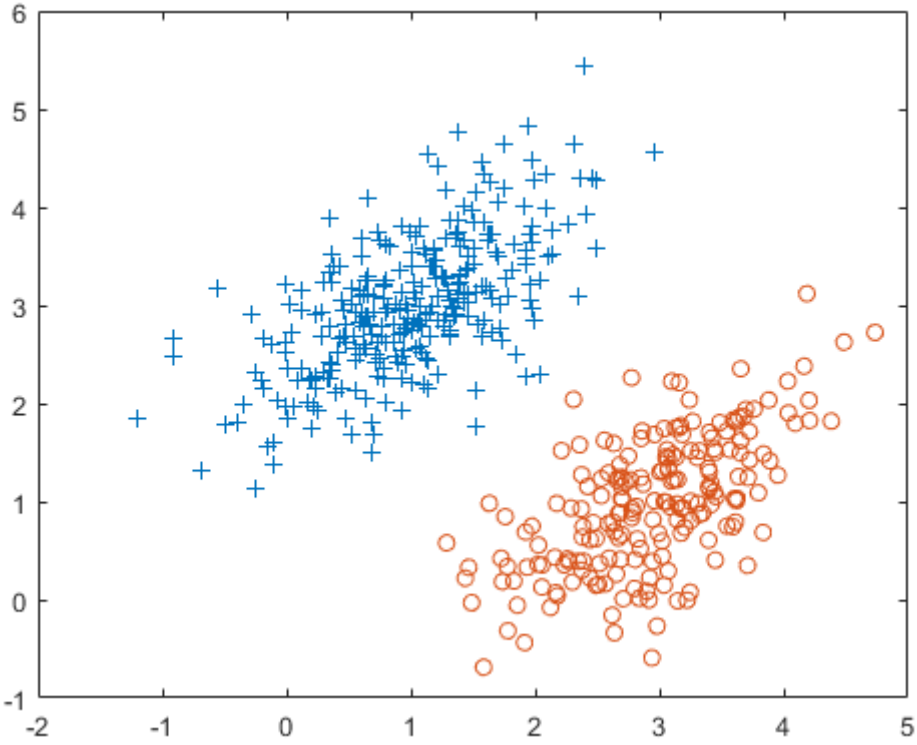


Figure 1. The training dataset of dataset 1.

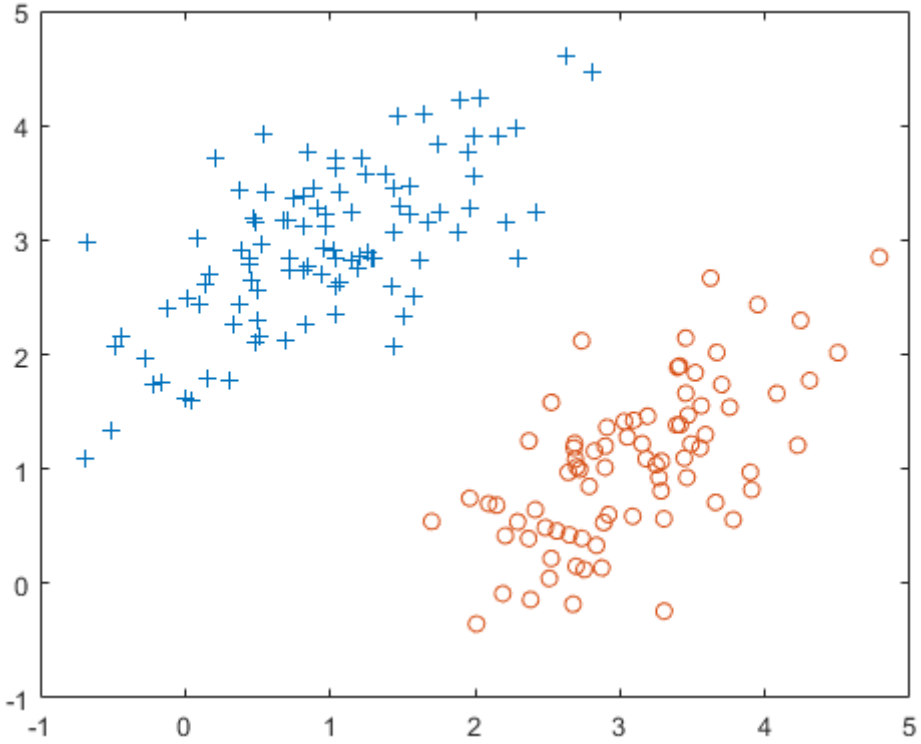


Figure 2. The testing dataset of dataset 1.

