Engineering Applications of Machine Learning and Data Analytics
Homework #5 [Due: 04/30/2021]

I acknowledge that this exam is solely my effort. I have done this work by myself. I have not consulted with others about this exam in any way. I have not received outside aid (outside of my own brain) on this exam. I understand that violation of these rules contradicts the class policy on academic integrity.

Name: ______________________________

Signature: ______________________________

Date: ______________________________

Instructions: There are four problems. X Partial credit is given for answers that are partially correct. No credit is given for answers that are wrong or illegible. Write neatly.

You must submit two PDFs on D2L. The first PDF has the results to the analytical questions as well as figures that are generated

Problem 1: ___________ Problem 2: ___________

Problem 3: ___________ :

Total: ___________

1 Semi-Supervised Learning [50pts]

This homework requires that you implement the self-training algorithm discussed in the semi-supervised learning lectures. You should refer back to your notes for the self-training pseudo-code.

[20pts] An Experiment on Synthetic Data

Generate 2D Gaussian data that is similar to the data set shown in Figure 1. Using data sets that have 1000 samples (500 from each class) implement the self-training algorithm, and test on a data set of 1000 samples. The requirements for this problem are as follows:

- Your implementation of self-training must use a classifier (e.g., neural network) that can give probabilities to select the data points that will have pseudo labels assigned to them. I will not make a restriction on the classifier other than the probabilities requirement. You will need to choose a suitable threshold to determine the data samples that will be labeled for the next round of self-training.
- Report the error of the self-training algorithm on the testing data at: (1) the first time a classifier is trained using only the labeled; (2) *at least one time* point during the self training process (i.e., when pseudo labels are used); and (3) after self-training is completed. Comment on the results.
- Perform an experiment reporting the above requirements with 10% and when 25% of the training data are labeled.

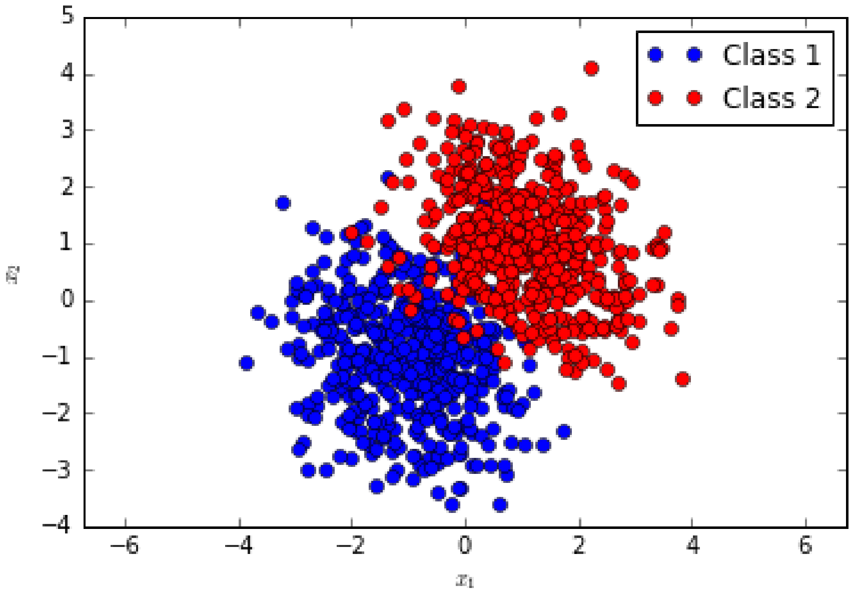

Figure 1: Example data set for problem 1.

[30pts] An Experiment on Real World Data

Implement the self-training algorithm using ten datasets available on the course Github repo. The requirements for this problem are as follows:

- You must report your results using 5-fold cross validation. In each cross validation step you can only use 15% of data as labeled. For example, if I have 500 data samples for training I must *randomly* select 15 data samples that I can use with my supervised classifier and the the 85 samples can be used in self-training.
- Write a brief discussion on wether semi-supervised helped on real-world data sets.