CISC Introduction to Machine Learning: Homework 3

Due: See Canvas

There are two parts: individual problems and group problems. Each student should upload one submission for individual problem. Each group should upload one submission for group problems.

Individual Problem (15pt+5pt)

Download Weka, a useful software.

Problem (15pt) We will practice Bagging and AdaboostM1. Details of AdaboostM1 are here. Experimental Settings:

- Collect one dataset from UCI machine learning repository or other places. Select an appropriate dataset. Not all datasets are viable for the settings in this experiment. Do not use very large datasets. Weka supports .arff file. You need convert your dataset to this format.
- Look at the **bagging** and **adaboostM1** function on Weka under the *meta* category. Select one algorithm supported by Weka.
- Please
  - run the algorithm itself
  - run bagging with numIterations=50
  - run adaboostM1 with weightThreshold=100000 and numIterations=50
  - note the error rates returned by Weka for each run.

What to Turn in

- your dataset in .arff format.
- (15pt) A report showing the algorithm you select and the error rates you got. Explain why bagging and boosting are useful or not useful depending on your observations.
- Extra credits. (5pt) If bagging helped in improving the performance in your experiment, can you find another algorithm where bagging will not work? If bagging did not help in improving the performance in your experiment, can you find another algorithm where bagging will help? Please show and discuss your findings, as the last part of your report.

Group Problem (35pt)

In this homework you will implement k-means for image segmentation.

- **Step 1.** Take **three** photos of something on campus at the University of Delaware. Do not use very large photos. For each photo,
  - Load your photo into computer and get the pixels each of which is given by (R,G,B). Use the libraries for processing image.
  - Apply K-means to the pixels, and rewrite the image where each pixel is replaced by the mean of its cluster. Now you have a new photo.
  - Try different k from $\{1, 2, 5, 10, 20\}$. Now you have many new photos.
- **Step 2.** Questions:
  - What did you find concerning the relationship between k and the size of the new image?
  - For each of your photos, which k is the best you think? Why?
- **Step 3.** Take a new photo of which you believe k=2 is the best, and verify it use your program in step 1.

What to Turn in

Please upload.

- Your **code and a Readme file** for compiling the code.
- Your original **photos** and produced photos.
- A pdf **report** of (a) your results in step 1, (b) your answers to step 2, and (c) you findings in step 3. You should also show your photos in your report.