Dataset 2 is not linearly separable. The raw dataset is in the file “dataset2_training.txt” and “dataset2_testing.txt”. These two datasets are visualized as follows. Each row in the dataset represents one data object. Each dataset contains 3 columns: the first column represents x axis coordinates, the second column represents the y axis coordinates, the third column represents the class label of the objects.

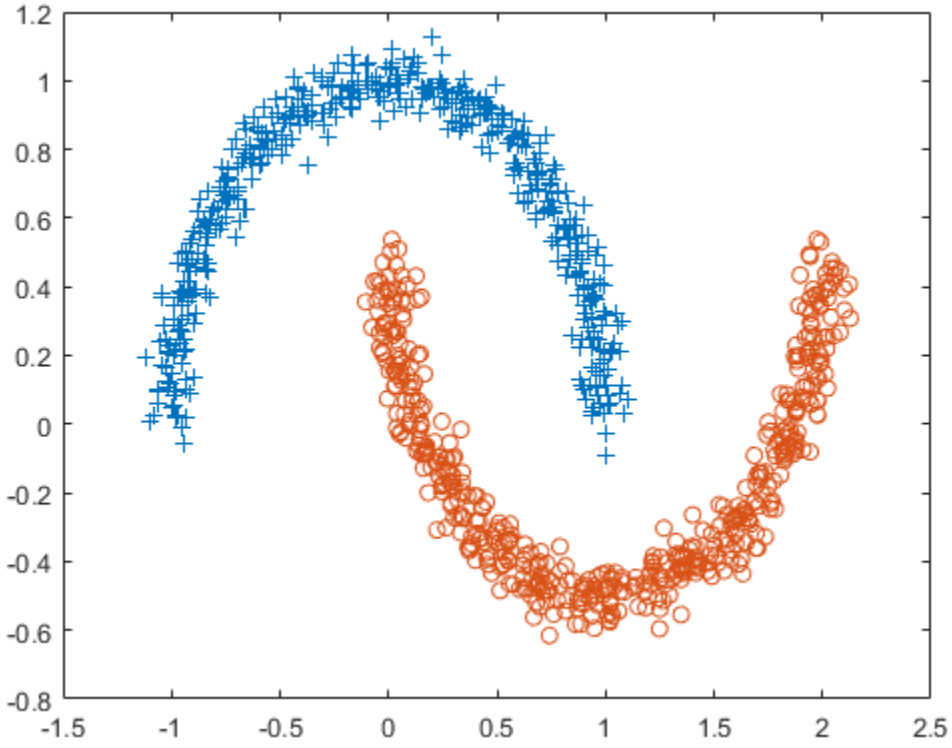


Figure 3. The training dataset of dataset 2.

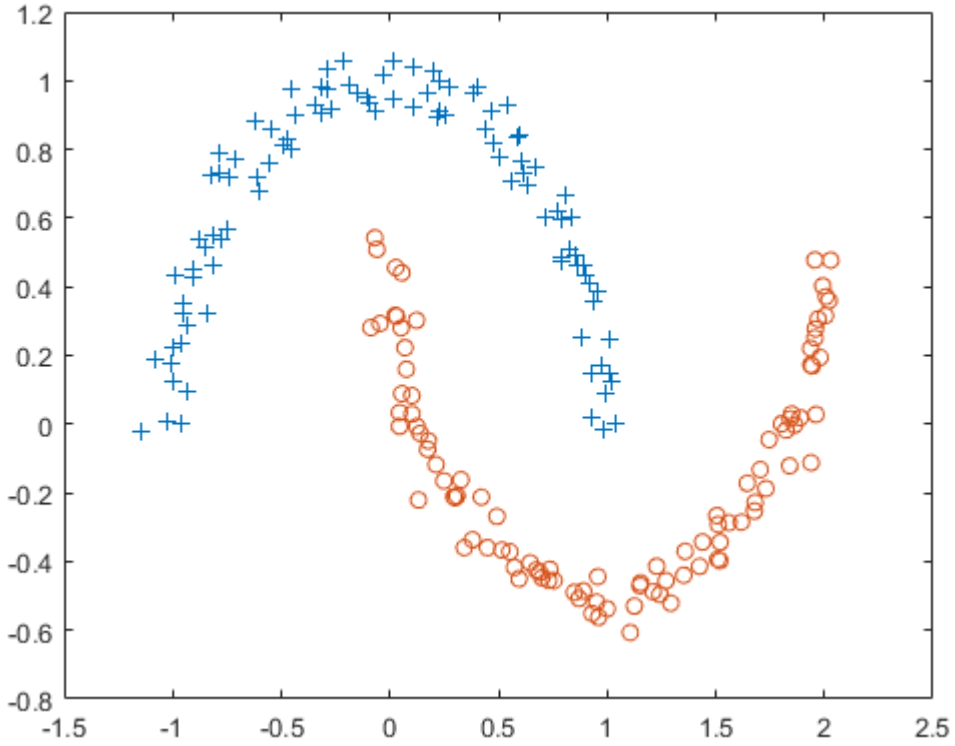
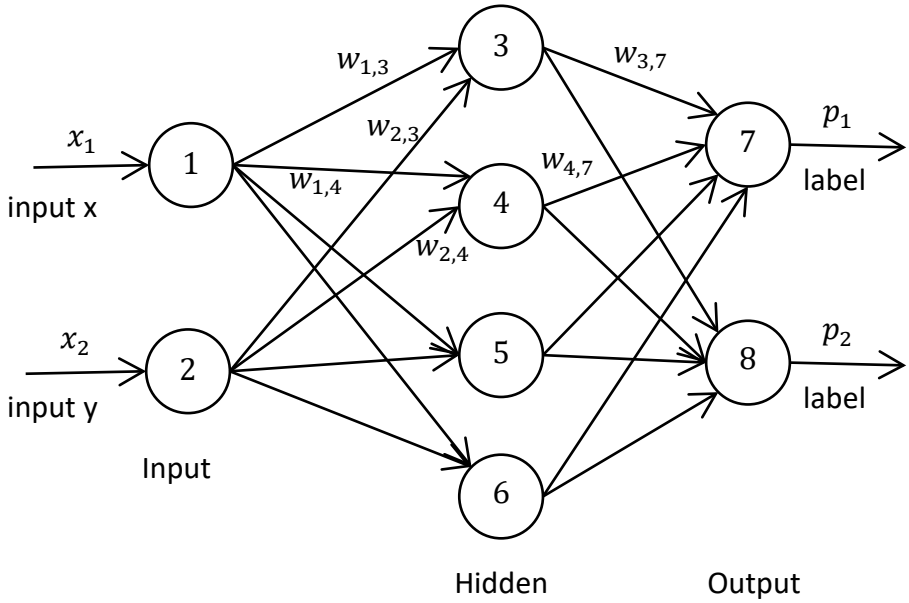


Figure 4. The testing dataset of dataset 2.

Problem 1 (50 points). Suppose Ben helps David design the following architecture of the neural network. It contains 1 hidden layer with 4 neurons.



Please implement the backpropagation algorithm and use the training datasets in dataset 1 and 2 to train the neural network, and then use the testing datasets in datasets 1 and 2 respectively to test the accuracy of the neural network you trained.

Please use the sigmoid function in the neurons in the hidden and output layers. Suppose the learning rate is set to 0.8. Suppose the parameter β in the logistic function is set to 1.

To visualize the boundary of the NN, we can try all points (x, y) for x in min:0.1:max and y in min:0.1:max for example, and the predict the label of each point in the grid. This will allow us to find the regions of the two classes partitioned by the NN trained by the datasets. Please provide the visualization results after you obtained the trained models from datasets 1 and 2.

Reference: How to Code a Neural Network with Backpropagation In Python (from scratch): https://machinelearningmastery.com/implement-backpropagation-algorithm-scratch-python/

Solutions:

Problem 2 (50 points). David plans to test the SVM on the two datasets. Please use SCIKIT-Learn SVM library to train the SVM model using dataset 1 and 2. And then test their accuracy using the testing datasets respectively.

Python SVM:

https://scikit-learn.org/stable/modules/svm.html

https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

Visualizations of the classification results and the boundary is highly appreciated by the TAs. The TAs will give some points for the visualization results.

Solutions: